Documenti digitali

a cura di Roberto Guarasci
in collaborazione con Antonietta Folino

ITER

Documenti digitali
a cura di Roberto Guarasci
in collaborazione con Antonietta Folino

EDITORE:
ITER Srl
Via F.lli Bressan 14
20126 Milano

ISBN: 978-88-903419-3-9

Finito di stampare: Aprile 2013

Indice

Roberto Guarasci
Documentazione e Scienze dell'Informazione “ 5

Madjid Ihadjadene - Laurence Favier
Scienze del Documento – Scienze dell'Informazione “ 33

Enrico De Giovanni
Il documento digitale: profili giuridici “ 109

Stefano Pigliapoco
Sistemi informativi e dematerializzazione “ 145

Eduardo De Francesco
I linguaggi di descrizione documentale “ 169

Giovanni Adamo
La Terminologia “ 215

Simonetta Montemagni
Estrazione Terminologica Automatica e Indicizzazione:
Scenari Applicativi, Problemi e Possibili Soluzioni “ 241

Giorgio Gambosi - Maurizio Lancia
Data Mining e Text Mining “ 285

Paolo Ferragina
Sui motori di ricerca “ 331

Mauro Guerrini
Classificazioni bibliografiche “ 371

Antonietta Folino
Tassonomie e thesauri “ 387

Vincenzo Loia
Le ontologie “ 445

Documentazione e Scienze dell'Informazione

ROBERTO GUARASCI*

Nel dicembre 2012, anche a seguito di pressanti richieste delle associazioni professionali e di categoria, l'Istituto Nazionale di Statistica (ISTAT) ha aggiornato la sua classificazione delle professioni (CP 2011)[1] inserendo – tra le altre – quella del *conservatore dei documenti digitali* prevista, fin dal 2005, dall'art. 44 bis del Decreto legislativo 82 del 7 marzo 2005 *Codice dell'Amministrazione Digitale* (CAD)[2]. Il dettato legislativo pre-

* Università della Calabria, Dipartimento di Lingue e Scienze dell'Educazione.

1 In Italia la classificazione delle professioni (<http://cp2011.istat.it/>) è redatta periodicamente dall'Istituto Nazionale di Statistica (ISTAT) ed è per la massima parte allineata con la ISCO -*International Standard Classification for Occupations* - standard di classificazione delle professioni elaborato dall'ILO -*International Labour Organization*-. Nel gennaio 2012 il Ministero dell'Università e della Ricerca Scientifica (MIUR) con nota del 31 gennaio 2012, protocollo n. 169 relativa all'offerta formativa 2012-2013 ha, per la prima volta, *«definito un protocollo d'intesa MIUR-ISTAT, con l'obiettivo di integrare il progetto Sistema informativo sulle professioni (<http://www.istat.it/it/archivio/18841>)con le informazioni inserite nella banca dati dell'Offerta Formativa (OFF), al fine di fornire un più efficace strumento di orientamento per gli studenti»*. Ciò nel tentativo di regolamentare l'estrema variabilità delle titolazioni e dei contenuti formativi dei percorsi di studio universitari costruendo una riferibilità tra le figure professionali in uscita e le richieste del mondo del lavoro.

2 Decreto legislativo 7 marzo 2005, n. 82, *Codice dell'amministrazione di-*

scriveva che questi operasse in sinergia con un'altra figura, prevista fin dal 2000[3], il *responsabile del servizio per la tenuta del protocollo informatico, della gestione dei flussi documentali e degli archivi* nonché con il responsabile per il trattamento dei dati personali di cui alla specifica normativa.

Tralasciando il responsabile del trattamento dei dati personali, soffermiamoci sulle altre due figure. In entrambi i casi le competenze richieste prevedevano un mix di saperi[4] che difficilmente trovava riscontro nei percorsi formativi delle università italiane se si eccettuavano sporadici casi di percorsi di presunta specializzazione post laurea nei quali si cercava – con risultati spesso discutibili – di colmare le evidenti lacune tecnologiche di pubblici funzionari con *skill* quasi sempre economico-giuridici. Se si prova a leggere in maniera comparata l'incipit del D.P.R. 428/98[5] con quello del citato D.P.R. 445/00 che lo sostituisce si coglie l'intento specifico del legislatore di marcare l'evoluzione del sistema di protocollo verso un più complesso sistema di gestione documentale indicando, contestualmente, una necessaria evoluzione delle competenze degli addetti perfettamente in linea con lo scenario europeo ma in controtempo rispetto a quello nazionale.

---

gitale, in Gazzetta Ufficiale del 16 maggio 2005, n. 112, Supplemento Ordinario n. 93.

3 Decreto del Presidente della Repubblica 28 dicembre 2000, n. 445, *Testo unico delle disposizioni legislative e regolamentari in materia di documentazione amministrativa*, in Gazzetta Ufficiale del 20 febbraio 2001, n. 42, Supplemento Ordinario n. 30, art. 61.

4 Il citato articolo 61 del D.P.R. 445/2000 esplicitamente prevedeva che alla gestione fosse preposto *un dirigente ovvero un funzionario* in possesso di competenze giuridiche, documentali e tecnico-informatiche.

5 Decreto del Presidente della Repubblica 20 ottobre 1998, n. 428, *Regolamento recante norme per la gestione del protocollo informatico da parte delle amministrazioni pubbliche*, in Gazzetta Ufficiale del 14 dicembre 1998, n. 291.

Nel D.P.R. 428/98, all'art. 1 si legge:

> *Ai fini del presente regolamento s'intende: per 'gestione dei documenti', l'insieme delle attività finalizzate alla registrazione di protocollo e alla classificazione, organizzazione, assegnazione e reperimento dei documenti amministrativi, formati o acquisiti dalle amministrazioni, nell'ambito del sistema di classificazione d'archivio adottato; per 'sistema di protocollo informatico' o, in forma abbreviata, 'sistema', l'insieme delle risorse di calcolo, degli apparati, delle reti di comunicazione e delle procedure informatiche utilizzati dalle amministrazioni per la gestione dei documenti*.

Nel D.P.R. 445/00 all'art. 1 la definizione che, nel testo previgente, era del *protocollo informatico* viene rititolata «*r) sistema di gestione informatica dei documenti*», lasciandone inalterata la valenza semantica. Già nel primo testo il protocollo informatico, pur mantenendo il nome storicamente consolidato e consacrato nella legislazione unitaria dal Regio Decreto del 25 gennaio 1900, n. 35[6], diventa cosa ben diversa dal registro di carico e scarico della corrispondenza ovvero del *«registro dal quale si possa facilmente ed in ogni momento rilevare l'insinuazione e l'evasione di ogni affare»* delle Istruzioni ai vice prefetti del regno d'Italia del gennaio del 1806 nelle quali, pure, oltre al carico e scarico della corrispondenza, aveva una embrionale funzione di controllo della produttività individuale attraverso la verifica *visuale* delle pratiche inevase, nel secondo la trasformazione si è completata. Questa diversa e corretta configurazione del *protocollo informatico* come sistema di gestione dei documenti non sarà però percepita né metabolizzata e si continuerà ad afferma-

6 Regio Decreto 25 gennaio 1900 n. 35, *Approvazione del regolamento per gli Uffici di registratura e di archivio delle Amministrazioni centrali,* in Gazzetta Ufficiale del 22 febbraio 1900, n. 44.

re una neutralità del mezzo tecnologico adoperato rispetto alle finalità da perseguire[7].

Il responsabile della conservazione, per come anche si evince dai requisiti richiesti a coloro i quali intendono iscriversi all'albo dei conservatori accreditati, dovrebbe quindi possedere competenze *«a livello gestionale, della conoscenza specifica nel settore della gestione documentale e conservazione dei documenti informatici e della dimestichezza con procedure di sicurezza appropriate e che sia in grado di rispettare le norme del CAD e le regole tecniche previste dall'art. 43 del CAD in materia di sistema di conservazione di documenti informatici»*[8]. Identici i requi-

[7] Ciò sarà alla base di non pochi equivoci e contenziosi tra i quali merita di essere segnalato quello intentato davanti al Garante per la protezione dei dati personali da una dipendente ENAC – Ente Nazionale per l'Aviazione Civile - in servizio presso l'aeroporto di Malpensa che lamenta come l'erronea configurazione dei diritti di accesso ai documenti protocollati abbia impropriamente diffuso dati riservati. Il Garante, sanzionando l'amministrazione prescrive, in aggiunta alla trasmissione degli atti all'autorità giudiziaria: *«a. quale misura opportuna, di effettuare un'attività formativa indirizzata al personale della sede di Malpensa che utilizza il sistema di gestione documentale anche con il ruolo 'impiegato', illustrandone compiutamente le funzionalità e le possibili implicazioni in relazione all'applicazione della disciplina di protezione dei dati personali (punto 4.4); b. quale misura opportuna, di dare attuazione con riguardo al complessivo funzionamento del sistema gestionale di documentazione alla disposizione, richiamata in motivazione, di cui all'art. 44, comma 1-bis, d. lg. n. 82/2005 (punti 5.2 e 5.3)»*. Nel ribadire il principio della necessaria sinergia operativa tra le tre figure di cui alla nostra premessa nel rispetto delle specifiche competenze e con l'obbiettivo comune della corretta gestione documentale dei procedimenti amministrativi dell'ente, esplicitamente afferma la necessità di rendere edotto il personale sulle funzioni e potenzialità del protocollo in relazione alla possibile diffusione di informazioni riservate.

[8] DIGITPA, Circolare 29 dicembre 2011, n. 59, *Modalità per presentare la domanda di accreditamento da parte dei soggetti pubblici e privati che svolgono attività di conservazione dei documenti informatici di cui all'ar-*

siti previsti dall'art. 5 della Deliberazione CNIPA n. 11/2004 del 19 febbraio 2004 *Regole tecniche per la riproduzione e conservazione di documenti su supporto ottico idoneo a garantire la conformità dei documenti agli originali*[9] nelle quali si specifica in aggiunta al comma 4 dello stesso articolo, che *«nelle amministrazioni pubbliche il ruolo di pubblico ufficiale è svolto dal dirigente dell'ufficio responsabile della conservazione dei documenti o da altri dallo stesso formalmente designati»*[10]. Le competenze e le figure professionali, quindi, sono puntualmente e ragionevolmente definite. Non così i *luoghi* nei quali acquisirle.

Fino ad oggi alle scelte non sempre lungimiranti del sistema formativo nazionale si aggiungeva anche una difficoltà formale. In Italia – come notavamo poc'anzi – la classificazione delle professioni[11] è redatta periodicamente dall'ISTAT. Nel gennaio

*ticolo 44-bis, comma 1, del decreto legislativo 7 marzo 2005, n. 82***,** in Gazzetta Ufficiale del 8 febbraio 2012, n. 32.
In attuazione di quanto disposto dal decreto legislativo 177 del 1 dicembre 2009, il Centro Nazionale per l'Informatica nella Pubblica Amministrazione (CNIPA) è stato trasformato in DigitPA - Ente nazionale per la digitalizzazione della Pubblica Amministrazione. Nel 2012 quest'ultimo è confluito nell'Agenzia per l'Italia Digitale.

9 Deliberazione CNIPA 19 febbraio 2004, n. 11/2004, *Regole tecniche per la riproduzione e conservazione di documenti su supporto ottico idoneo a garantire la conformità dei documenti agli originali*, in Gazzetta Ufficiale del 9 marzo 2004, n. 57.

10 Nel caso dell'affidamento all'esterno del servizio, previsto espressamente nello stesso testo, l'Agenzia delle Entrate con risoluzione n. 364 del 03.10.2008 ha espressamente chiarito che unico responsabile della corretta tenuta e conservazione delle scritture contabili e di tutti i documenti fiscalmente rilevanti rimane il contribuente, anche se ha esternalizzato a più fornitori il processo di conservazione. Eventuali inadempienze degli outsourcers, pertanto, non potranno essere opposte all'Amministrazione finanziaria per giustificare irregolarità o errori nella tenuta e nella conservazione della contabilità.

11 <http://cp2011.istat.it/>.

2012 il Ministero dell'Università e della Ricerca Scientifica (MIUR) con nota del 31 gennaio 2012, protocollo n. 169 relativa all'offerta formativa 2012-2013 ha, per la prima volta, *«definito un protocollo d'intesa MIUR-ISTAT, con l'obiettivo di integrare il progetto Sistema informativo sulle professioni con le informazioni inserite nella banca dati dell'Offerta Formativa (OFF), al fine di fornire un più efficace strumento di orientamento per gli studenti»*. Ciò nel tentativo di regolamentare l'estrema variabilità delle titolazioni e dei contenuti formativi dei percorsi di studio universitari costruendo una riferibilità tra le figure professionali in uscita e le richieste del mondo del lavoro. Nella progettazione di un percorso di studi diventa quindi obbligatorio il perfetto allineamento tra le figure professionali in uscita e la classificazione delle professioni che – come già detto – ha inserito il *Conservatore dei documenti digitali* solo dal dicembre 2012 colmando un vuoto quasi decennale tra la prescrizione normativa e il recepimento della figura professionale. Eppure oltre alle previsioni di obbligatorietà il complesso quadro normativo nazionale dalla Legge 241/90[12] in poi per continuare con il Decreto Legislativo 150/09[13] e finire con la Legge 221/2012[14], da molti anni presupponeva sempre di più la disponibilità di informazioni quantitative da parte dei decisori delle pubbliche ammi-

[12] Legge 7 Agosto 1990 n. 241, *Nuove norme in materia di procedimenti amministrativi e di diritto di accesso ai documenti amministrativi*, in Gazzetta Ufficiale del 18 agosto 1990, n. 192.

[13] Decreto Legislativo 27 ottobre 2009, n. 150, *Attuazione della legge 4 marzo 2009, n. 15, in materia di ottimizzazione della produttività del lavoro pubblico e di efficienza e trasparenza delle pubbliche amministrazioni*, in Gazzetta Ufficiale del 31 ottobre 2009, n. 254, Supplemento Ordinario n. 197.

[14] Legge 17 dicembre 2012, n. 221, *Conversione in legge, con modificazioni, del decreto-legge 18 ottobre 2012, n. 179, recante ulteriori misure urgenti per la crescita del Paese*, in Gazzetta Ufficiale del 18 dicembre 2012, n. 294, Supplemento Ordinario n. 208.

nistrazioni non mai esplicando compiutamente a quali profili curricolari queste crescenti attività erano in carico. La corretta realizzazione del ciclo delle performance o l'applicazione delle norme per il contrasto alla corruzione nelle pubbliche amministrazioni – per citarne solo alcune – non possono prescindere dalla disponibilità puntuale ed aggiornata dell'informazione documentale. La realizzazione del Fascicolo Sanitario Elettronico, previsto espressamente come obbligo dalla citata Legge 221/2012 non è realizzabile senza una corretta definizione degli oggetti documentali che lo compongono ed in assenza di una codifica standardizzata delle informazioni necessarie per evitare ambiguità ed errori in situazioni critiche e di emergenza. Esso poi altro non è che un sottoinsieme del più generale Fascicolo del Cittadino del quale l'Anagrafe Nazionale della Popolazione Residente costituisce l'ineliminabile presupposto, alla stessa stregua del Fascicolo dello Studente, del Fascicolo Previdenziale, ecc. Unica Anagrafica con basket documentali settoriali a somiglianza di quanto accadeva nelle strutture documentali analogiche nelle quali si realizzavano un fascicolo e tanti sottofascicoli, partizioni logico-fisiche del primo del quale – nel cartaceo – si percepiva materialmente l'unicità e la riconducibilità parte-tutto nella materialità dei contenitori che racchiudevano gli atti. In un punto preciso della storia della nazione l'inefficienza e l'elefantiasi della macchina burocratica dello Stato è stata personificata nella gestione cartacea dei procedimenti amministrativi quasi come se un supporto scrittorio potesse, per la sua stessa natura, essere il responsabile di inadeguatezze strutturali di ben altra origine e dimensione.

Nella lunghissima storia dei materiali scrittori i periodi di crisi e di transizione sono sempre stati attraversati da demonizzazioni più o meno evidenti e manifeste di ciò che subentrava rispetto a ciò che cessava. *«Un frate dominicano della fine del XV secolo. Filippo di Strata, persino disse: Se il mondo è andato avanti perfettamente per 6000 anni senza la stampa, perché*

cambiare adesso?» (Hibbits, 1977)[15]. La paura era quella di perdere il controllo sull'intero ciclo produttivo del libro e produrre opere di qualità più scadente in quanto prive del controllo culturale del copista. Inizialmente queste fosche previsioni si avverarono in più di un caso e i primi tipografi commisero errori spesso grossolani anche a causa di una certa dose di improvvisazione, ma era la mancata specializzazione umana non il mezzo utilizzato la causa come dimostrò il prosieguo della storia dalla quale pare non abbiamo però tratto grandi insegnamenti. Tornando ai giorni nostri, un momento di snodo è sicuramente rappresentato dagli anni Ottanta del XX secolo.

Il 28 marzo 1980 il Consiglio Nazionale delle Ricerche trasmetteva al Ministro per la ricerca scientifica la proposta di un progetto finalizzato sull'organizzazione e il funzionamento della Pubblica Amministrazione elaborata da un comitato presieduto da Aldo M. Sandulli. La proposta incontrò il favore dell'allora ministro Giancarlo Tesini che dette la sua approvazione e, con successiva deliberazione del febbraio 1982, il presidente del C.N.R. Ernesto Quagliariello nominò una commissione – presieduta da Mario Nigro – per la realizzazione dello studio di fattibilità[16] che venne, poi, il 6 marzo 1985, approvato dal Comitato

[15] BERNARD J. HIBBITS, *Oggi come ieri. Scettici, scribi e la fine delle riviste giuridiche*, in «Informatica e Diritto», vol. 9, n. 2, 1977, p. 81.

[16] Ne facevano parte, tra gli altri, Sabino Cassese, Giuseppe De Rita, Elio Gizzi, Alberto Zuliani oltre allo stesso Aldo M. Sandulli. Finalità principale del progetto era *«la sistemazione delle conoscenze disponibili e lo svolgimento di tutte le ulteriori rilevazioni necessarie ad ottenere una rappresentazione, la più possibile compiuta, della struttura e dell'azione amministrativa e di valutare la funzionalità della pubblica amministrazione, anche in vista della formulazione di proposte relative alle maggiori disfunzioni ed alle procedure, dirette ad assicurare maggiore efficienza»*. C.N.R., *Progetto finalizzato sull'organizzazione ed il funzionamento della pubblica amministrazione – Relazione finale sull'attività svolta e sui risultati conseguiti*, rapporto interno s.d., 1995-96, pp. 8-9.

interministeriale per la programmazione economica. La dotazione finanziaria quinquennale di circa trenta miliardi di lire era indice della rilevanza del progetto nelle strategie di ricerca ed innovazione della nazione in quel periodo. Nel luglio 1985 la responsabilità delle attività venne affidata a Giuseppe Guarino e, in seguito alle sue dimissioni, a Sabino Cassese che lo porterà a conclusione nei primi anni '90.

Ciò che emerge con chiarezza nella relazione finale è la netta bipartizione tra il documento amministrativo, fulcro e base del procedimento, e il dato/informazione che – in quel contesto – ha finalità eminentemente statistiche e di supporto alla decisione. Questa necessaria divisione concettuale – anche riconducibile al quadro normativo vigente – è sottolineata dagli stessi estensori della relazione allorquando si parla dell'amministrazione finanziaria specificando che

> *la strutturazione dell'intero sistema informativo intorno alle esigenze, spesso contingenti, di gestione di un quadro normativo sempre più dettagliatamente specificato, ha oggettivamente limitato le capacità dell'amministrazione finanziaria di armonizzare i diversi piani della propria azione (applicazione della normativa, organizzazione e procedimenti, acquisizione e sviluppo della tecnologia) attraverso la crescita di un sistema informativo integrato e polifunzionale*[17].

[17] Nella relazione finale sull'attività svolta e sui risultati conseguiti viene enunciata una articolata definizione di procedura amministrativa: *«Muovendo da un approccio multidisciplinare e constatati i diversi significati attribuiti al termine "procedura" nei differenti contesti disciplinari in cui viene normalmente usato, si è adottata una definizione ampia, comprensiva dei seguenti aspetti: il procedimento amministrativo; il meccanismo di decisione inteso come l'insieme di regole e criteri attraverso cui l'espletamento di una attività porta ad un determinato esito; le modalità del processo produttivo, cioè l'organizzazione delle risorse utilizzate; l'insieme di informazioni raccolte e mobilitate all'interno del processo amministrativo»*, 1992. C.N.R., *Progetto finalizzato sull'organizzazione cit.*, pp. 73-75.

I sistemi informativi della Pubblica Amministrazione, fin dal loro nascere, saranno spesso considerati semplici gestori di processi contingenti dei quali non sempre si riuscirà a cogliere la visione d'insieme anche per la frammentarietà e disorganicità delle singole realizzazioni. I legislatori degli anni successivi, pur delineando un quadro normativo per molti versi completo anche se non sempre compiutamente operativo, seguiranno una strada simile, concependo gli interventi normativi sulla innovazione e digitalizzazione della Pubblica Amministrazione, non come un necessario adeguamento dell'esistente al mutare degli scenari e dei contesti ma come generazione ex novo di un corpus autonomo che, proprio in quanto tale, mostrerà tutti i suoi limiti specie quando non accompagnato da specifiche misure sanzionatorie. Le competenze professionali richieste seguiranno questa artificiale dicotomia tra il tecnicismo del dato e la concettualità dei contenuti.

La già notata visione efficientista che vedeva nella produzione legislativa specializzata e nell'uso massivo dell'informatica uno dei modi per ovviare alla inefficienza dei servizi offerti dall'amministrazione pubblica determinò la convinzione che il contenuto dell'azione amministrativa fosse ipostatico ovvero trasferibile liberamente in *contenitori* tra loro diversi non tenendo in nessun conto l'esperienza umana ed il contesto socio-culturale nel quale l'azione stessa si era originariamente collocata ne, tantomeno, i mezzi tecnologici utilizzati.

«*Miliardi a raffica contro la burocrazia. Servono per i computer che renderanno veloci rimborsi IVA, pratiche di pensione, rinnovo patenti*» titolava il «Corriere della Sera» del 4 maggio 1993; *«Lo Stato gioca la 'carta' dell' informatica per ridare efficienza alla pubblica amministrazione»* sempre il «Corriere della Sera» del 15 dicembre 1994.

Su questo substrato la profonda crisi politica del paese nei primi anni Novanta ottunde la capacità programmatoria di lungo respiro e accentua stereotipi e modelli duali che hanno avuto una

vita spesso commisurata alle risorse stanziate per la loro sperimentazione. *«I criteri allocativi [esplicitava sempre il progetto C.N.R.] sono risultati condizionati in modo determinante e crescente dagli interessi delle imprese»*[18]. La divaricazione sempre più marcata tra dati/informazioni e documenti se da un lato ha protratto nel tempo l'ambiguità semantica e concettuale tra i due termini ha però anche attenuato gli effetti del taylorismo informativo nei pubblici uffici rendendo meno evidente la costruzione dell'ufficio-fabbrica di molte realtà private ed è stata perciò spesso accettata se non incoraggiata.

In quegli anni l'informatica diventò quindi l'alfiere di una *nuova Italia* anche perché l'industria informatica nazionale, produttrice di apparati più che di soluzioni, intravide in quella favorevole contingenza una opportunità difficilmente ripetibile di conquistare definitivamente un mercato gestito dall'offerta e non dalla domanda. Non il ripensamento di una società in transizione verso il digitale ma un digitale applicato in maniera posticcia ad una amministrazione che apparentemente lo subiva ma in effetti lo rifiutava non riuscendo a percepirne le motivazioni, la logica e, prima di tutto non avendo evidenza dei miglioramenti reali nella gestione dei procedimenti amministrativi. Sono gli anni delle centinaia di miliardi di vecchie lire destinate al progetto *Giacimenti culturali*. Quasi trent'anni dopo – nel 2012 – resta la considerazione della Corte dei conti, nella memoria del procuratore generale Salvatore Nottola al giudizio sul rendiconto dello Stato «*Nonostante vari tentativi di giungere a una stima attendibile dei beni culturali, non esiste oggi una catalogazione definitiva specie per i reperti archeologici. Inoltre, per i grandi musei statali non esiste una stima del valore delle opere possedute*». Un'orgia di spesa informatica e infinite difficoltà nella gestione del bene rinveniente (le banche dati) non ancora completamente

[18] C.N.R., *Progetto finalizzato sull'organizzazione, cit.*, p. 79.

conclusa e affidata per una parte non trascurabile al volontariato ed all'iniziativa dei singoli. Questo precario equilibrio riuscirà a sopravvivere senza grandi traumi fino alla fine del decennio complice anche la preminenza data al documento digitalizzato più che al digitale nativo. All'avvento del nuovo millennio eventi internazionali di grande impatto mediatico evidenzieranno l'impossibilità del perdurare della separazione. Il ruolo delle mail nella mancata previsione del fallimento della Enron nel 2001 e la funzione dei social network nell'affaire Strauss-Kahn nel 2011[19] saranno solo i punti estremi di una diversa percezione della inevitabile riconsiderazione della dualità tra il documento ed i suoi avatar: dati e informazioni. Purtroppo gran parte dei danni erano già stati fatti e la contrazione generale delle risorse limiterà la possibilità di correzioni e ripensamenti organici degli scenari (Guarasci, 2012)[20].

Eppure fino allo scoppio della seconda guerra mondiale la situazione italiana nella gestione dell'informazione documentale e nella formazione delle connesse figure professionali non solo era pienamente allineata a quella degli altri paesi europei ma anzi, per molti versi, costituiva un modello di integrazione culturale tra le scienze, resa viepiù evidente e necessaria all'indomani della prima guerra mondiale quando la sottovalutazione dei saperi tecnologici ed il mancato raccordo tra scienza ed industria avevano evidenziato la superiorità delle potenze dell'Asse (Maiocchi, 2004)[21]. A cominciare dal primo dopoguerra saranno proprio le *hard science* quelle che incentiveranno e trarranno i maggiori benefici dalla costruzione di centri di documentazione e sistemi

[19] <http://affordance.typepad.com/mon_weblog/2011/05/dsk-le-temps-reel-documente.html>.

[20] Cfr. ROBERTO GUARASCI, *Documenti, Dati, Viste Documentali*, in «e-Health care», n. 17, 2012, pp. 81-88.

[21] ROBERTO MAIOCCHI, *Scienza e Fascismo*, Roma, Carocci, 2004, p. 21 e sgg.

di gestione dell'informazione documentale dal Centro di Documentazione Elettrotecnica dell'Università di Padova al Centro di Fotodocumentazione Scientifico Tecnica del Politecnico di Milano al Centro di Documentazione della Glaxo di Verona[22].

Emblematica sarà l'esperienza del Centro Nazionale di Documentazione Scientifica del C.N.R. (Guarasci, 2011)[23] che fornirà informazione documentale a supporto dei diversi ambiti della scienza e della tecnica non solo in Italia ma anche in molti paesi europei costruendo un modello culturale ispirato all'Istituto Internazionale di Bibliografia di Bruxelles. Nel periodo immediatamente precedente lo scoppio della seconda guerra mondiale il Centro diventerà punto di aggregazione e di coordinamento di tutte le iniziative similari esistenti sul territorio nazionale e pubblicherà anche la *Bibliografia Internazionale dell'Ingegneria e dell'Industria*. Le necessità dell'industria pesante, le esigenze dell'attività di riarmo e la connessa politica dell'autarchia produrranno in Italia un momento di grande attenzione verso la gestione dell'informazione documentale a fini brevettuali e di trasferimento tecnologico. Ogni azienda ed ogni ente degno di questo nome avranno un proprio centro di documentazione piccolo o grande che sia e verrà dato anche avvio ad una iniziativa di censimento di quelli esistenti. Quando il 28 dicembre 1944, finita la Guerra, si insedia il nuovo presidente Gustavo Colonnetti questi, il 24 gennaio 1945, scrivendo al Ministro dell'Aeronautica avrà a dire: *«Uno degli istituti da potenziare è il Centro Nazionale di Documentazione Tecnica»*[24].

---

22 *Gli Organismi italiani di Documentazione*, in La Documentazione in Italia, Roma, C.N.R., 1952, pp. 183-204.

23 Cfr. ROBERTO GUARASCI, *La memoria della scienza: l'Archivio Tecnico Italiano e Il Centro Nazionale di Documentazione Scientifica*, in Archivi Privati, Guarasci R., Pasceri E. (a cura di), Roma, C.N.R., 2011, pp. 195-219.

24 Archivio Centrale dello Stato (ACS), C.N.R., Centro Nazionale di Documentazione Scientifica (CNT), busta 943, fasc. 2.

Nonostante queste affermazioni, due anni più tardi, il 29 dicembre 1947, il responsabile della struttura Uberto Bajocchi, scrivendo al collega ungherese J. Zeleny, illustrerà la mutata situazione dell'immediato dopoguerra:

> *prima della guerra i centri di documentazione tecnica erano più o meno coordinati dal Consiglio Nazionale delle Ricerche attraverso l'organizzazione che vi scrive, cioè attraverso il Centro Nazionale di Documentazione Tecnica che rappresenta l'Italia in seno alla Federazione Internazionale di Documentazione con sede all'Aja. Dopo la guerra l'attività di tutti i centri è diminuita ed il collegamento di quelli scientifici a questo è divenuto meno serrato. Infatti le difficoltà di comunicazioni con l'estero e le difficoltà di comunicazione anche all'interno aggravano oltre ogni dire la situazione delle organizzazioni documentarie […]. Prima della guerra esso pubblicava una grande Bibliografia Internazionale dell'ingegneria e dell'Industria che usciva ogni due mesi con otto parti a bimestre e – a fine anno – indici alfabetici e sistematici. Il 31 dicembre 1943 questa pubblicazione è finita […]. Nel luglio 1945 profittando del materiale bibliografico messo a disposizione dalle ambasciate d'America e di Inghilterra e dal British Council questo Centro ha ripreso la pubblicazione di una piccola bibliografia internazionale limitata alla lingua italiana, francese ed inglese*[25].

Il Centro di Documentazione e, più in generale, la Documentazione Italiana non riescono – nel secondo dopoguerra – a fare quel salto di qualità che, in altre nazioni europee, darà autonoma configurazione alla disciplina permettendo l'estrema affermazione di Robert Pages secondo il quale archivisti e conservatori di musei sono professioni *pré-documentalistes* ed i bibliotecari *un*

[25] ACS, C.N.R., CNT, Busta 943A bis, fasc.4.

cas particulier de documentaliste (Briet, 1951)[26]. L'attività, embrionalmente perseguita, di riflessione metodologica sull'estrazione ed elaborazione dell'informazione documentale verrà progressivamente abbandonata e sarà, in prosieguo di tempo, causa della completa espropriazione del ruolo anche in concomitanza con la brusca accelerazione tecnologica degli anni successivi e con l'inizio della dicotomia dato/documento. Mentre in altri contesti nazionali i Documentalisti riescono a connotarsi autonomamente all'interno delle scienze dell'Informazione o vengono riassorbiti, de facto, nelle discipline bibliografiche e biblioteconomiche contribuendo, però, ad una sostanziale evoluzione di queste verso ambiti tecnologicamente avanzati, in Italia – tranne che in casi isolati – essi si dissolvono nel mondo delle scienze del libro che, fortemente condizionate dal contesto culturale delle scienze umane nel quale si trovavano ad essere inserite, accentuano in quegli anni gli aspetti conservativi rispetto a quelli operativi (Alberani, Poltronieri, 2003)[27]. Paradigmatica – in tal senso – è l'evoluzione dell'*American Library Association* dalla quale gemma la *Special Libraries Association*, poi – nel 1937 – l'*American Documentation Institute* e infine *l'American Society for Information Science*. Ancora nel 1952, definendo le connotazioni professionali del documentalista a servizio dell'industria, Alessandra Omodei precisava prima di tutto le motivazioni alla base dell'esistenza della disciplina che rintracciava nell'ovvia necessità della definizione dello stato dell'arte: *«Per studiare una nuova lavorazione occorre raccogliere, nel più breve tempo possibile, e con una spesa ragionevole, dati tecnici ed economici che spesso non rientrano neppure nella particolare competen-*

[26] SUZANNE BRIET, *Qu'Est-Ce Que la Documentation*, Parigi, Edit, 1951, p. 20.

[27] Cfr. VILMA ALBERANI, ELISABETTA POLTRONIERI, *La Documentazione rispetto alle altre discipline dell'Informazione*, in «AIDA Informazioni», a. 21, n. 3, 2003, pp. 19-47.

za di chi compie la ricerca […] si deve ricorrere quindi all'aiuto di un esperto di documentazione» (Omodei, 1952)[28]. Specificava poi che il documentalista: *«non deve sostituirsi, nello studio di un problema, a chi gli chiede notizie bibliografiche; deve essere guida intelligente nella ricerca di quanto a quel problema si riferisce sia in generale sia sotto quel determinato punto di vista teorico o pratico o economico che più interessa»*[29]. Una funzione quindi di intermediazione informativa – un *knowledge worker*[30] ante litteram – capace di coniugare conoscenze di dominio e abilità tecniche di ricerca documentale con una spiccata connotazione applicativa. *«E' il buon senso non il teoricismo che deve guidare il documentatore che deve immedesimarsi nelle esigenze di coloro che gli richiedono una documentazione, assumerne la mentalità»*[31]. Sono gli ultimi lampi. L'esplosione informativa dei decenni successivi determinerà una rapida evoluzione del modello di organizzazione e gestione della conoscenza sia dal punto di vista della struttura che dei tempi di riposta. Ad un modello sostanzialmente gerarchico, piramidale, se ne sostituirà rapidamente uno reticolare e multicentrico nel quale il valore e la determinazione dei nodi cognitivi, le categorie, non saranno stabili o a variabilità lenta ma tenderanno assumere una dinamica sempre più accelerata all'interno della quale non sempre sarà possibile individuare con certezza una logica associativa. Dai *le-*

---

28 ALESSANDRA OMODEI, *La Documentazione e l'industria*, in La Documentazione in Italia, Roma, C.N.R., 1952, p. 132.

29 *Ibidem*.

30 Il termine *knowledge worker*, ovvero lavoratori della conoscenza, è stato coniato da Peter Drucker nel 1993 ma, nella sua accezione originaria, non definiva una specifica figura professionale inquadrabile in un sistema classificatorio bensì un insieme di professioni coinvolte nella gestione dell'informazione ovvero nella produzione di valore attraverso sistemi di capitalizzazione della conoscenza e di utilizzo dell'informazione per finalità operative e programmatorie.

31 OMODEI, A., *op. cit.*, p. 133.

gami deboli di (Granovetter, 1973)[32] ci si incammina verso i *lampi* di (Barabasi, 2011)[33].

> *Questo modello [scrive nel 1994 Paolo Bisogno] è in via di sostituzione sotto la spinta di un rapido sviluppo tecnologico. La diversificazione dei supporti, l'accrescersi e il variare delle modalità di distribuzione, la disponibilità di teorie facili, amichevoli, ha certamente avuto come effetto immediato lo sviluppo dell'accesso e della consultazione diretta e una modifica e un ampliamento delle modalità di distribuzione gestite in modo autonomo e parallelo* (Bisogno, 1994)[34].

Il modello multicentrico tende – specie nella sua fase di avvio – ad eliminare ogni intermediazione cognitiva vissuta come distorcente ed antidemocratica.

---

32 Cfr. MARK S. GRANOVETTER, *The Strength of Weak Ties*, in «American Journal of Sociology», vol. 78, n. 6, maggio 1973, pp. 1360-1380.

33 Cfr. ALBERT LASZLO BARABASI, *Lampi*, Frediani S. (traduzione di), Torino, Einaudi, 2011.

34 PAOLO BISOGNO, *Il Futuro della Memoria – elementi per una teoria della Documentazione*, Milano, Franco Angeli, 1994, p. 14.
Significativa ma isolata resterà l'esperienza di Paolo Bisogno e del suo Istituto di Studi sulla Ricerca e Documentazione Scientifica (ISRDS). Istituito inizialmente come Laboratorio sulla Ricerca e sulla Documentazione rappresenterà il tentativo di far diventare la Documentazione non più attività di servizio ma oggetto di studio all'interno del panorama delle competenze del C.N.R. Il Laboratorio, nel febbraio 1976, diventerà Istituto di Studi sulla Ricerca e Documentazione Scientifica (ISRDS). Nel 2001 all'Istituto, diventato nel frattempo Istituto Sperimentale, verrà accorpato l'Istituto di ricerca sulla dinamica dei sistemi economici e la nuova realtà assumerà il nome di *Istituto di studi economici sull'innovazione e le politiche della ricerca*. Rimarrà in vita fino al 2003 quando sarà definitivamente soppresso. Cfr. anche CARLA BASILI, *Tappe salienti della Documentazione nel periodo 1983-2003 nella rassegna dei convegni AIDA, ASIS e FID*, in «AIDA Informazioni», a.21, n. 3, 2003, pp. 105-117.

Lentamente il termine Documentazione diventa sempre meno indicativo di una disciplina e la connessa figura professionale, il documentalista, sopravvive solo in pochissimi ambiti professionali di nicchia. Sempre più spesso il termine informazione prima soppianta come sinonimo d'uso quello di documentazione poi ne diventa l'antitesi positiva. Documentazione diventa un termine talmente generico da necessitare di puntuali specificazioni per ridurne l'ambiguità concettuale.

> *La ragione [afferma J. Meyriat] non sta soltanto nella forza delle abitudini, oppure nell'attaccamento degli interessi ad un vocabolo che ha richiesto molto tempo per essere riconosciuto, ma soprattutto nella polisemia della parola informazione. Per la maggior parte dei nostri contemporanei la parola evoca, innanzitutto, il contenuto dei giornali e degli altri media. Quando si vuole far comprendere che si tratta dell'informazione cui si rivolgono le attività documentarie, si è portati a precisare con un epiteto di valore del sostantivo: informazione specializzata, informazione scientifica e tecnica, informazione professionale. In conclusione, se il termine documentazione è spesso sentito come limitativo e ristretto, quello di informazione ha un contenuto troppo largo e indefinito. Così si è portati volentieri ad accoppiarli, nell'idea che la vicinanza di ciascuno dei due aiuti alla comprensione dell'altro* (Meyriat, 1996)[35].

Archivisti e Bibliotecari pur se con accentuazioni diverse riusciranno a mantenere, fino ad oggi, un loro ruolo ed una loro fisionomia solo continuando a privilegiare gli aspetti storico-conservativi rispetto a quelli operativi. Diventano i custodi del patrimonio librario e documentale, pubblico e privato, del quale ga-

[35] JEAN MEYRIAT, *La Documentazione: elementi per un riesame*, in La Documentazione in Italia, Paci A.M. (a cura di), Milano, Franco Angeli, 1996, p. 100.

rantiscono la fruizione in un ambiente a volte digitale ma sempre parallelo al e rigidamente circoscritto in modo da fare della peculiarità un ulteriore tratto distintivo e *difensivo* della professione. Le *morte spoglie del passato* e le *bianche case dei morti* si configurano come l'ultima frontiera (Cassese, 1949)[36]. Il progressivo utilizzo del Web in sostituzione dei cataloghi bibliografici è – oggi – il primo sintomo che questa strategia di approccio non è più in grado di reggere e la progressiva disponibilità di libri e documenti in formato digitale nativo rappresenterà un ulteriore momento di crisi di quel modello faticosamente costruito spostando l'asse delle competenze verso settori spesso totalmente estranei ai professionisti degli archivi e delle biblioteche[37].

Mentre in altre nazioni il termine *Documentazione* si fonde naturalmente in quello di Scienze dell'Informazione, in Italia il concetto di *information science* o di *sciences et techniques de l'information et de la documentation* (Guinchat, Menou, 2003)[38] è diventato un Giano bifronte: da un lato l'Informatica e dall'altro l'Ingegneria dell'informazione entrambe sostanzialmente autoconsistenti. Il Documentalista e la Documentazione sono spariti senza lasciare nemmeno traccia nel lessico d'uso[39]. Pochissi-

---

36 Leopoldo Cassese, *Intorno al concetto di materiale Archivistico e materiale Bibliografico*, in «Notizie degli Archivi di Stato», a. 9, 1949, p. 34.

37 Cfr. Mauro Guerrini, Tiziana Possemato, *Linked Data: Un nuovo alfabeto del web semantico*, in «Biblioteche Oggi», aprile 2012, pp. 7-15; Carlo Bianchini, *Library Linked Data e il Futuro delle Biblioteche*, 2012. <https://sites.google.com/site/homepagecarlobianchini/Ricerca/library-linked-data-e-il-futuro-delle-biblioteche>.

38 Cfr. Claire Guinchat, Michel Menou, *Introduction générale aux sciences et techniques de l'information et de la documentation*, Parigi, 1990; Cfr. anche Carla Basili (Ed.), *Information Literacy in Europe: a first insight into the state of the art of information literacy in the european union*, Roma, 2003.

39 *«Il documentalista italiano pensa di sé di essere nato troppo tardi e di far parte di una élite dispersa e sparuta, tanto da non avere l'energia e la mo-*

mi si sono addolorati per la perdita perché è mancata anche la memoria di quello che era stato e di quello che poteva essere.

Eppure nella mappa epistemologica delle interconnessioni esistenti tra le Scienze dell'informazione elaborata da Peter Ingwersen anche ai fini dell'analisi bibliometrica (Almind, Ingwersen, 1997)[40] è chiaramente esplicitata una interconnessione multipla tra le scienze dell'informazione e una necessarietà di apporti disciplinari diversi, dalla linguistica alla psicologia cognitiva, alla tecnologia (Baldazzi, 2002)[41].

Ancora una volta le università non hanno fatto eccezione e non sono state estranee al più generale processo di disallinea-

tivazione per fondare una corporazione: ci si riferisce ovviamente ai bibliotecari che si augurano di avere presto un ordine professionale, una specie di salvagente vieux jeu nella bufera – con inondazioni – dell'unificazione europea. Tuttavia il documentalista italiano esiste e cresce in un contesto internazionale, ma è maggiormente travagliato, che non i colleghi stranieri, dagli squilibri italiani e soprattutto dalle caratteristiche poco 'europee' degli studi universitari e post-universitari in Italia».
VALENTINA COMBA, *Esperienze e prospettive di formazione professionale nel settore privato per i documentalisti*, in Informazione e Documentazione: temi trasversali di formazione, Roma, 1992, pp. 57-58.
A conclusione della Tavola Rotonda sul tema: La disciplina Documentazione nelle Università Italiane, Paolo Bisogno affermava: *«La formazione dei documentalisti è appannaggio del settore privato» … «alle poche cattedre di biblioteconomia, bibliografia, alle pochissime di documentazione, non si affianca una organizzazione didattica articolata.. non abbiamo una politica della documentazione né una politica delle biblioteche…»*.
Paci M.A. (a cura di), *La Documentazione nelle Università Italiane*, Roma, 1989, p. 75; Cfr. anche ROBERTO GUARASCI, *Libri, Documenti e altre storie. L'insegnamento della Documentazione nelle università italiane*, in «AIDAInformazioni», a. 21, n. 3, luglio-settembre 2003, pp. 47-59.

40 Cfr. TOMAS C. ALMIND, PETER INGWERSEN, *Infometric analyses on the World Wide Web: Methodological approaches to 'webometrics'*, in «Journal of Documentation», a. 53, n. 4, 1997, pp. 404-426.

41 ANNA BALDAZZI, *Le scienze dell'informazione e le teorie della transizione*, in «AIDAInformazioni», a.20, n. 1, gennaio - marzo 2002, p. 28.

mento delle competenze ma anzi lo hanno in molti casi perpetuato pienamente allineandosi al trend del paese. Le Scienze dell'Informazione hanno assunto una connotazione esclusivamente tecnicistica e saperi umanistici e saperi tecnologici, lungi dal completarsi, si sono rigidamente divisi e l'idea – di per sé corretta – della valutazione prevalentemente quantitativa ha fatto il resto. Costruiti degli ambiti disciplinarmente circoscritti all'interno dei cui confini culturali era necessario chiedere di essere valutati in ordine alla scientificità della propria produzione e della propria attività di ricerca si è accelerato un processo di ripiegamento. I pochi tentativi di trasversalità esistenti sono caduti davanti allo spettro dell'eterodossia e della non congruità. Il panorama dei corsi di studio universitari ha espunto la gran parte delle sinergie tra le scienze del testo e del documento e le scienze dell'informazione lasciando al mondo extra universitario la formazione e l'acquisizione delle competenze non solo obbligatorie ma sempre più richieste dalla società civile e dal mondo delle imprese. Lontano e di più ampia portata risuona il monito dell'Unione Europea della necessità di una comune politica per la gestione delle competenze e dei posti di lavoro in ambiente digitale con l'avvertenza di una fortissima carenza di figure professionali entro il 2015[42]. Periodicamente però riaffiorano dei *lampi*. Google nel 2012 riscopre Paul Otlet come teorico del Web[43]. La metafora dei *Linked Open Data* è quella di un *cervel-*

[42] Per approfondimenti si rimanda al seguente url <http://europa.eu/rapid/press-release_IP-12-1389_it.htm>.

[43] Per approfondimenti si rimanda al seguente url <http://museumpublicity.com/2012/03/14/mundaneum-and-google-announce-partnership>.
«Nonostante alcune evidenti differenze nel pensiero e nei progetti di Paul Otlet, Vannevar Bush ed Eugene Garfield è dunque possibile rinvenire gli elementi essenziali a caratterizzare la filosofia e i principi che stanno alla base del web».
FRANCESCA DI DONATO, *La Scienza e la Rete*, Firenze University Press, 2009, p. 51.

lo del mondo cara al Mundaneum ed all'Istituto Internazionale di Bibliografia. I *Knowledge Worker* di (Drucker, 1993)[44], così come i Documentalisti del Centro Nazionale di Documentazione del C.N.R. negli anni '50 estrapolano informazioni dai diversi media per aumentare la catena del valore, il processo di creazione della conoscenza teorizzato da (Nonaka, Takeuchi, 2009)[45] orientato all'estrazione di informazioni dall'interno delle organizzazioni presenta fortissime similitudini con gli assiomi di base della Documentazione della prima metà del secolo scorso. Del Resto (Rayward, 1997)[46] aveva già suggerito che le teorie di Otlet e dei suoi contemporanei costituivano le teorie di base delle scienze dell'informazione specificando poi che

> *accettiamo che scienza dell'informazione sia un termine ora convenzionalmente adottato sulla base dei tentativi degli ultimi cinquanta anni o quasi per studiare in modo formale e rigoroso procedimenti, tecniche, condizioni ed effetti che si rendono necessari per migliorare l'efficacia dell'informazione, variamente definita e compresa, come dispiegata ed usata per una serie di scopi connessi ai bisogni individuali, sociali e collettivi* (Rayward, 1996)[47].

Ciò però non individua una ragion d'essere della disciplina

---

44 Cfr. PETER DRUCKER, *Post-Capitalist Society*, New York, Harper Business, 1993.

45 Cfr. IKUJIRO NONAKA, HIROTAKA TAKEUCHI, *The Knowledge Creating Company*, Stella G. (traduzione di), Milano, Guerini, 2009.

46 W.BOYD RAYWARD, *The origin of information science and the International Institute of Bibliography/International Federation for Information and Documentation (FID)*, in «Journal of the American Society for Information Science», vol. 48, n. 4, 1997, pp. 289-300.

47 W.BOYD RAYWARD, *The History and historiography of information science: some reflections*, in «Information Processing and Management», vol. 31, n. 1, 1996, pp. 3-17.

ma ne connota al più l'essere stata la base teorica di una multiforme produzione intellettuale che, presentando significative aree di sovrapposizione, si è poi evoluta secondo direttive di sviluppo proprie sia in ragione della finalità d'uso che delle diverse matrici filosofiche e concettuali di partenza[48]. La ridefinizione di un ruolo e di uno spazio risiede però proprio nel riconoscere il valore di queste necessarie aree di sovrapposizione culturale con altri ambiti disciplinari riappropriandosi del ruolo di mediazione informativa non più tra il documento e l'utente finale ma tra i documenti e il web dei dati. Fin dal suo nascere il documentalista – o comunque vogliamo oggi chiamarlo – è stato legato ai dati e non ai documenti, all'informazione e non ai supporti materiali.

L'estrazione ed elaborazione dell'informazione documentale, la costruzione delle strutture di classificazione ed i formalismi di rappresentazione sono sicuramente alcuni degli elementi di una nuova frontiera verso la quale andare con la certezza che la figura professionale non sarà più monolitica ma dovrà necessariamente disaggregarsi in unità professionali che condividono una base comune di conoscenze e delle abilità peculiari del dominio prevalente.

«L'un des élements du changement donc devrait être repésenté par le concept collaboration» (Boadas I Raset, 2010)[49], afferma l'*International Council on Archives* testimoniando come la consapevolezza stia progressivamente estendendosi a tutte le scienze del testo e del documento. L'evoluzione del web e l'aumento della pervasività del digitale che, inizialmente, hanno rappresentato un momento di rottura diventano ora un elemento di spinta alla riappropriazione di un ruolo ed alla definizione di una nuova mappa epistemologica delle competenze in un universo virtuale nel quale è sempre più pressante la richiesta di sinergie,

48 Cfr. anche ALBERANI, V., POLTRONIERI, E., *op.cit*.

49 JOAN BOADAS I RASET, *De Quoi les citoyens ont-ils-besoin? S'adapter ou disparaître!*, in «Comma», vol. 1, 2010, p. 106.

di collegamenti, di strutture di classificazione delle quali il Memex di Vannevar Bush rappresenta l'archetipo[50].

Nella prima metà dell'Ottocento il capitano William Allen in navigazione sul Niger (Allen, Thompson, 1848)[51] scoprì che i tamburi parlanti usati dagli indigeni *primitivi* erano tonali e disambiguavano il significato mediante frasi di contesto. Trasmettevano frasi intere e complesse contestualizzate non parole semplici o messaggi standardizzati di avvertimento, di gioia o di paura. Erano in grado di creare neologismi e di esprimere significati complessi. Nello stesso periodo Samuel F.B. Morse era alle prese con il suo codice telegrafico che doveva viaggiare lungo i fili del telegrafo. Anche in questo caso ognuno ignorava l'esistenza dell'altro, lontani dal sogno visionario di Xanadu[52] metafora di un mondo atomicamente disaggregato nel quale tracciatori di strade con diverse uniformi e diversi nomi convergono verso un unico punto d'arrivo percorrendo insieme solo brevi tratti del cammino nella consapevolezza che *«les usagers ne sont pas intéressés par notre profession. C'est l'information qui les intéresse»* (Boadas I Raset, 2010)[53]. Avranno le università la lungimiranza di raccogliere la sfida?

---

50 Cfr. VANNEVAR BUSH, *As We May Think*, in «Atlantic Magazine», luglio 1945. <http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>.

51 Cfr. WILLIAM ALLEN, THOMAS R.H. THOMPSON, *A Narrative of the expedition to the River Niger in 1841*, Londra, R. Bentley, 1848.

52 Nel 1965, Ted Nelson, propone la prima riflessione teorica organica sul concetto di ipertesto, e introduce i due termini *hypertext* e *hypermedia*. L'idea di Nelson è quella di realizzare «*una rete mondiale che possa essere utilizzata da centinaia di milioni di utenti simultaneamente, costituita dall'insieme degli scritti, delle immagini, dei dati conservati in tutto il mondo*». Egli battezza *Xanadu* il sistema che dovrebbe gestire la rete, permettendo di utilizzare la ragnatela di rimandi ipertestuali per reperire in maniera facile ogni documento.
<http://www.sfc.keio.ac.jp/~ted/>.

53 BOADAS I RASET, J., *op. cit.*, p. 105.

Bibliografia

ALBERANI, V., POLTRONIERI, E., *La Documentazione rispetto alle altre discipline dell'Informazione*, in «AIDA Informazioni», a. 21, n. 3, 2003, pp. 19-47

ALLEN, W., THOMPSON, T.R.H., *A Narrative of the expedition to the River Niger in 1841*, Londra, R. Bentley, 1848

ALMIND, T. C., INGWERSEN, P. *Infometric analyses on the World Wide Web: Methodological approaches to 'webometrics'*, in «Journal of Documentation», a. 53, n. 4, 1997, pp. 404-426

ARCHIVIO CENTRALE DELLO STATO (ACS), C.N.R., Centro Nazionale di Documentazione Scientifica (CNT), busta 943, fasc. 2

ARCHIVIO CENTRALE DELLO STATO (ACS), C.N.R., Centro Nazionale di Documentazione Scientifica (CNT), Busta 943A bis, fasc. 4

BALDAZZI, A., *Le scienze dell'informazione e le teorie della transizione*, in «AIDAInformazioni», a.20, n. 1, gennaio-marzo 2002, pp. 25-30

BARABASI, A.L., *Lampi*, Frediani S. (traduzione di), Torino, Einaudi, 2011

BASILI C., (Ed.), *Information Literacy in Europe: a first insight into the state of the art of information literacy in the european union*, Roma, 2003

BASILI, C., *Tappe salienti della Documentazione nel periodo 1983-2003 nella rassegna dei convegni AIDA, ASIS e FID*, in «AIDA Informazioni», a.21, n. 3, 2003, pp. 105-117

BIANCHINI, C., *Library Linked Data e il Futuro delle Biblioteche*, 2012. <https://sites.google.com/site/homepagecarlobianchini/Ricerca/library-linked-data-e-il-futuro-delle-biblioteche>

BISOGNO, P., *Il Futuro della Memoria – elementi per una teoria della Documentazione*, Milano, Franco Angeli, 1994

BOADAS I RASET, J., *De Quoi les citoyens ont-ils-besoin? S'adapter ou disparaître!*, in «Comma», vol. 1, 2010, pp. 103-108

BRIET, S., *Qu'Est-Ce Que la Documentation*, Parigi, Edit, 1951

BUSH, V., *As We May Think*, in «Atlantic Magazine», luglio 1945 <http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>

CASSESE, L., *Intorno al concetto di materiale Archivistico e materiale Bibliografico*, in «Notizie degli Archivi di Stato», a. 9, 1949

COMBA, V., *Esperienze e prospettive di formazione professionale nel settore privato per i documentalisti*, in Informazione e Documentazione: temi trasversali di formazione, Roma, 1992, pp. 57-58

CONSIGLIO NAZIONALE DELLE RICERCHE, *Progetto finalizzato sull'organizzazione ed il funzionamento della pubblica amministrazione – Relazione finale sull'attività svolta e sui risultati conseguiti*, rapporto interno s.d., 1995-96

Decreto del Presidente della Repubblica 20 ottobre 1998, n. 428, *Regolamento recante norme per la gestione del protocollo informatico da parte delle amministrazioni pubbliche*, in Gazzetta Ufficiale del 14 dicembre 1998, n. 291

Decreto del Presidente della Repubblica 28 dicembre 2000, n. 445, *Testo unico delle disposizioni legislative e regolamentari in materia di documentazione amministrativa*, in Gazzetta Ufficiale del 20 febbraio 2001, n. 42, Supplemento Ordinario n. 30

Decreto Legislativo 27 ottobre 2009, n. 150, *Attuazione della legge 4 marzo 2009, n. 15, in materia di ottimizzazione della produttività del lavoro pubblico e di efficienza e trasparenza delle pubbliche amministrazioni*, in Gazzetta Ufficiale del 31 ottobre 2009, n. 254, Supplemento Ordinario n. 197

Decreto legislativo 7 marzo 2005, n. 82, *Codice dell'amministrazione digitale*, in Gazzetta Ufficiale del 16 maggio 2005, n. 112, Supplemento Ordinario n. 93

Deliberazione CNIPA 19 febbraio 2004, n. 11/2004, *Regole tecniche per la riproduzione e conservazione di documenti su supporto ottico idoneo a garantire la conformità dei documenti agli originali*, in Gazzetta Ufficiale del 9 marzo 2004, n. 57

Di Donato, F., *La Scienza e la Rete*, Firenze, University Press, 2009

DIGITPA, Circolare 29 dicembre 2011, n. 59, *Modalità per presentare la domanda di accreditamento da parte dei soggetti pubblici e privati che svolgono attività di conservazione dei documenti informatici di cui all'articolo 44-bis, comma 1, del decreto legislativo 7 marzo 2005, n. 82***,** in Gazzetta Ufficiale del 8 febbraio 2012, n. 32

Drucker, P., *Post-Capitalist Society*, New York, Harper Business, 1993

Gli Organismi italiani di Documentazione, in La Documentazione in Italia, Roma, C.N.R., 1952, pp. 183-204

Granovetter, M.S., *The Strength of Weak Ties*, in «American Journal of Sociology», vol. 78, n. 6, maggio 1973, pp. 1360-1380

Guarasci, R., *Documenti, Dati, Viste Documentali*, in «e-Health care», n. 17, 2012, pp. 81-88

Guarasci, R., *La memoria della scienza: l'Archivio Tecnico Italiano e Il Centro Nazionale di Documentazione Scientifica*, in Archivi Privati, Guarasci R., Pasceri E. (a cura di), Roma, C.N.R., 2011, pp. 195-219

Guarasci, R., *Libri, Documenti e altre storie. L'insegnamento della Documentazione nelle università italiane*, in «AIDAInformazioni», a. 21, n. 3, luglio-settembre 2003, pp. 47-59

Guerrini, M., Possemato, T., *Linked Data: Un nuovo alfabeto del web semantico*, in «Biblioteche Oggi», aprile 2012, pp. 7-15

GUINCHAT, C., MENOU, M., *Introduction générale aux sciences et techniques de l'information et de la documentation*, Parigi, 1990

HIBBITS, B.J., *Oggi come ieri. Scettici, scribi e la fine delle riviste giuridiche*, in «Informatica e Diritto», vol. 9, n. 2, 1977

Legge 17 dicembre 2012, n. 221, *Conversione in legge, con modificazioni, del decreto-legge 18 ottobre 2012, n. 179, recante ulteriori misure urgenti per la crescita del Paese*, in Gazzetta Ufficiale del 18 dicembre 2012, n. 294, Supplemento Ordinario n. 208

Legge 7 Agosto 1990 n. 241, *Nuove norme in materia di procedimenti amministrativi e di diritto di accesso ai documenti amministrativi*, in Gazzetta Ufficiale del 18 agosto 1990, n. 192

MAIOCCHI, R., *Scienza e Fascismo*, Roma, Carocci, 2004

MEYRIAT, J., *La Documentazione: elementi per un riesame*, in La Documentazione in Italia, Paci A.M. (a cura di), Milano, Franco Angeli, 1996

NONAKA, I., TAKEUCHI, H., *The Knowledge Creating Company*, Stella G. (traduzione di), Milano, Guerini, 2009

OMODEI, A., *La Documentazione e l'industria*, in La Documentazione in Italia, Roma, C.N.R., 1952

PACI M.A. (a cura di), *La Documentazione nelle Università Italiane*, Roma, 1989

RAYWARD, W.B., *The History and historiography of information science: some reflections*, in «Information Processing and Management», vol. 31, n. 1, 1996, pp. 3-17

RAYWARD, W.B., *The origin of information science and the International Institute of Bibliography/International Federation for Information and Documentation (FID)*, in «Journal of the American Society for Information Science», vol. 48, n. 4, 1997, pp. 289-300

Regio Decreto 25 gennaio 1900 n. 35, *Approvazione del regolamento per gli Uffici di registratura e di archivio delle Amministrazioni centrali*, in Gazzetta Ufficiale del 22 febbraio 1900, n. 44

Sitografia

< http://cp2011.istat.it/>

<http://www.istat.it/it/archivio/18841>

<http://affordance.typepad.com/mon_weblog/2011/05/dsk-le-temps-reel-documente.html>

<http://museumpublicity.com/2012/03/14/mundaneum-and-google-announce-partnership>

<http://www.sfc.keio.ac.jp/~ted/>

<http://europa.eu/rapid/press-release_IP-12-1389_it.htm>

Scienze del documento – Scienze dell'informazione*

MADJID IHADJADENE** - LAURENCE FAVIER***

Introduzione

Le sfide e le opportunità scientifiche, politiche, economiche e culturali che la rivoluzione digitale comporta sono notevoli. I cambiamenti verificatisi hanno confuso i punti di riferimento nel campo delle scienze dell'informazione e della comunicazione tanto sul piano delle professioni che su quello della ricerca. I professionisti dell'informazione si trovano di fronte ad una nuova *ecologia del sapere*[1], che li pone in concorrenza con nuovi attori, ma che al tempo stesso offre loro nuove opportunità (Miège, 2004)[2]. La questione delle professioni e dello zoccolo delle competenze emerge nuovamente. Paradossalmente, la rilevanza assunta dal digitale non ha permesso, fino a questo momento, il consolidamento di tale campo professionale e di ricerca.

Le professioni dell'informazione sono molteplici. Esse ri-

* Traduzione dal francese all'italiano di Antonietta Folino. Il testo originale segue la traduzione.

** Laboratoire Index-paragraphe, Université de Paris 8.

*** Laboratoire Geriico, Université de Lille 3.

1 Per *ecologia del sapere* si intende l'insieme delle istituzioni che permettono l'accesso, la produzione e l'utilizzo delle conoscenze per finalità di apprendimento e innovazione [*N.d.T.*].

2 Cfr. BERNARD MIÈGE, *L'information – communication, objet de connaissance*, Bruxelles, De Boeck & INA, 2004.

guardano tanto la documentazione e la biblioteconomia, quanto la gestione degli archivi. Storicamente (Accart, 2000)[3], l'insegnamento della documentazione e della biblioteconomia è ripartito tra l'offerta universitaria e quella delle *grandes écoles et instituts*[4] (l'ENSB[5] diventato ENSSIB[6]; l'*Institut National des Techniques Documentaires* et l'*Ecole des bibliothécaires-documentalistes* legata all'istituto cattolico di Parigi). Le *unità di formazione e ricerca* (UFR)[7] rilasciano formazioni il cui livello è pari a quello del *master*[8].

La formazione orientata alla ricerca è assicurata dalle scuole dottorali. A questo panorama è necessario aggiungere le formazioni brevi erogate dagli *Instituts Universitaires de Technologies*[9] (due anni post-maturità) e gli *Instituts Universitaires Pro-*

---

3 Cfr. JEAN-PHILIPPE ACCART, *Bibliothécaire, documentaliste: même métier?*, in «Bulletin des bibliothèques de France», vol. 45, n. 1, 2000, pp. 88-93.

4 In Francia, le *grandes écoles* sono istituti di istruzione superiore di livello universitario (5 anni post-maturità), di grande prestigio accademico e con un sistema di accesso molto selettivo [*N.d.T.*].

5 *École Nationale Supérieure de Bibliothécaires*.

6 *École Nationale Supérieure des Sciences de l'Information et des Bibliothèques*.

7 Unità costituite da dipartimenti di formazione e laboratori di ricerca [*N.d.T.*].

8 Il *Master* corrisponde alla Laurea Magistrale nel sistema universitario italiano [*N.d.T.*].

9 Gli *Instituts universitaires de technologies (IUT)* rilasciano, al termine di un biennio, il *Diplôme Universitaire de Technologie* (DUT) e offrono una formazione professionalizzante nei settori terziario e tecnologico. Il loro obiettivo principale è quello di preparare gli studenti ad entrare direttamente nel mondo del lavoro, rendendoli autonomi e capaci di prendere iniziative. Al termine della formazione, tuttavia, è possibile anche accedere a percorsi universitari, quali una *licence professionnelle* (percorso di studi orientato all'inserimento nella vita professionale, di durata annuale e di livello corrispondente a tre anni post-maturità <http://www.etudinfo.com/diplome/licence-professionnelle/>) o un'*école d'ingénieur* (offrono una soli-

fessionnalisés[10] (tre anni post-maturità), senza dimenticare i corsi preparatori ai percorsi formativi di *professeur-documentaliste*[11]. Questi attori dell'informazione afferiscono a diverse associazioni professionali (ADBS[12], ABF[13], Fabden[14], ADBU[15], AAF[16], ecc.).

---

da formazione teorica nelle materie scientifiche e una significativa esperienza pratica relativa al mestiere dell'ingegnere. Sono sottoposte al controllo della *Commission Nationale des Titres d'Ingénieurs* – CTI che valida i diplomi rilasciati agli studenti <http://www.campusfrance.org/fr/page/les-ecoles-d%E2%80%99ingenieur>) o *de commerce* (orientate alla formazione di manager e di dirigenti d'azienda <http://www.lesmetiers.net/orientation/p1_325333/bien-choisir-son-ecole-de-commerce>). <http://ressources.campusfrance.org/catalogues_recherche/diplomes/fr/iut_fr.pdf> [*N.d.T.*].

10 Gli *Instituts Universitaires Professionnalisés (IUP)* sono orientati verso formazioni prevalentemente professionalizzanti, infatti almeno un terzo della sua durata deve essere svolto in contesti lavorativi e deve prevedere la presenza di professionisti che esercitano la propria attività in ambienti legati alla specializzazione dell'istituto. <http://www.label-iup.org/> [*N.d.T.*].

11 I *professeur-documentaliste* hanno la responsabilità, nei *collèges* (corrisponde alle scuole secondarie inferiori nel sistema di istruzione italiano. Ha una durata di quattro anni ed è rivolto a ragazzi di età compresa tra gli 11 e i 15 anni) o nei licei del centro di documentazione e di informazione (CDI). Essi formano gli studenti alla gestione e alla cultura dell'informazione e contribuiscono a far loro acquisire le competenze relative alle Tecnologie dell'Informazione e della Comunicazione (TIC), quali saper informarsi e documentarsi e adottare un atteggiamento responsabile nell'utilizzo di internet. <http://www.iufm.fr/devenir-ens/choisir/documentaliste.html> [*N.d.T.*].

12 *Association des professionnels de l'information et de la documentation.* <http://www.adbs.fr/>.

13 *Association des bibliothécaires français. <http://www.abf.asso.fr/>*.

14 *Fédération des associations des enseignants- documentalistes de l' Éducation Nationale*. <http://www.fadben.asso.fr>.

15 *Association des directeurs & personnels de direction des bibliothèques universitaires et de la documentation.* <http://abdu.fr>.

16 *Association des archivistes français.* <http://www.archivistes.org/>.

Nel dominio degli archivi, l'*Ecole nationale des chartes*[17], creata nel 1821, garantisce la formazione dei responsabili della conservazione del patrimonio nazionale. Dall'inizio degli anni '80, la formazione degli archivisti è stata diversificata. Le università propongono più di 15 percorsi formativi di livello *master* nell'ambito delle professioni degli archivi e del patrimonio (Defrance, 2002)[18]. Tali percorsi sono spesso erogati nei dipartimenti di Storia o di Scienze dell'Informazione e della Comunicazione (SIC). Si tratta di due concezioni, che riteniamo complementari, del ruolo degli archivisti e della loro organizzazione. Per alcuni, l'archivista è prima di tutto un custode della memoria storica, per cui la conoscenza della storia è imprescindibile. La seconda visione tende, invece, ad accostare la formazione degli archivisti a quella dei bibliotecari-documentalisti (Favier, 1993)[19]. Gli studenti, al termine di questi percorsi formativi, partecipano ai concorsi indetti dalle amministrazioni pubbliche locali o si inseriscono nel settore privato (Defrance, 2002)[20].

Sebbene i tre settori professionali si appoggino su saperi teorici e competenze equivalenti, (Palermiti, Polity, 2002)[21] dimostrano come le professioni dell'informazione siano in realtà scisse e come manchi in Francia, fin dai tempi del dopoguerra, una visione globale e unificata di queste stesse professioni. Essi te-

17 <http://www.enc.sorbonne.fr/>.

18 Cfr. JEAN-PIERRE DEFRANCE, *La formation archivistique en France: l'exemple du Bureau des métiers et de la formation de la Direction des Archives de France*, in «Archives» vol. 34 n. 1-2, 2002-2003, pp. 81-99.

19 Cfr. Favier J. (a cura di), *La pratique archivistique française*, Parigi, Archives nationales, 1993.

20 DEFRANCE, J-P., *op.cit.*

21 Cfr. ROSALBA PALERMITI, YOLLA POLITY, *Dynamiques de l'institutionnalisation sociale et cognitive des sciences de l'information en France*, in Les origines des sciences de l'information et de la communication en France, regards croisés, Boure R. (a cura di), Lille, Presses universitaires du Septentrion, 2002, pp. 95-123.

stimoniano la quasi-assenza del mondo delle biblioteche e degli archivi nel settore della ricerca nelle scienze dell'informazione. Queste professioni, secondo gli autori:

> *[…] se distinguent par des discours et des images stéréotypées qui reflètent davantage une appartenance aux lieux d'exercice qu'à celle d'une même famille professionnelle [...]. Les raisons de ces divergences et de ces oppositions s'expliquent historiquement. Elles sont liées particulièrement à des politiques publiques incohérentes, tant au niveau des statuts des personnels qu'au niveau d'un système de formation centralisé, fermé sur lui-même et peu progressif, créant notamment une relative confusion chez les employeurs et une absence de clarté dans la carte des formations*[22].

Sebbene le professioni di archivista, bibliotecario e documentalista si siano differenziate notevolmente, la riflessione sul loro avvenire deve necessariamente essere comune (Melot, 2005)[23]. Quest'ultimo stima che in futuro tali professioni potrebbero essere chiamate ad incontrarsi. Un simile avvicinamento è favorito da un contesto che non rimette in discussione le specificità de-

---

22 *Ivi*, p. 113.
[…] si differenziano per i discorsi e le immagini stereotipate che riflettono più un'appartenenza ai luoghi d'esercizio che ad una stessa famiglia professionale […]. Le ragioni di queste divergenze ed opposizioni vanno lette in chiave storica. Esse sono legate in particolare a delle politiche pubbliche incoerenti, tanto a livello degli statuti del personale, quanto a livello di un sistema formativo centralizzato, chiuso su se stesso e poco evolutivo, creando una relativa confusione tra gli impiegati e una mancanza di chiarezza nell'offerta formativa.

23 MICHEL MELOT, *Archivistes, documentalistes, bibliothécaires: Compétences, missions et intérêts communs*, in «Bulletin des Bibliothèques de France», t. 50, n. 5, Parigi, 2005, p. 9.

gli utenti (Wiegandt, 2005)[24]. La maggior parte dei professionisti dell'informazione svolge, secondo (Fondin, Rouault, 1998)[25] sia funzioni di intermediazione documentaria o culturale tra gli utenti e le risorse informative, sia funzioni di conservazione del patrimonio documentale. Queste professioni si sono evolute attraverso l'integrazione di ulteriori competenze legate all'ambito imprenditoriale, in risposta all'emergere di nuove esigenze da parte degli utenti e alla generalizzazione delle tecnologie dell'informazione e della comunicazione (TIC). Gli autori includono in questa categoria le professioni che operano a monte del trattamento dell'informazione (gestione editoriale, gestione di siti web, gestione di basi di dati documentali) o a valle del suo utilizzo, estendendo quelle della documentazione alle professioni della veglia[26] e dell'*intelligence économique*[27], così come della

[24] CAROLINE WIEGANDT, *Bibliothécaires et documentalistes: deux métiers qui se rapprochent*, in «Bulletin des Bibliothèques de France», t. 50, n. 5, Parigi, 2005, p. 18.

[25] HUBERT FONDIN, JACQUES ROUAULT, *L'information: l'arlésienne de l'interdiscipline des sciences de l'information et de la communication*, documento dattiloscritto, 1998, p. 2.

[26] *«Veille informationnelle: Processus continu et dynamique faisant l'objet d'une mise à disposition personnalisée et périodique de données ou d'information/renseignement, traitées selon une finalité propre au destinataire […]».*
SERGE CACALY, YVES-FRANÇOIS LE COADIC, PAUL-DOMINIQUE POMART, ERIC SUTTER, *Dictionnaire de l'Information*, ed.2, Parigi, Armand Colin, 2006, p. 247.
Veglia informativa: Processo continuativo e dinamico finalizzato alla messa a disposizione personalizzata e periodica di dati o informazioni, trattati in base a obiettivi propri del destinatario […].

[27] *«Intelligence économique: Ensemble des actions coordonnées de recherche, de traitement et de distribution en vue de son exploitation, de l'information utile aux acteurs économiques […]».*
CACALY, S., et alii, *op. cit.*, p. 127.
Intelligenza economica: Insieme di azioni coordinate di ricerca, tratta-

gestione dei contenuti aziendali e del knowledge management (*Comité National d'Evaluation*, 1993)[28].

L'evoluzione[29] delle professioni dell'informazione è costantemente esaminata attraverso studi condotti dall'ADBS, dall'ANPE (*Agence Nationale pour l'Emploi* diventata poi *Pôle Emploi*) o da società di consulenza (*Serda*, *Cepid*, *Histen Riller*). Da questi studi emerge, peraltro, una incertezza sulle possibilità occupazionali dei professionisti dell'informazione. (Lebigre, 2011)[30], citando alcune statistiche dell'INSEE (*Institut national de la statistique et des études économiques*) e gli studi effettuati dalla società *Serda*, afferma che i professionisti dell'informazione che ad oggi esercitano in Francia sono tra le 35.000 e le 80.000 unità. Un terzo di essi lavora nel settore privato e si tratta, generalmente, di documentalisti. Tra i punti di convergenza di tali lavori, (Lebigre, 2011)[31] e (Stiller, 2001)[32] citano la diluizione delle funzioni informazione-documentazione all'interno delle organizzazioni, determinata dal fatto che un terzo degli intervistati lavora in maniera autonoma, al di fuori di centri di documentazione. Quasi un terzo delle aziende non dispone più di un servizio di documentazione (Lebigre, 2011)[33] e nei casi in cui è presente,

---

mento e distribuzione dell'informazione utile ad attori economici in vista del suo struttamento […].

28 Cfr. COMITÉ NATIONAL D'EVALUATION (CNE), *Les sciences de l'information et de la communication*, Rapporto di valutazione, 1993, p. 13.

29 Per una panoramica delle qualifiche professionali in Europa Cfr. BARRY MAHON, *The disparity in professional qualifications and progress in information handling: a European perspective*, in «Journal of information science», vol. 34, n. 4, 2008, pp. 567-575.

30 LOÏC LEBIGRE, *Communautés de l'info-doc: un équilibre subtil*, in «Documentaliste-Sciences de l'Information», vol. 48, n. 2, 2011, p. 24.

31 LEBIGRE, L., *op.cit.*, p. 27.

32 HENRI STILLER, *La fonction Information-Documentation dans les grandes entreprises: une enquête*, in «Documentaliste-Sciences de l'Information», vol. 38, n. 3-4, 2001, p. 223.

33 LEBIGRE, L., *op.cit.*, p. 23.

si osserva un calo nelle dimensioni delle strutture, pur sapendo che, in termini di settori d'attività, la situazione è composita e variegata. Così secondo (Stiller, 2001):

> *Dans certaines entreprises, la fonction I-D est parfaitement reconnue: en plus des fonctions classiques, on lui confie des missions de veille, on l'associe au processus de gestion des connaissances, on la fait participer au déploiement des NTIC. Dans d'autres, à l'inverse, cette fonction est diluée au sein des services, où aucune structure propre ne la prend en charge*[34].

Le evoluzioni economiche (crisi, globalizzazione), insieme alla rivoluzione digitale, sono, secondo (Michel, 2011)[35] e (Michel, 2003)[36], all'origine della messa in discussione delle modalità di intervento professionale della documentazione. Secondo l'autore è il modello di centralizzazione in uso nei centri di documentazione e sviluppato nel periodo successivo alla prima fase di informatizzazione delle pratiche documentali ad essere reso obsoleto dai cambiamenti attuali. La catena documentale generata da questo modello risulta oggi inadatta. Tali studi mostrano anche un'evoluzione delle attività nell'ambito delle funzioni documentali: quelle legate alla catalogazione e alla raccolta di informazione decrescono in favore dell'analisi e della valorizza-

---

34 STILLER, H., *op.cit.*, p. 223.
In alcune aziende, la funzione I-D (informazione-documentazione) è perfettamente riconosciuta: oltre alle missioni classiche le si affidano anche compiti di veglia, le si associa il processo di gestione delle conoscenze, la si coinvolge nel dispiegamento delle NTIC (Nuove tecnologie di informazione e di comunicazione).

35 Cfr. JEAN MICHEL, *Crise économique, crise de la profession... constats et perspectives d'évolution*, in «Documentaliste-Sciences de l'Information», vol. 49, n. 3, 2011, pp. 4-7.

36 Cfr. JEAN MICHEL, *Les documentalistes: l'urgence d'une reconnaissance sociale*, in «Hermès», n. 35, 2003, pp. 185-193.

zione dei documenti (Lebigre, 2011)[37] (Stiller, 2001)[38]. Questa diversificazione nelle professioni dell'informazione rappresenta un'opportunità per l'inserimento professionale. Alcune funzioni editoriali, di comunicazione relativa al web, di gestione strategica dell'informazione completano la panoplia di professioni tradizionali (Lebigre, 2011)[39]. La sfida per i professionisti dell'informazione consiste nel dimostrare che le proprie competenze restano applicabili al di là del centro di documentazione tradizionale (Stiller, 2011)[40]. Tra le prospettive identificate a seguito di tali studi, ne emergono due in particolare: la prima consiste nel padroneggiare strumenti, metodi e organizzazioni delle professioni del web, mentre la seconda nella comprensione dei bisogni propri alle professioni aziendali. I professionisti dell'informazione, infatti, associano a competenze specifiche nelle scienze dell'informazione una conoscenza, spesso approfondita, di un settore d'attività.

Dopo aver presentato la situazione delle professioni dell'informazione in Francia, ci si soffermerà, nel seguito, sulla condizione dell'attività di ricerca nella scienza dell'informazione, mettendo in discussione i fondamenti teorici e accademici di questa giovane disciplina.

1. Dalla documentazione alle scienze dell'informazione

1.1 *Fondamenti teorici*

Sebbene l'idea innovativa di una *scienza dell'informazione* si debba *all'Union Française des Organismes de Documentation*

37 LEBIGRE, L., *op.cit.*, p. 27.
38 STILLER, H., *op.cit.*, p. 223.
39 LEBIGRE, L., *op.cit.*, p. 23.
40 Cfr. STILLER, H., *op.cit.*

(creata nel 1932) fin dal primo dopoguerra (Fayet-Scribe, 2000)[41], bisognerà attendere gli anni '70 del secolo scorso affinché essa acquisti *un'esistenza accademica* in Francia nell'ambito del più ampio dominio delle *scienze dell'informazione e della comunicazione*. Solo negli Stati Uniti la scienza dell'informazione e delle biblioteche (*Library and Information Science*) riuscirà a svilupparsi indipendentemente dall'università, anche se, negli anni '90 essa vedrà decrescere la propria importanza a vantaggio di altre denominazioni che eludono qualsiasi riferimento alla biblioteca (*school of information, school of information sciences, Ischools,* ecc.).

I fondamenti teorici delle scienze dell'informazione sono stati elaborati all'interno di un progetto politico. Infatti, il concetto di *documentazione* e il significato moderno di *informazione* si fondano sulla costruzione della Società delle Nazioni (SDN) e, all'interno di questa, sulla cooperazione intellettuale internazionale[42], grazie al notevole operato dei due avvocati belgi, Paul Otlet e Henri Lafontaine. Se la Società delle Nazioni e il pacifismo di questo periodo ebbero fine a causa di un fallimento politico e della seconda guerra mondiale, il lavoro di Paul Otlet continua a far riflettere i propri contemporanei per la sua modernità: la separazione del contenuto e del supporto, l'idea di ipertesto e la prefigurazione del web, la nozione di *linguaggio documentale*, la mediazione documentale. I primi tre temi sono compresi

41 SYLVIE FAYET-SCRIBE, *Histoire de la documentation en France. Culture, science et technologie de l'information: 1895-1937*, Parigi, CNRS Editions, 2000, p. 52.

42 Nel 1921 viene creata una Commissione internazionale di cooperazione intellettuale da parte dell'assemblea della Società delle Nazioni e, nel 1926, l'Istituto internazionale di cooperazione intellettuale che nasce a Parigi la rende permanente. Questa commissione è considerata come l'inizio di ciò che diventerà l'UNESCO dopo la guerra, mentre la SDN diverrà l'Organizzazione delle Nazioni Unite.

nel concetto moderno di *documentazione*, mentre l'ultimo si pone nell'ottica di un nuovo rapporto con il sapere che passa attraverso l'*informazione* così come essa viene definita in questo contesto *pre-informatico*.

Vedremo come i fondamenti teorici delle scienze dell'informazione e della documentazione si siano costituiti a partire dalla doppia eredità del Repertorio Bibliografico Universale di Paul Otlet e di un legame con lo studio della comunicazione specifico di una concezione universitaria francese del dominio.

1.1.1 *L'eredità del Repertorio Bibliografico Universale*

Il concetto di documentazione

Il progetto di Repertorio Bibliografico Universale (RBU) che forgiarono i due avvocati è un rinnovamento dell'enciclopedismo del secolo dei Lumi, nella misura in cui esso mira a raccogliere la *science vivante*[43] attraverso la creazione di un nuovo sistema nel quale non si tratta più di sintetizzare il sapere, ma di dare accesso alle pubblicazioni originali che l'hanno formulato. In tal modo *«L'ensemble de tous les écrits pourra, en un certain sens, être considéré comme formant un seul grand livre, un livre aux proportions formidables, aux chapitres en nombre quasi illimité»* (Otlet, 1903)[44]. La documentazione si colloca a metà strada tra l'enciclopedia e la bibliografia: è al tempo stesso il *grand livre* e l'insieme dei lavori di cui si compone. Ma essa non si riduce né

[43] In merito a questa espressione, cfr. FRANÇOISE LEVIE, *L'homme qui voulait classer le monde*, Bruxelles, Les impressions nouvelles, 2006.

[44] PAUL OTLET, *Les sciences bibliographiques et la documentation*, in «Bulletin de l'Institut International de Bibliographie», n. 8, Bruxelles, 1903, p. 136.
L'insieme di tutti gli scritti potrà, in un certo senso, essere considerato come parte di un solo grande libro, un libro dalle proporzioni enormi, con un numero di capitoli quasi illimitato.

all'uno né all'altro, poiché non è né un'impresa di divulgazione scientifica (di riformulazione dei lavori), né un inventario delle pubblicazioni esistenti (la bibliografia). Essa è *«la carte immense des domaines du savoir, avec tout le complexe des divisions et des subdivisions de leurs territoires»*, grazie alla quale

> *nous pourrions localiser tout naturellement chacun des travaux dans quelqu'une des circonscriptions. Nous les verrions s'y rattacher aux travaux similaires pour compléter ce qui furent écrits antérieurement, et à leur tour servir de lien entre ces données du passé et les progrès de l'avenir* (Otlet, 1903)[45].

Il RBU rappresenta un'innovazione sotto molti punti di vista. Sul piano materiale, introduce la cassettiera per la conservazione delle schede mobili in uso fino all'informatizzazione dei cataloghi. Questa modalità di concepire la documentazione si contrappone alla presentazione della bibliografia sottoforma di volumi cumulativi. Sul piano organizzativo, il RBU impone la standardizzazione delle attrezzature impiegate. La documentazione, anche in questo caso, si discosta dalla tradizione bibliografica, che rimane un'impresa individuale: essa applica i metodi industriali del tempo che danno vita alle prime agenzie di normalizzazione. Otlet e Lafontaine, d'altronde, erano ammiratori di Ford e del fordismo. La cooperazione internazionale, che si incarna nell'Unione delle Associazioni Internazionali, si ottiene grazie a questo forzo di normalizzazione che resterà una preoccupazione impor-

[45] *Ivi*, p. 135.
La mappa immensa dei domini del sapere, con tutto il complesso delle divisioni e delle suddivisioni dei loro territori.
Potremmo localizzare con tutta semplicità ciascun lavoro in qualsiasi circoscrizione. Li vedremmo riallacciarsi a lavori simili per completare quelli scritti precedentemente e, a loro volta, fungere da legami tra questi dati del passato e i progressi del futuro.

tante della documentazione. Sul piano funzionale, il RBU è un sistema di ricerca dell'informazione e non solo una classificazione ragionata dei libri nelle biblioteche. In tal senso esso prefigura il web, come fatto notare da numerosi autori e personalità: da Boyd Rayward[46] a Vinton Cerf (inventore del protocollo TCP/IP) e Thierry Geerts (direttore di Google Belgio)[47]. Le relazioni tra le conoscenze vengono stabilite attraverso la Classificazione Decimale Universale (CDU), la quale non è considerata solo una tecnica di classificazione, ma una vera e propria lingua universale: *«Créer une classification synthétique avec notation concise des idées c'est doter l'esprit d'une véritable langue écrite universelle capable d'agir puissamment sur la forme elle-même de la Pensée»* (Otlet, 1935)[48]. Ciò rappresenta la principale differenza con la Classificazione Decimale Dewey (CDD), sebbene la CDU vi si ispiri direttamente. Quest'ultima, infatti, introduce un paradigma nuovo, malgrado la continuità con l'impresa di Dewey. Essa trasforma lo schema strettamente gerarchico ed enumerativo della CDD in una lingua capace di esprimere relazioni di altra natura tra i soggetti (in particolare l'introduzione del principio delle faccette tramite le tavole ausiliarie e la definizione di simboli per rappresentare le relazioni). La CDU diventa anche una sorta di *esperanto* nel quale ciascun soggetto deve poter essere codificato in relazione al sistema universale delle scienze.

---

46 Cfr. W. BOYD RAYWARD, *Visions of Xanadu: Paul Otlet (1868-1944) and Hypertext*, in «Journal of the American Society for Information Science», vol. 4, n. 4, 1994, pp. 235-250; W. BOYD RAYWARD, *Knowledge organization and a new world polity: the rise and fall of the ideas of Paul Otlet*, in «Transnational Associations», n. 1-2, 2003, pp. 4-15.

47 Cfr. JEAN-PIERRE STROOBANTS, *Le Web, une histoire belge*, in «Le Monde», 03/11/2012.

48 PAUL OTLET, *Monde. Essai d'universalisme*, Bruxelles, Editiones Mundaneum, 1935, p. 389.

Lo schema *Les étapes de l'organisation documentaire* (Figura 1) riassume questa nuova idée della documentazione.

Figura 1. ©Archives Mundaneum.

Quest'opera, destinata a raccogliere il sapere del mondo, necessita di nuove definizioni:

> *Nous entendons par le terme général information les données de toute nature, faits, théories nouvelles, qui, parvenus à l'intelligence humaine, constituent des notions, des éclaircissements, des directives pour la conduite et l'action; d'autre part, nous entendons par documentation l'ensemble des moyens propres à transmettre, à communiquer, à répandre les informations, (livres, périodiques, catalogues, textes et images, documents de toutes formes)* (Otlet, 1917)[49].

Creare una classificazione sintetica con una annotazione concisa delle idee significa dotare lo spirito di una vera e propria lingua scritta universale capace di agire notevolmente sulla forma stessa del Pensiero.

[49] PAUL OTLET, *L'information et la documentation au service de l'industrie,*

Un nuovo rapporto con il sapere: dal RBU al Mundaneum

Il RBU permette di immaginare l'aggiornamento perpetuo delle conoscenze, *l'effemerografia*[50], secondo il principio stabilito da Charles Limousin[51] nel 1900 (Fayet-Scribe, 2000)[52]. I *current contents*[53], l'analisi documentale, la *revue à decou-*

in Bulletin de la Société d'encouragement pour l'industrie nationale, Parigi, Renouard, 1917, p. 518.
Intendiamo con il termine generico *informazione* dati di qualsivoglia natura, fatti, teorie nuove, che acquisiti dall'intelletto umano, costituiscono nozioni, chiarimenti, direttive per il comportamento e l'azione; d'altra parte, con documentazione intendiamo l'insieme dei mezzi atti a trasmettere, comunicare e diffondere le informazioni (libri, periodici, cataloghi, testi e immagini, documenti di qualsiasi natura.

50 Calco del francese *éphémérographie*. Il significato di tale concetto si desume da quello di effemeride, al quale è strettamente legato. Le effemeridi, infatti, erano dei libri impiegati nell'antichità per la registrazione periodica degli atti del re. <http://www.treccani.it/vocabolario/effemeride/>.

51 Charles Limousin metteva in evidenza nel 1900 come fosse necessario redigere delle *éphémérographie* degli articoli contenuti nelle pubblicazioni speciali, scientifiche, letterarie e artistiche. Molti di questi articoli, infatti, non essendo raccolti in un volume a seguito della loro pubblicazione sulle riviste, diventano estremamente effimeri e rischiano di essere dimenticati, causando una notevole perdita di informazione. La ricerca all'interno delle riviste, dal canto suo, richiede enormi quantità di tempo e sforzi considerevoli. Secondo Limousin, quindi, dovrebbero esistere in tutte le biblioteche, delle schede sulle quali riportare una breve analisi di ciascun articolo e il rinvio alla pubblicazione, al volume, alla pagina, ecc. in maniera tale da facilitare la ricerca e il recupero da parte del lettore. Tali schede assolverebbero alla funzione di *éphémérographie*. Il *Bulletin des Sommaires*, ideato dallo stesso Limousin, ne rappresenta un esempio.
Cfr. M. Charles Limousin, *L'éphémerographie: Bibliographie des journaux et pubblications périodiques*, in «Bulletin de l'IIB», 1900, pp. 147-149.

52 Fayet-Scribe, S., *op. cit.*, pp. 96-97.

53 «*I Current contents […] sono bollettini che con periodicità molto fitta forniscono in formato cartaceo o elettronico le Table of contents (Toc), ovvero i sommari, talvolta arricchiti con abstract, di un certo numero di pe-*

per[54] diventano opportunità per accedere rapidamente all'informazione ed evitare la perdita di denaro. Il RBU e le associazioni internazionali che vi contribuiscono rispondono a questi bisogni, che sono propri della società industriale in pieno sviluppo. Ma, al di là delle operazioni di aggiornamento, emerge l'idea che la documentazione implichi una mediazione tra i lettori e i documenti. Tale idea è presente nei testi di Otlet relativi alla creazione di un *Office de Documentation industrielle* che attribuiscono ai suoi operatori la qualifica di *intermédiaires vivants entre le public et les documents*[55]:

> *L'Office est actif tandis que la bibliothèque est passive; il repose sur l'idée que les personnes auxquelles il est destiné doivent être amenées à agir dans une direction donnée et, pour cela, qu'elles doivent être incitées à la faire, que l'Office doit les aider, de toute manière, dans leur effort, pour se représenter les choses, non pas d'une manière quelconque mais le plus exactement possible. Pour cela, les agents de l'Office sont des in-*

riodici specializzati in un determinato settore».
<http://www.laterza.it/bibliotecheinrete/Cap09/Cap09_16.htm>.

54 Il concetto di *revue à découper* è introdotto in un articolo di Charles Didier (Cfr. CHARLES DIDIER, *La Revue à découper, note sur un mode plus rationnel de publier les articles de revue*, in «Bulletin de l'IIB», 1898, pp. 175-182.), il quale illustra le modalità con le quali organizza la propria documentazione personale : ripartizione in dossier (*articles à découper*) classificati in base alla classificazione decimale.
FAYET-SCRIBE, S., *op. cit*., pp. 96-97.
Questo concetto rientra nella medesima logica dell'éphémérographie, ovvero l'organizzazione degli articoli pubblicati in riviste per favorirne il recupero e la disseminazione. In LIMOUSIN, C., *op. cit*., infatti, si fa riferimento ad agenzie bibliografiche che di leggere le riviste, di partizionarle e di inviare agli abbonati gli articoli che trattano temi di loro interesse. Tali agenzie prendono infatti il nome di *Découpeurs*.

55 Intermediari viventi tra il pubblico e i documenti.

> *termédiaires vivants entre le public et les documents. Les questions leur sont posées sous la forme concrète des cas appliqués. Ne pouvant tout connaître ni tout retenir, ces agents doivent pouvoir se retourner vers des sources autorisées […], des livres, des dossiers, des répertoires contenant des données antérieurement élaborées […]* (Otlet, 1917)[56].

Sebbene la mediazione documentale sia rappresentata da nuovi mestieri, quali gli operatori ai quali oggi ci si riferirebbe con il termine *documentalisti*, o da motori di ricerca e da altri strumenti (annuari, segnalibri, portali…) noti ai nostri giorni, il regno dell'informazione e del documento implica una mediazione sempre più complessa per accedere alle risorse e per intuire i percorsi di ricerca degli utenti, ai quali Otlet si riferisce con il termine *pubblico*.

Piuttosto che dall'accesso all'informazione, il cuore del progetto della documentazione è rappresentato dall'istruzione. Otlet immagina l'integrazione della documentazione in un progetto più vasto, il *Mundaneum* o *Palais Mondial*[57], il quale dovrebbe riunire un centro di documentazione, una biblioteca, un centro di cultura scientifica, un museo e una biblioteca internazionale e

---

[56] OTLET, P., *op. cit.*, 1917, p. 52.
L'*Office* è attivo, mentre la biblioteca è passiva: esso si basa sull'idea che le persone alle quali si rivolge debbano essere indotte ad agire in una direzione data e, che, perciò, debbano essere incitate a seguirla, che l'*Office* debba aiutarle, in qualunque modo, nel loro sforzo di rappresentarsi le cose, non in maniera qualunque, ma il più esattamente possibile. Per tali ragioni, gli agenti dell'*Office* sono degli intermediari viventi tra il pubblico e i documenti. Le domande vengono loro poste secondo la forma concreta dei casi applicativi. Non potendo conoscere o ricordare tutto, tali agenti devono poter far riferimento a fonti autorizzate […], libri, dossier, repertori contenenti dati precedentemente elaborati […].

[57] Palazzo mondiale.

dovrebbe declinarsi in *Mundaneum* locali (che convergerebbero in un *Mundaneum* centrale) grazie alle associazioni internazionali:

> *De même que les États ont organisé la Société des nations et les Unions officielles des gouvernements qui en dépendent, de même les Associations internationales ont à se relier entre elles en une Union des Associations internationales et à faire du Mundaneum leur œuvre capitale.*
> …
> *Il est devenu nécessaire d'établir des liens entre ces organisations et d'en faire les partie d'une institution générale: a) en prenant pour base les quatre grandes institutions intellectuelles qui ont noms Bibliothèque, Musée, Université, Académie, Société scientifique [...]* (Otlet, 1935)[58].

La convergenza che si osserva oggi tra documentazione, biblioteche, archivi e musei, legata alla digitalizzazione delle risorse documentali, si avvicina a questa concezione premonitrice. Non potendo immaginare una rete virtuale come luogo di tale convergenza, egli pensa ad una città mondiale nella quale si dispiega la *rete mundaneum* illustrata nella figura di seguito riportata (Figura 2).

---

[58] Otlet, P., *op.cit.*, 1935, p. 449.
Così come gli Stati hanno organizzato la Società delle Nazioni e le *Unions Officielles des Gouvernements* che da essa dipendono, così le Associazioni internazionali devono riunirsi tra loro in una *Union des Associations internationales* e fare del *Mundaneum* la loro opera capitale.
È diventato necessario stabilire delle relazioni tra queste organizzazioni e trasformarle in parti di un'istituzione generale: a) assumendo come punto di partenza le quattro grandi istituzioni intellettuali denominate Biblioteca, Museo, Università, Accademia, Società scientifica […].

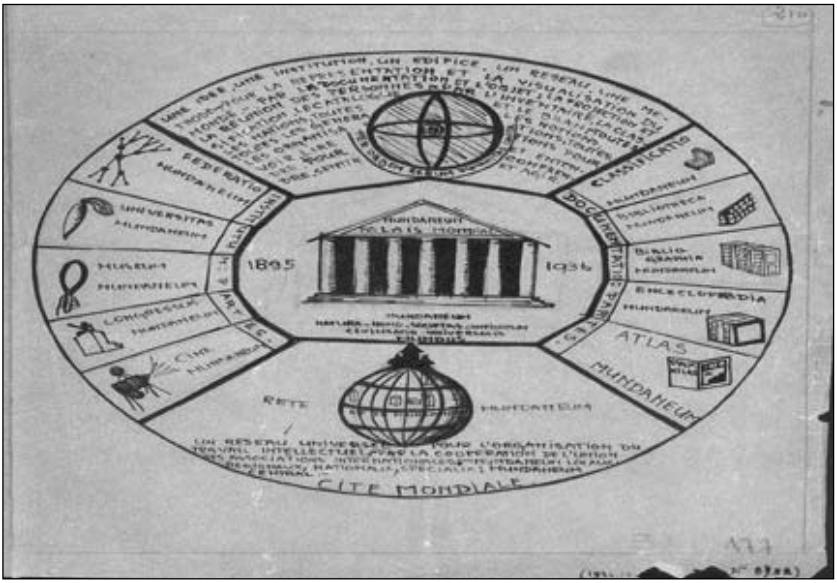


Figura 2. ©Archives Mundaneum.

Il RBU è, quindi, un elemento di un'utopia più complessa orientata verso l'istruzione lungo tutto l'arco della vita, tema che non può che considerarsi contemporaneo:

> *le Mundaneum repose sur le principe que l'éducation n'est pas simplement utilitaire. [...] L'éducation continue s'impose donc à chacun comme un devoir individuel et social de perfectionnement, développé et continué à tout âge* (Otlet, 1935)[59].

[59] *Ibidem*.
Il Mundaneum si basa sul principio che l'istruzione non è semplicemente utilitaria […]. L'istruzione continua, quindi, si impone a ciascuno come un dovere individuale e sociale di perfezionamento, sviluppato e perseguito ad ogni età.

Questa fu l'ambizione politica, al contempo pacifista e universalista, che diede vita ai fondamenti teorici delle scienze dell'informazione così come vengono oggi insegnate nelle università, anche se la rivoluzione cibernetica, la nascita dell'informatica e la generalizzazione della digitalizzazione sembrano esserne stati i principali fomentatori.

La nascita accademica delle scienze dell'informazione e della comunicazione (SIC) in Francia non ha rivendicato in maniera prioritaria questa eredità. Qualunque sia l'interpretazione che è possibile dare a questo accostamento di informazione e comunicazione, caratteristico del contesto francese, per la sua teorizzazione ci si riferisce piuttosto all'eredità del modello della teoria matematica della comunicazione di Shannon e Weaver (il modello e le sue critiche)[60].

1.1.2 *Dalla documentazione alle scienze dell'informazione e della comunicazione*

Secondo (Fondin, 2006)[61] la teoria matematica della comunicazione funge da referente teorico per le scienze dell'informazione (SI), non per necessità concettuale, ma per il principio di dare una parvenza scientifica alla teorizzazione dell'*attività documentale*:

---

60 Contrariamente a ciò che affermano numerosi autori. Si veda FIDELIA IBEKWÉ-SAN-JUAN, *The French Conception of Information Science: Une exception française?*, in «Journal of the American Society for Information Science and Technology», vol. 63, n. 9, 2012, p. 1693: «*Indeed, there is a widespread belief among members of the academic community in France that the Anglophone conception of information science is very different from theirs, in that it is rooted mainly in Shannon's mathematical theory of communication*». Si veda anche HUBERT FONDIN, *La science de l'information ou le poids de l'histoire*, in «Les Enjeux de l'information et de la communication», 2006, pp. 1-19. <http://w3.u-grenoble3.fr/les_enjeux/ 2005/Fondin/home.html>.

61 Cfr. FONDIN, H., *op. cit.*

> *Pour la quasi-totalité des chercheurs en SI, le référent théorique, parce qu'il faut un référent théorique si l'on veut faire "scientifique", est, implicitement et même très souvent explicitement, la théorie de l'information de Claude E. Shannon. L'activité documentaire est assimilée à un codage/décodage des signes d'un message qu'il faut réaliser dans les meilleures conditions techniques. […] On est là dans un modèle scientifique tel que le XX^{ème} siècle l'a forgé et imposé comme modèle à toutes les sciences, qu'elles soient exactes ou sociales, autrement dit quel que soit le type de phénomènes à étudier*[62].

La scienza dell'informazione anglosassone non ha avuto necessità di un referente teorico per trovare un modello di scientificità. Se essa ha riunito, negli anni '50, ingegneri e specialisti dell'organizzazione delle conoscenze (indicizzazione, linguaggi documentali) provenienti da differenti orizzonti culturali, è stato in fin dei conti per creare un dominio specifico: le scienze dell'informazione e delle biblioteche.

In Francia la definizione delle SIC avviene, secondo molti autori, in favore di una preponderanza della comunicazione, della quale la teoria matematica di Shannon si presenta spesso come il naturale fulcro, pur assumendo spesso una posizione di critica nei suoi riguardi. A proposito di questa posizione di preminenza

---

62 *Ivi*, p. 16.
Per la quasi totalità dei ricercatori in SI il riferimento teorico, dal momento che un riferimento teorico è necessario se si vuole essere *scientifici*, è, implicitamente, ma molto spesso anche esplicitamente, la teoria dell'informazione di Claude E. Shannon. L'attività documentale è assimilata a una codifica/decodifica dei segni contenuti in un messaggio, che bisogna eseguire nelle migliori condizioni tecniche […]. Ci si riferisce al modello scientifico che è stato forgiato e imposto dal XX secolo a tutte le scienze, siano esse esatte o sociali, in altre parole, a qualsiasi tipo di fenomeno da studiare.

(Courbières, 2010)[63] afferma che le *scienze della comunicazione* giocano il ruolo di *scienza-madre* anche per coloro che, come J. Meyriat, hanno determinato lo sviluppo delle SI in Francia. Così (Meyriat, 1993)[64] fa della *communicologie* la scienza a partire dalla quale si sviluppano le altre branche (*medialogie*, *informatologie*, *communicologie fonctionnelle*). La *communicologie* si definisce come lo studio de *«la communication sous tous ses aspects»*[65] e non include nessun campo specifico per le SI. La *documentologie* (oggetto della *medialogie* così come gli studi sulla stampa e la comunicazione di massa) si distingue dalla scienza del contenuto semantico della comunicazione (*informatologie*) e della comunicazione funzionale (funzionamento dei sistemi di comunicazione che assolvono a compiti particolari, tra cui la comunicazione scientifica e la pubblicità). In sintesi, ciò che rientra oggi nel campo delle scienze dell'informazione nei corsi universitari si trova qui scisso nelle diverse branche dello studio della comunicazione.

Un altro eminente fondatore delle SIC, Robert Escarpit, conferma questo primato della comunicazione. La sua opera, *L'Information et la Communication. Théorie générale*[66], copre un ampio spettro che si estende dalla teoria di Shannon e il *sogno ciberne-*

---

63 Cfr. CAROLINE COURBIÈRES, *Femme en contextes: la conception stéréotypée du féminin au travers du langage documentaire (1958-2008),* Mémoire pour l'habilitation à diriger des recherches en Sciences de l'Information et de la Communication, 2010, Università di Toulouse le Mirail.

64 Cfr. JEAN MEYRIAT, *Un siècle de documentation: la chose et le mot,* in «Documentaliste-Sciences de l'Information», vol. 30, n. 4-5, luglio-ottobre 1993, pp. 192-198.

65 La comunicazione in tutti i suoi aspetti.
MEYRIAT, J., *op. cit.*, citato da COURBIÈRES, C., *op. cit.*, p. 64.

66 Cfr. ROBERT ESCARPIT, *L'Information et la Communication. Théorie générale*, 6 ed., Parigi, Hachette, 1991.

tico[67] allo studio della cultura di massa, passando per l'analisi del linguaggio[68] e della documentazione. Due capitoli riguardano il campo delle scienze dell'informazione: *L'information et le document*, *Les problèmes documentaires*. La comunicazione è il fulcro della citata teoria di Shannon: essa è definita come *un cas particulier du transport*[69], quello dell'informazione. I limiti del modello termodinamico ed entropico emergono nel momento in cui essi vengono confrontati con le *vivant* e la *pensée:*

> *Il est donc probable que, dès qu'on passe du mécanique au vivant et plus encore du vivant à l'humain, un nouvel outil épistémologique devient nécessaire. Cet outil est sans doute la notion d'information. Nous pouvons hasarder l'hypothèse que cette notion est structurellement liée à celle d'entropie, mais qu'elle en diffère par le fait qu'elle prend en compte deux éléments laissés de côté par la science physique: la vie et surtout la pensée* (Escarpit, 1991)[70].

---

67 Si veda il capitolo *Le rêve cybernetique*.

68 Si vedano i capitoli *Langage et langages* e *La communication et l'événement*.

69 «*La communication est un cas particulier du transport. Notre langage courant emploie d'ailleurs le terme de 'communications' pour désigner les routes, les chemins de fer, les canaux ou les lignes aériennes*».
ESCARPIT, R., *op. cit.*, p. 7.
La comunicazione è un caso particolare di trasporto. Il nostro linguaggio corrente d'altronde utilizza il termine comunicazione per indicare le strade, le ferrovie, i canali o le linee aeree.

70 *Ivi*, pp. 19-20.
È quindi probabile che, dal momento in cui si passa dal meccanico al vivente e, ancora di più, dal vivente all'umano, divenga necessario un nuovo strumento epistemologico. Questo strumento è rappresentato indubbiamente dalla nozione di informazione. Possiamo azzardare l'ipotesi che questa nozione sia strutturalmente legata a quella di entropia, ma che essa ne differisca poiché prende in considerazione due elementi lasciati da parte dalla scienza fisica: la vita e soprattutto il pensiero.

«La communication est un acte et l'information est son produit» (Escarpit, 1991)[71]: sia che fallisca, sia che abbia successo, questo atto genera ugualmente informazione. Ciò rappresenta una differenza rispetto al funzionamento di un sistema telegrafico. In questo quadro generale la nozione di documento va definendosi in relazione alle specificità del *vivant* e del *temporel*. Un sistema di comunicazione esiste in quanto tale solo se è percepito da una *persona* umana, *«c'est-à-dire par une entité autonome, à la fois mécanique, vivante et pensante, toujours en lutte contre le temps. Toute description humaine d'un tel système tend donc à se soustraire à l'événement […] ou tout au moins à le prévoir, à le dominer, à le manipuler»* (Escarpit, 1991)[72]. Il documento è quindi definito come un *anti-événement*. Non si tratta di un'iscrizione della memoria che salta fugacemente da un messaggio ad un altro, ma è la forma materiale della *mémoire des données*[73], memoria al contempo individuale e culturale. Un passo decisivo è stato compiuto quando l'uomo *«a institué le document, cumulation de traces fixes et permanentes où les réponses données en feedback à travers le temps restent disponibles pour une lecture […]*[74]» (Escarpit, 1991). I *problemi documentali* riguardano esattamente tale feedback. Non vi sono dunque elementi per una scienza dell'informazione, ma la documentazione occupa indubbiamente un posto all'interno delle SIC.

---

71 *Ivi*, p. 100.
La comunicazione è un atto e l'informazione è il suo prodotto.

72 ESCARPIT, R., *op. cit*., p. 61.
Ovvero da un'entità autonoma, al tempo stesso meccanica, viva e pensante, sempre in lotta contro il tempo. Ciascuna descrizione umana di un simile sistema tende quindi a sottrarsi all'evento […] o per lo meno a prevederlo, a dominarlo, a manipolarlo.

73 Memoria dei dati.

74 ESCARPIT, R., *op. cit*., p. 62.
Ha istituito il documento, accumulo di tracce fisse e permanenti nel quale le risposte date come feedback nel corso del tempo restano disponibili per una lettura […].

Nel definire tali scienze, i francesi si sono orientati più verso lo studio della comunicazione in tutte le sue forme che verso l'eredità storica della documentazione e dell'idea delle scienze dell'informazione così come concepita dall'UFOD[75] all'inizio del XX secolo.

1.2 *Fondamenti accademici*

Non esiste in Francia un equivalente delle *library and information sciences* anglosassoni. Le scienze dell'informazione rappresentano in origine un punto di incontro tra professionisti dell'informazione e ricercatori (Couzinet, 2002)[76]. Intesa come un fenomeno sociotecnico, secondo (Ollivier, 2007), *«l'information est désormais quelque chose qui circule et se voit traduit dans le cadre de médiations et d'interactions complexes qui dépasse largement la conception purement technique dans laquelle on l'a longtemps isolée»*[77]. Le ricerche condotte nelle scienze dell'informazione e della comunicazione rientrano all'estero nell'ambito di discipline quali le scienze dell'informazione, gli studi dei media, le scienze della comunicazione, gli studi culturali, la sociologia, le scienze politiche, gli studi di letteratura o la semiotica (Jeanneret, 2001)[78].

---

75 *Union Française des Organismes de Documentation*.

76 Cfr. VIVIANE COUZINET, *Convergences et dynamiques nationales: pour une mise en visibilité des recherches en sciences de l'information*, in Recherches récentes en Sciences de l'information. Convergences et dynamiques: Actes du colloque Mics-Lerass, Università Paul Sabatier, Toulouse, 21-22 mars 2002, Parigi, ADBS Éditions, pp. 9-14.

77 Cfr. BRUNO OLLIVIER, *Les sciences de la communication. Théories et acquis*, Parigi, Armand Colin, 2007, p. 181.
L'informazione è ormai qualcosa che circola e che viene tradotta nell'ambito di mediazioni e interazioni complesse che superano largamente l'idea puramente tecnica nella quale è stata isolata per lungo tempo.

78 Cfr. YVES JEANNERET, *Les sciences de l'information et de la communication: Une discipline méconnue en charge d'enjeux cruciaux*, in «La lettre d'inforcom», n. 60, 2001, pp. 3-45.

I lavori nel campo della documentazione sono integrati nella 71esima sezione del CNU (*Conseil National des Université*)[79]: le Scienze dell'Informazione e della Comunicazione. Essa è stata ufficialmente creata in qualità di disciplina accademica nel gennaio del 1975 ad opera di personalità eminenti quali Robert Escarpit, Jean Meyriat, Roland Barthes e Fernand Terrou con l'obiettivo di dare una cornice universitaria a un insieme di formazioni composite tra le quali la storia del libro, la biblioteconomia o il giornalismo. Nel 1978 è stata fondata la *Société française des sciences de l'information et de communication* (SFSIC). Tuttavia, è importante qui ricordare le ricerche nell'ambito dell'informatica documentale e dei sistemi di organizzazione della conoscenza condotte a partire dagli anni Cinquanta da pionieri quali Gardin, Pages, Levy (Boure, 2002)[80].

Allo stesso modo, secondo il CNU, rientrano nel campo delle SIC i lavori (CNE, 1993)[81]:

A. Les études sur les notions d'information et de communication, sur leurs relations, sur la nature des phénomènes et des pratiques ainsi désignés, de même que les différentes approches scientifiques qui s'y appliquent.

B. L'étude, d'une part, des processus, des productions et des usages de l'information et de la communication, d'autre part, de la conception et de la réception de celles-ci. Ainsi que l'étude des processus de médiation et de médiatisation.

C. L'étude des acteurs, individuels et institutionnels, de l'information et de la communication, l'étude des profession-

79 <http://www.cpcnu.fr/>.

80 Boure R. (a cura di), *Les origines des Sciences de l'information et de la communication. Regards croisés*, Villeneuve d'Ascq, Presses Universitaires du Septentrion, 2002.

81 Cfr. COMITÉ NATIONAL D'EVALUATION (CNE), *op. cit.*

nels (dont notamment les journalistes) et de leurs pratiques.

D. *L'étude de l'information, de son contenu, de ses systèmes sous l'angle des représentations, des significations ou des pratiques associées.*
E. *L'étude des médias de communication et des industries culturelles sous leurs divers aspects*[82].

La particolarità di queste raccomandazioni consiste nel prendere in considerazione l'insieme della catena dell'informazione e della sua comunicazione. La creazione della sezione ha permesso di istituzionalizzare la ricerca accademica dando alle università la possibilità di rilasciare titoli di *doctorat de troisième cycle*[83] che vanno così a costituire un vivaio di ricercatori (CNE, 1993)[84].

---

82 <http://www.cpcnu.fr/web/section-71>.
A. Gli studi sulle nozioni di informazione e di comunicazione, sulle relazioni tra le stesse, sulla natura dei fenomeni e delle pratiche così designate, così come i differenti approcci scientifici che vi si applicano; B. Lo studio, da una parte, dei processi, delle produzioni e degli usi dell'informazione e della comunicazione, dall'altra, dell'elaborazione e della ricezione di queste ultime. Analogamente, lo studio dei processi di mediazione e di mediatizzazione; C. Lo studio degli attori, individuali e istituzionali, dell'informazione e della comunicazione, lo studio dei professionisti (tra cui in particolare i giornalisti) e delle loro pratiche; D. Lo studio dell'informazione, del suo contenuto, dei suoi sistemi dal punto di vista delle rappresentazioni, dei significati o delle pratiche associate; E. Lo studio dei mezzi di comunicazione e delle industrie culturali nei loro differenti aspetti.

83 Fino al 1984 esistevano in Francia tre tipi di dottorato: *Doctorat de troisième cycle*, della durata di uno o due anni, *Doctorat d'ingénieur*, accessibile da chi aveva conseguito un titolo di studio in ingegneria e il *Doctorat d'État*, indispensabile per accedere alla professione di *Professeur des Universités*.
<http://www.e-tud.com/encyclopedie-education/?136-doctorat> [*N.d.T.*].

84 Comité National d'Evaluation (CNE), *op. cit.*, p. 15.

Per quanto concerne le pubblicazioni, la 71esima sezione del CNU possiede una lista di riviste riconosciute come riviste di riferimento nella disciplina dell'informazione e della comunicazione. Ne esistono due categorie: la prima comprende le riviste accademiche di riferimento principalmente sulle scienze della comunicazione, mentre la seconda contiene riviste professionali che pubblicano articoli scientifici. Al fine di valutare la produzione scientifica dei ricercatori, l'Aeres[85] ha integrato ulteriori riviste internazionali in scienze dell'informazione, oltre a due riviste professionali pubblicate in Francia: *Documentaliste-Sciences de l'Information*, creata nel 1964 e pubblicata dall'ADBS, e il BBF (*Bulletin des Bibliothèques de France*), creato nel 1956 e pubblicato dall'Enssib.

Analizzando la natura degli articoli pubblicati nella rivista *Documentaliste-Sciences de l'information* a partire dal 1964, (Couzinet, 2000)[86] individua tre periodi: il periodo 1964-1976, nel quale gli articoli sono redatti quasi esclusivamente da professionisti; il periodo 1976-1989, quando i ricercatori iniziano a pubblicare sul giornale (da uno a cinque articoli all'anno) e, infine, il periodo che inizia nel 1990, durante il quale il numero di articoli pubblicati dai ricercatori tende ad eguagliare quello degli articoli pubblicati dai professionisti. Un'analisi di 318 articoli pubblicati in quattro anni dal BBF (Couzinet, 2000)[87] dimostra che 52 articoli (16,35%) sono stati scritti da ricercatori. Per quanto riguarda le tesi di dottorato, (Polity, 2001)[88] dà prova, attraverso l'analisi di 90 tesi discusse nell'ambito delle scienze dell'informazione per il periodo compreso tra il 1971 e il 2000,

85 *Agence d'évaluation de la recherche et de l'enseignement supérieur*.

86 Cfr. VIVIANE COUZINET, *Médiations hybrides: le documentaliste et le chercheur de sciences de l'information*, Parigi, ADBS éditions, 2000.

87 Cfr. COUZINET, V., *op. cit.*, 2000.

88 YOLLA POLITY, *Les bibliothèques, objets de recherche universitaire*, in «Bulletin des bibliothèques de France», vol. 46, n. 4, 2001, p. 66.

che la maggior parte di esse verte sull'*intelligence économique*, la sociologia della lettura e l'informatica documentale. Sulla base della distinzione operata da Richard Whitley tra l'istituzionalizzazione cognitiva e l'istituzionalizzazione sociale, (Palermiti, Polity, 2000) constata

> *[…] qu'il n'y a pas eu pour les SI de corrélation entre une certaine institutionnalisation cognitive amorcée dès le début de ce siècle, et leur reconnaissance universitaire qui est un des facteurs académiques déterminant de l'institutionnalisation sociale. Cette dernière s'est faite en l'absence d'une fraction importante du monde de l'information, celle du secteur des bibliothèques et des archives, que les sciences de l'information n'ont pas su ou pas pu intégrer*[89].

È importante sottolineare che l'attuazione di riforme strutturali nelle università francesi (creazione dell'Aeres, diminuzione dei budget, individualizzazione della valutazione, modifica dello statuto dei ricercatori, concorrenza intra e interuniversitaria) hanno modificato le condizioni di accesso e di diffusione delle conoscenze scientifiche (Devroey, 1999)[90]. I recenti cambiamenti istituzionali, in particolare le procedure di valutazione normate dall'Aeres, possono sfociare in nuove modalità di argomenta-

---

89 PALERMITI, R., POLITY, Y., *op. cit.*, p. 96.
[…] che non ci sia stata per le SI una correlazione tra una certa istituzionalizzazione cognitiva avviata fin dall'inizio di questo secolo e il loro riconoscimento universitario, che è uno dei fattori accademici determinanti dell'istituzionalizzazione sociale. Quest'ultima si è delineata in assenza di una frazione importante del mondo dell'informazione, quella del settore delle biblioteche e degli archivi, che le Scienze dell'Informazione non hanno saputo o non hanno potuto integrare.

90 Cfr. JEAN-PIERRE DEVROEY, *La place de la bibliothèque dans la formation documentaire à l'université*, 1999.
<http://www.fr.ch/bcu/a/pub_elec/etudes_et_recherche_info.pdf>.

zione, ma anche in un'evoluzione delle prassi di valutazione *scientifica* nella nostra disciplina.

Da un punto di vista istituzionale e con riferimento ai lavori di (Cardy, Froissart, 2002)[91], (Cardy, Froissart, 2006)[92], la ricerca in scienze dell'informazione e della comunicazione è un dominio in pieno sviluppo e sempre più riconosciuto in Francia. Il numero di assunzioni e di offerte di lavoro è in crescente aumento. Dal 1977 al 2005 il numero di posti è aumentano annualmente del 10%, a partire dai 43 *enseignants-chercheurs*[93] del 1997 fino ai 663 del 2005. Sempre secondo (Cardy, Froissart, 2006) *«le nombre de postes universitaires en Infocom a dépassé les domaines tels que la philosophie et les sciences politiques, et est proche de la sociologie et la linguistique»*[94]. L'analisi dei profili professionali nel periodo 1995-2001 mette in evidenza, secondo gli autori, come la documentazione e le NTIC rappresentino il 23% dei posti. Le tematiche di ricerca degli *enseignants-chercheurs* in informazione-documentazione (Fondin,

---

91 Cfr. HÉLÈNE CARDY, PASCAL FROISSART, *Les enseignants-chercheurs en Sciences de l'information et de la communication. Portrait statistique*, in Les recherches en information et communication et leurs perpectives: Histoire, objet, pouvoir, méthode. Actes du XIII Congrès national des sciences de l'Information et de la communication, Marsiglia, 7-9 ottobre 2002, SFSIC, 2002, pp. 353-362.

92 Cfr. HÉLÈNE CARDY, PASCAL FROISSART, *SIC: cartographie d'une discipline*, in Sciences de l'information et de la communication. Objets, savoirs, discipline, Olivesi S. (a cura di), Grenoble, Presse Universitaire de Grenoble, 2006, pp. 259-278.

93 Ricercatori tenuti anche a svolgere attività di docenza in ambito universitario [*N.d.T.*].

94 CARDY, H., FROISSART, P., *op. cit*., 2006.
Il numero di posti universitari nell'ambito dell'*Infocom* (*Sciences de l'Information et de la Communication* - Scienze dell'Informazione e della Comunicazione) è maggiore rispetto a settori quali la filosofia e le scienze politiche ed è prossimo a quelli della sociologia e della linguistica.

Rouault, 1998)[95] (CNE, 1993)[96] riguardano la rappresentazione delle conoscenze e l'organizzazione del sapere, la comunicazione uomo-macchina (analisi dei bisogni degli utenti, analisi delle pratiche informatiche, ecc.); l'impiego di strumenti informatici per la gestione dei processi (concezione, sviluppo e valutazione dei dispositivi informatici, ecc.); l'informazione e la sua gestione (lavori sulla veglia e sulla gestione dell'informazione e delle istituzioni, ecc.) e infine l'informazione nella società (economia dell'informazione, diritto dell'informazione, divario digitale[97], ecc.). Il successo di internet ha ispirato opere che mettono nuovamente in dubbio il concetto di informazione. Faremo riferimento nel seguito a lavori recentemente condotti in contesto francese e concernenti due tematiche importanti relative rispettivamente all'evoluzione delle professioni e della disciplina.

2. Sfide e prospettive per le scienze dell'informazione

L'universo documentale a partire dal quale si è definito, nel corso di un intero secolo, un corpus di teorie, strumenti, metodi e norme è oggi stravolto dal digitale. Questa *(r)évolution*[98] del digitale ha condotto alcuni autori a parlare di *redocumentarisation*[99] (si veda la Figura 3), altri ancora di *bibliothécarisa-*

95 Cfr. FONDIN, H., ROUAULT, J., *op. cit*.
96 Cfr. COMITÉ NATIONAL D'EVALUATION (CNE), *op. cit.*
97 *«Il divario digitale rappresenta lo scarto tra chi ha pieno accesso alle tecnologie e alle possibilità offerte dal digitale e chi invece ne è escluso»*. <http://www.agendadigitale.regione.lombardia.it/cs/Satellite?c=Page&childpagename=DG_01%2FMILayout&cid=1213474655678&p=1213474655678&pagename=DG_01Wrapper> [*N.d.T.*].
98 Rivoluzione-evoluzione.
99 *«L'objectif de la documentarisation est d'optimiser l'usage du document en permettant un meilleur accès à son contenu et une meilleure mise en*

tion[100] e/o di *googlisation*[101] del mondo (Bazin, 2006)[102], dal momento che il trattamento dell'informazione (che per molto tempo è stato un dominio di competenza dei bibliotecari), è sempre meno appannaggio degli specialisti dell'informazione, diventando una sfida sociale nel senso più lato.

	Documentarisation	**Redocumentarisation**
Dates	Tournant XIX^{e}-XX^{e}	Tournant XX^{e}-XXI^{e}
Quelques figures	M. Dewey, P. Otlet, O. Lafontaine, W Carnegie	T. Berners-Lee, T. Nelson, B. Gates, S. Brin
Quelques techniques	Classification, Indexation, Langages documentaires, Thésaurus...	Protocoles Web (Html, Url) Web 2.0, Web sémantiqueOntologies...
Quelques réalisations	Réseau mondial de bibliothèques	Google, Wikipédia
Les modernités	L'esprit scientifique, la raison-logique, l'État-nation, les votes, l'industrie, l'auteur...	Le savoir limité, la raison-statistique, l'individu, les opinions, les services, la réflexivité...
Quelques objets documentaires concernés	Les revues, les règlements, les contrats, les brevets, les œuvres, les médias et l'imprimerie	Les pré-publications, les formulaires, les sources ouvertes, les wikis, les blogues et le web

Figura 3. Documentarisation - Redocumentarisation[103].

contexte. Le numérique, par nature, implique une re-documentarisation. […]En effet, bien des unités documentaires du Web ne ressemblent plus que de très loin aux documents traditionnels. […] Il s'agit alors d'apporter toutes les métadonnées indispensables à la reconstruction à la volée de documents et toute la tracabilité de son cycle».
Cfr. JEAN-MICHEL SALAÜN, *La redocumentarisation, un défi pour les sciences de l'information*, in «Études de Communication» n. 30, 2007, p. 2.
L'obiettivo della documentalizzazione è quello di ottimizzare l'utilizzo del documenti permettendo un migliore accesso al suo contenuto e una sua migliore contestualizzazione. Il digitale, per sua natura, implica una ri-documentalizzazione. […] Effettivamente, gran parte delle unità documentali del Web non sono più assimilabili, se non lontanamente, ai documenti tradizionali. […] Si tratta, quindi, di apportare tutti metadati indispensabili alla ricostruzione immediata di documenti e alla tracciabilità del loro ciclo.

100 <http://abfblog.wordpress.com/2009/06/page/3/>.

101 In senso generale si intende l'uso dei motori di ricerca [*N.d.T.*].

102 PATRICK BAZIN, *L'avenir incertain des bibliothèques*, 2006, p. 105. <http://cla.univ-fcomte.fr/gerflint/Perou2/Bazin.pdf>.

103 SALAÜN, J-M. *op. cit.*, 2007, p. 4.

Questo *nouvel ordre documentaire*[104] interessa organizzazioni sociali e ideologiche diverse: da qui il bisogno di una critica di questa ragione digitale (Paul, Perriault, 2004)[105].

2.1 *Immaginare le metamorfosi del documento*

I lavori recentemente condotti in Francia in merito al documento, a partire dalle riflessioni di (Roger T. Pédauque, 2006)[106], hanno messo in evidenza l'avvento di nuove argomentazioni relative soprattutto alla durata, alla temporalità, alla metamorfosi e all'*ibridazione* del documento e dello spazio documentale. Sul

[104] *«L'ordre documentaire ancien, l'écosystème documentaire était non marchand. Le nouvel ordre documentaire s'appuie sur des stratégies commerciales, avec un besoin de captation de l'attention. Il est donc nécessaire de se tourner vers des architectes de l'information, c'est à dire des personnes compétentes dans trois domaines: la documentation (repérer, organiser des ressources), le web (connaissance des outils) et une expérience d'utilisateur (construire des services intuitifs)».*
JEAN-MICHEL SALAÜN, *Les trois facettes du document numérique et le nouvel ordre documentaire*, Intervento al Seminario «Ressources numériques au CDI», 10-11 maggio 2012.
L'*ordine documentale* antico, l'ecosistema documentale era non commerciale. Il nuovo *ordine documentale* si basa su strategie commerciali, con un bisogno di catturare l'attenzione. È quindi necessario rivolgersi ad architetti dell'informazione, ovvero a persone competenti in tre domini: la documentazione (individuare, organizzare le risorse), il web (conoscenza degli strumenti) e un'esperienza da utente (costruire servizi intuitivi).

[105] Cfr. Paul V., Perriault J. (a cura di), *Critique de la raison numériqu*e, in «Hermès», vol. 39, Parigi, CNRS Editions, 2004.

[106] Cfr. ROGER T. PÉDAUQUE, *Le Document à la lumière du numérique: forme, texte, médium: comprendre le rôle du document numérique dans l'émergence d'une nouvelle modernité*, C&F éditions, 2006.
'Roger T. Pédauque è uno pseudonimo collettivo utilizzato da una rete di scienziati francofoni che lavorano nei diversi domini delle scienze umane e sociali e delle scienze e tecniche dell'informazione e della comunicazione [*N.d.T.*]'.

piano tecnologico, negli ultimi dieci anni si è assistito allo sviluppo di nuovi modelli di ricerca che sfruttano principalmente la strutturazione dei documenti, una maggiore personalizzazione delle funzionalità, così come l'integrazione della ricerca d'informazione su nuovi supporti come il PDA[107]. Le tecnologie digitali associate ad internet hanno progressivamente portato ad una destrutturazione completa della nozione originaria di documento, ma non alla sua soppressione o cancellazione (Ihadjadene, Chaudiron, 2008)[108]. Tale processo è innegabilmente legato ad un contesto tecnologico particolare, ma è al tempo stesso indotto da pratiche informative nuove (il deposito dei pre-print negli archivi aperti ad esempio) e da un'evoluzione dei modelli economici dei grandi editori (la vendita di articoli o capitoli).

Nel dominio dell'IST (Informazione Scientifica e Tecnica) un aspetto di tale cambiamento è legato alla rivisitazione delle condizioni della ricerca scientifica, che causa il problema della collocazione degli specialisti dell'informazione. In effetti, i lavori attuali nell'ambito delle *e-sciences*[109] e dell'informatica umanistica mettono in dubbio l'approccio scientifico e gli strumenti che permettono di validare il processo di raccolta, analisi, trattamento e restituzione dei dati. Al di là del documento digitale (articolo, monografia,…) o del dato grezzo, si tratta d'ora in avanti di proporre un accesso ai protocolli di esperienza, alle raccolte di dati e agli applicativi software utilizzati negli esperimenti scientifici. Si crea in tal modo un continuum nel *costrutto* scientifico nel quale i dati osservati, i protocolli, le ipotesi, le inferenze, i ri-

[107] *Personal Digital Assistant.*

[108] Cfr. MADJID IHADJADENE, STEPHANE CHAUDIRON, *L'Étude des dispositifs d'accès à l'information électronique: approches croisées*, in Problématiques émergentes dans les sciences de l'information, Papy F. (a cura di), Parigi, Hermès-Lavoisier, 2008, pp. 183-207.

[109] Le *e-sciences* sono le scienze che si basano sulla produzione e sulla diffusione di grandi quantità di dati digitali [*N.d.T.*].

sultati, ecc. si articolano in una struttura reticolare. Per il bibliotecario o per il documentalista, la sfida non consisterà più semplicemente nella gestione della pubblicazione dei risultati della propria ricerca, quanto piuttosto nella gestione degli stessi dati di ricerca. Nel dominio delle SUS (Scienze Umane e Sociali), a partire dalla metà degli anni '80, sono state avviate delle ricerche sulla modellizzazione a priori dei testi scientifici finalizzata a proporre una scrittura multimediale adatta ad Internet che tentasse al tempo stesso di assicurare la facile manipolazione delle costruzioni scientifiche e l'accesso all'insieme dei dati sui quali esse si fondano (Gardin, Roux, 2004)[110]. Darnton, citato da (Renoult, 2001)[111], mostra come Internet sia in grado di modificare profondamente non solo l'economia delle riviste, ma anche la loro capacità di argomentazione scientifica. Egli fa riferimento ad una trasformazione della monografia scientifica in una gerarchia a sei livelli (stampato, livello pedagogico, livello teorico o storico, apparato critico, note, appendici). Questa analisi evidenzia come la nozione classica di monografia o dello stesso documento si stia trasformando e sviluppando a vantaggio della nozione evolutiva di risorsa digitale. Non si tratta solo di mettere online dei documenti, ma di costituire una vera e propria gestione editoriale dei contenuti digitali che apra così a nuove pratiche di scrittura il cui potenziale non è sempre pienamente sfruttato (Bachimont, 1999)[112].

---

[110] Cfr. JEAN-CLAUDE GARDIN, VALENTINE ROUX, *The Arkeotek project: a european network of knowledge bases in the archaeology of techniques*, in «Archeologia e Calcolatori», vol. 15, 2004, pp. 25-40.

[111] DANIEL RENOULT, *Les bibliothèques numériques*, in Des Alexandries 1. Du livre au texte, Giard L., Jacob C. (a cura di), Parigi, Bibliothèque National de France, 2001.

[112] Cfr. BRUNO BACHIMONT, *Bibliothèques numériques audiovisuelles: des enjeux scientifiques et techniques*, in «Document Numérique», vol. 2, n. 3-4, 1999, pp. 219–242.

Infine, si pone il problema dell'appartenenza di un documento ad una collezione nel senso biblioteconomico del termine. Non ci dilungheremo su questo punto, già più volte trattato, ma ricorderemo soltanto che, nell'ambito dell'informazione scientifica e tecnica, questa decostruzione dell'*effetto collezione* è particolarmente forte. (Wiegandt, 1995)[113] sottolinea la necessità per gli specialisti dell'informazione di liberarsi dalla localizzazione e dalla nozione di possesso a favore della nozione di accesso. Secondo l'autrice questa evoluzione rappresenta una svolta notevole che turba il rapporto tra il professionista e la collezione. Le nozioni classiche tradizionali di documento, di frammento documentale, di collezione sono stravolte dal digitale. Il supporto scrittorio e la forma semiotica scelta limitano, secondo (Bachimont, 1999)[114] l'espressione del contenuto e le sue condizioni di ricezione, interpretazione e comunicazione.

Immaginare le metamorfosi del documento implica una riflessione sulle possibilità di cooperazione e di convergenza tra gli archivi, le biblioteche e i musei (ALM secondo l'acronimo anglosassone). (Yarrow et alii, 2008)[115] e (Zorich et alii, 2008)[116] descrivono lo stato delle cooperazioni e delle collaborazioni di queste istituzioni culturali. Gli autori considerano la collaborazione come un continuum che va dal semplice contatto tra le istituzioni di base ad una convergenza e interdipendenza che permette alle istituzioni stesse di trascendere le proprie missioni d'origine e i propri ruoli.

---

113 WIEGANDT, C., *op. cit.*, pp. 17-18.

114 BACHIMONT, B., *op. cit.*, p. 223.

115 Cfr. ALEXANDRA YARROW, BARBARA CLUBB, JENNIFER-LYNN DRAPER, *Bibliothèques publiques, archives et musées: tendances en matière de collaboration et de coopération*, Rapporto IFLA-Comité permanent de la section des bibliothèques publiques, n. 109, 2008.

116 Cfr. DIANE ZORICH, GÜNTER WAIBEL, RICKY ERWAY, *Beyond the Silos of the LAMs: Collaboration among Libraries, Archives and Museums*, Dublino, OCLC Research, 2008.

2.2 *Approcci infocomunicativi*

Se in passato la peculiarità francese, ovvero le scienze dell'informazione e della comunicazione come frutto di una congiuntura particolare e istituzionale che ha articolato due discipline fortemente isolate all'estero, era difficilmente sostenibile sul piano scientifico, è diventata oggi appropriata per studiare e per fornire dei chiarimenti sulle modifiche sociali e tecnologiche indotte dalla crescente digitalizzazione delle attività culturali e commerciali. Il superamento dell'angusto ambito di studio dei documenti o dell'artefatto per interessarsi ad altre dimensioni, quali l'uso, le pratiche informative, le politiche d'informazione o le questioni etiche o economiche mostra, al contrario, il bisogno di un approccio infocomunicativo (Miège, 2004)[117]. Quest'ultimo ricorda che le evoluzioni attuali dell'informazione specializzata (commercializzazione, internazionalizzazione, diffusione tramite nuovi media) avvicinano i settori della documentazione-biblioteca a quello delle industrie culturali. Altri autori hanno proposto di arricchire le ricerche in scienze dell'informazione con le riflessioni teoriche generate dalla comunicazione (Chaudiron, Ihadjadene, 2010)[118]. Gli studi in scienze della comunicazione hanno descritto e analizzato per circa trent'anni le reazioni dei destinatari dei programmi di fronte alle offerte che erano state loro indirizzate. Questi lavori (studi di ricezione, sociologia della cultura e delle pratiche culturali, sociologia degli usi, analisi delle mediazioni, teoria critica, ecc.) rappresentano un notevole apporto per una migliore comprensione e modellizzazione delle pratiche informative. Per molto tempo, i lavori relativi all'informazione sono stati concepiti senza prendere in considera-

[117] Cfr. MIÈGE, B., *op. cit*., 2004.

[118] Cfr. STEPHANE CHAUDIRON, MADJID IHADJADENE, *De la recherche d'information aux pratiques informationnelles*, in «Étude de Communications», vol. 35, 2010, pp. 13-29.

zione le funzioni della comunicazione che la accompagnano (Fondin, 2006)[119]. Questa è probabilmente una delle ragioni per le quali la maggior parte dei modelli informativi minimizza le attività di condivisione e di comunicazione dell'informazione e privilegia spesso l'attività di accesso e di ricerca.

Pertanto, i cambiamenti tecnici dei dispositivi di produzione, di ricerca, di condivisione e di diffusione dell'informazione inducono a trasformazioni organizzative, favorendo l'emergere di un nuovo ecosistema informativo, che rappresenta un terreno privilegiato per rinnovare gli approcci sulle pratiche informative. Questi nuovi dispositivi (Web 2.0, reti sociali, CMS[120], software di filtraggio, ecc.) banalizzano le modalità di produzione dell'informazione e riuniscono le funzionalità di ricerca e di edizione. In tal senso, i sistemi attuali di accesso all'informazione (in particolare i motori di ricerca) presentano le caratteristiche di un nuovo media (di un web-media piuttosto). (Salaün, 2006)[121] (Halavais, 2009)[122] sottolineano che il mondo del web prende in prestito dalla radio-televisione il modello economico, essendo questo fondato su un'economia dell'attenzione, ma che si rifà anche al modello della biblioteca attraverso la condivisione e la messa in comune delle conoscenze. I suddetti dispositivi sono dei media, ovvero degli oggetti che, riprendendo (Jeanneret, 2009)[123],

[119] Cfr. FONDIN, H., *op. cit.*

[120] *Content Management Systems.*

[121] Cfr. JEAN-MICHEL SALAÜN, *Web-média, synthèse*, 2006. <http://grds04.ebsi.umontreal.ca/jms/index.php/2006/11/09/116-web-media-synthese>.

[122] Cfr. ALEXANDER HALAVAIS, *Search Engine Society*, Cambridge, Polity Press, 2009.

[123] Cfr. YVES JEANNERET, *La relation entre médiation et usage dans les recherches en information-communication en France*, in «RECIIS Electronic Journal of Communication Information & Innovation in Health», vol. 3, n. 3, 2009. <http://www.reciis.cict.fiocruz.br/index.php/reciis/article/view/276/320>.

non solo si associano alla socialità, ma la producono attraverso, in particolare, mediazioni simboliche e rappresentative. Il collegamento con gli approcci comunicativi può mettere in evidenza delle sfide legate alla mediazione dei saperi, alla gestione dell'identità digitale e al controllo e gestione della capacità di decisione. Superando il loro status di semplici riceventi, gli utenti giocano ormai un ruolo attivo nella produzione, classificazione e valutazione dell'informazione. Gli individui, ma anche le organizzazioni, mettono in atto delle strategie o delle politiche per facilitare lo scambio di informazioni e la relativa comunicazione attraverso piattaforme di intermediazione (Intranet, reti sociali, ecc.) (Georges, 2010)[124].

Un percorso di ricerca che riteniamo stimolante e motivante consiste nello studio delle pratiche informative a partire dalla prospettiva della mediazione digitale[125]. In effetti, in un istituto culturale o in un'organizzazione (in particolare per la gestione

[124] Cfr. FANNY GEORGES, *Identités virtuelles. Les profils utilisateur du Web 2.0*, Questions théoriques, 2010.

[125] «*La médiation numérique est l'utilisation d'outils numériques afin de faire se rencontrer une offre et une demande de contenus dans le cadre d'une politique documentaire. Il s'agit d'organiser l'interaction entre le public – réel ou 'virtuel' – les professionnels de l'information et de la documentation et les contenus. La médiation numérique est donc une voie de renouvellement de nos pratiques pour être présent dans l'environnement web de l'usager/internaute et répondre aux besoins d'orientation dans les contenus disponibles. […]*».
<http://www.adbs.fr/la-mediation-numerique-des-collections-90920.htm>.
La mediazione digitale consiste nell'utilizzo di strumenti digitali orientati a far incontrare un'offerta e una domanda di contenuti nell'ambito di una politica documentale. Si tratta di organizzare l'interazione tra il pubblico – reale o virtuale – i professionisti dell'informazione e della documentazione e i contenuti. La mediazione digitale è quindi una modalità di rinnovamento delle nostre pratiche per essere presenti nell'ambiente web dell'utente/internauta e rispondere ai bisogni d'orientamento nei contenuti disponibili.

del relativo contenuto informativo), la scelta delle norme e degli standard, le politiche di indicizzazione e di classificazione, le operazioni di inventario e di costituzione di collezioni non sono neutre, ma vengono mediate da logiche sociali. Infatti, secondo (Dufrene, 2007), tali operazioni non sono *«indépendants du cadre dans lesquels ils s'élaborent; leur sens est relatif à des conditions d'énonciation propres à des types d'institutions et à des situations historiques»*[126]. L'interesse di un approccio che rientri nella prospettiva della mediazione è quello di avviare una critica delle condizioni nelle quali i testi, i dati e i dispositivi sono prodotti, al fine di circoscrivere meglio la loro interpretazione (Jeanneret, 2009)[127].

Il concetto di mediazione, largamente impiegato nei paesi francofoni fin dagli anni '90, è stato oggetto essenzialmente di due tipi di approcci: il passaggio dal paradigma sistema (orientato alla collezione) al paradigma utenti (Le Coadic, 1997)[128] ha

«La médiation numérique est un dispositif technique, éditorial ou interactif mis en œuvre par des professionnels de l'information-documentation, favorisant l'accès organisé ou fortuit à tout contenu proposé par une bibliothèque, son appropriation et sa dissémination».
LIONEL DUJOL, *La médiation numérique: l'exemple des Médiathèques du Pays du Romans, France*, in «Argus», vol. 39, n. 3, 2011, p. 18.
La mediazione digitale è un dispositivo tecnico, editoriale o interattivo al quale ricorrono i professionisti dell'informazione-documentazione, favorendo l'accesso organizzato o casuale a ciascun contenuto proposto da una biblioteca, la sua appropriazione e disseminazione.

[126] BERNADETTE DUFRENE, *Intérêts d'une approche sociohistorique des questions de médiation culturelle*, in Quelles approches de la médiation culturelle?, Parigi, L'harmattan, 2007, p. 151.
Indipendenti dal contesto nel quale vengono eseguite. Il loro significato è correlato a condizioni di enunciazione proprie a tipologie di istituzioni e situazioni storiche.

[127] JEANNERET, Y., *op. cit.*, 2009.

[128] Cfr. YVES-FRANÇOIS LE COADIC, *Usages et usagers de l'information*, Parigi, ADBS: Nathan, 1997.

rafforzato anche, nel caso delle biblioteche, la funzione di accoglienza e il ruolo del bibliotecario come guida delle innovazioni, in particolare quelle tecniche (Chourrot citato da [Ihadjadene, Dufrene, 2011])[129], mentre l'*événementialisation*[130] della cultura, che interessa tutte le istituzioni di conservazione, dà origine ad una serie di studi sul ruolo culturale in materia di animazione e di valorizzazione. Questa mediazione tradizionale implica un'interazione in spazi particolari, quali biblioteche e centri di documentazione, in cui l'utente ha un bisogno informativo immediato. Tali interazioni sono ugualmente arricchite dalla presenza fisica delle collezioni (Demeurisse et alii, 2008)[131]. Le fonti, immediatamente disponibili, possono aiutare a chiarire molti aspetti fin dall'inizio della ricerca. La missione di queste ultime si è evoluta dalla semplice messa a disposizione di informazioni alla costruzione di legami e interazioni tra bisogno ed uso dell'informazione (Ihadjadene, Chaudiron, 2001)[132].

L'emergere di pratiche di *automédiation*[133] non modifica tut-

[129] MADJID IHADJADENE, BERNADETTE DUFRENE, *Les médiations en bibliothèque: une logique de service public?*, in «Argus», vol. 39, n. 3, 2011, p. 21.

[130] Trasformazione in evento anche di ciò che non lo è propriamente [*N.d.T.*].

[131] Cfr. JOSIANE SENIÉ DEMEURISSE, ISABELLE FABRE, CÉCILE GARDIÈS, *Organisation du savoir et médiation documentaire: exemple du traitement des périodiques d'histoire dans deux bibliothèques universitaires*, in Médiations et Usages des Savoirs et de l'Information: un dialogue France-Brésil, Marteleto R.M., Thiesen I. (a cura di), 2008, pp. 379-392.

[132] MADJID IHADJADENE, STEPHANE CHAUDIRON, *La recherche et la diffusion d'information sur Internet: vers de nouvelles médiations?*, in Emergences et continuité dans les recherches en information et communication: Actes du XII Congrès national des sciences de l'information et de la communication, Unesco, 10-13 gennaio 2001, Parigi, SFSIC, p. 169.

[133] *«Ce sont ces mouvements que nous approchons par la notion paradoxale d'auto-médiation dans ses deux dimensions en partie contradictoires-d'autonomisation et d'automatisation de la médiation. On parlera d'autonomisation lorsque l'acteur médiatise lui-même l'événement et construit ainsi directement l'espace de sa communication/diffusion. On caractéri-*

te quelle che rientrano nel campo della mediazione (Miège, 2008)[134]. Molti osservatori avevano previsto che con l'avvento internet lo scambio diretto tra le persone avrebbe fatto svanire gli intermediari. Ora, con il passare del tempo, si sviluppa il fenomeno inverso: gli intermediari, seppure in forma differente, sono in pieno sviluppo. La natura della mediazione professionale è diversa: essa riguarda al tempo stesso le funzioni di rappresentazione delle conoscenze e quelle di formazione e di informazione (portali comunitari). Se le interazioni sociali giocano un ruolo essenziale nei comportamenti legati al consumo di informazioni da parte degli internauti, ogni intermediazione efficace si basa oggi su una gestione e uno sfruttamento attento dei metadati di un dato settore e su una messa in atto di una ge-

sera l'automatisation de la médiation par l'usage direct de logiciels permettant d'accomplir directement une tâche (par exemple, rechercher des informations sur Internet). On suivra ces mouvements dont la synthèse donne corps au concept d'auto-médiation, mouvements qui s'expriment dans la substitution directe d'intermédiaires mais aussi, bien que moins explicitement, dans l'évitement d'institutions par des relations horizontales».
<http://hypermedia.univ-paris8.fr/weissberg/presence/3.html>.
Sono questi i movimenti ai quali ci approcciamo attraverso la nozione paradossale di auto-mediazione nelle sue due dimensioni – in parte contraddittorie – di autonomizzazione e automazione della mediazione. Si parlerà di autonomizzazione quanto l'attore stesso mediatizza l'evento e costruisce direttamente anche lo spazio della sua comunicazione/diffusione. L'automazione della mediazione sarà caratterizzata dall'uso diretto di software che permettono di assolvere direttamente ad un compito (ad esempio cercare delle informazioni su internet). Si seguiranno questi movimenti la cui sintesi da vita al concetto di auto-mediazione, movimenti che si esprimono nella sostituzione diretta degli intermediari ma anche, pur se meno esplicitamente, nell'evitare le istituzioni attraverso relazioni orizzontali.

134 Cfr. BERNARD MIÈGE, *Médias, médiations et médiateurs, continuités et mutations*, in «Réseaux», vol. 2, n. 148-149, 2008, pp. 117-146.

stione editoriale dei contenuti informativi. Le industrie culturali sono così un esempio sorprendente dell'impatto della digitalizzazione su una catena del valore e della ridefinizione della funzione di intermediazione (Gensollen, 2009a)[135]. Secondo quest'ultimo, è attraverso la gestione dei metadati e delle piattaforme di intermediazione che si stabilisce la *socialisation des goûts*[136] o il decentramento dei saperi. È indubbiamente in que-

[135] Cfr. MICHEL GENSOLLEN (a), *Le web relationnel: vers une économie plus sociale?*, in Le Web relationnel, mutation de la communication?, Millerand F., Proulx S., Québec, Presses de l'Université du Québec, 2009.

[136] *«Nous avons désigné du terme de «méta-information», d'une part, les données qui permettent la labellisation des produits (en particulier des produits innovants et des biens d'expérience). […] Celui de la méta-information ex ante, le conseil avant l'achat, le bouche-à-oreille, la critique institutionnelle, qui ont toujours existé, se prolongent en ligne, sur des plateformes où les consommateurs qui ont déjà acheté les produits les décrivent, les commentent et, éventuellement, les recommandent à d'autres consommateurs. Pour les biens d'expérience (par exemple, la qualité d'un film qui vient de sortir), comme pour les biens innovants, pour lesquels les consommateurs potentiels ne disposent d'aucune représentation de leur éventuelle utilité, une telle phase d'acculturation est essentielle. A la production d'innovations doit correspondre, en quelque sorte, une «production sociale des goûts»*.
MICHEL GENSOLLEN (b), *A quoi rassemblera le monde numérique, en 2030?*, in «Réalités Industrielles», maggio 2009, p.16.
Abbiamo designato con il termine meta-informazione, da una parte, i dati che permettono l'etichettatura dei prodotti (in particolare i prodotti innovativi e i beni di esperienza).[…] Quello della meta-informazione ex ante, il consiglio prima dell'acquisto, il passaparola, la critica istituzionale, che sono sempre esistiti, continuano online, sulle piattaforme nelle quali i consumatori che hanno già acquistato i prodotti li descrivono, li commentano ed, eventualmente, li raccomandano ad altri consumatori. Per i beni di esperienza (per esempio, la qualità di un film appena uscito), come per i beni innovativi, per i quali i potenziali consumatori non dispongono di nessuna rappresentazione della loro utilità, tale fase di acculturazione è essenziale. Alla produzione dell'innovativo deve corrispondere, in qualche modo, una *produzione sociale dei gusti*.

sto contesto che entrano in gioco i nuovi attori della mediazione. I termini impiegati per identificarli – la cui precisione costituisce di per sé un'opportunità teorica rispetto alla molteplicità dei discorsi – sono diversi: *community managers*, *infomediari*, *go-between*, *formatori*.

La questione della mediazione digitale può contribuire a chiarire molte sfide attuali. Da una parte, essa appartiene ad una categoria di attività nella quale il lavoro non è stabilizzato (Aubouin et alii, 2010)[137] e ci si può interrogare circa la sua appartenenza a ciò che alcuni sociologi di professione chiamano *les métiers flous*[138], dei quali questa indeterminatezza può generare numerose interpretazioni (Champy, 2009)[139] (Denis, Pontille, 2012)[140]. Gli attori hanno grandi margini di manovra per definire la loro professione quotidiana (Boussard et alii, 2010)[141]. Si

[137] Nicolas Aubouin, Frédéric Kletz, Olivier Lenay, *Médiation culturelle: l'enjeu de la gestion des ressources humaines*, in «Culture études», vol. 1, n. 1, 2010, p. 2.

[138] *«Chefs de projets, médiateurs, agents de développement, chargés de mission… des métiers qui ne sont plus 'nouveaux' depuis plusieurs décennies mais qui demeurent bien entourés de 'flou' se sont multipliés dans l'action publique. […], flou des positions dans l'organisation, flou des statuts d'emploi et finalement flou de l'identité et du métier».*
<http://www.octares.com/boutique_fiche.asp?IdProd=106>.
Cfr. Gilles Jeannot, *Les Métiers flous. Travail et action publique*, Toulouse, Editions Octarès, 2005.
Capi progetto, mediatori, agenti di sviluppo, incaricati di missione … mestieri che non sono più *nuovi* già da diversi decenni, ma che restano attorniati di vago si sono moltiplicati nell'azione pubblica. […] vaghezza delle posizioni all'interno dell'organizzazione, vaghezza degli statuti professionali e infine vaghezza di identità e di mestiere.

[139] Cfr. Florent Champy, *La sociologie des professions*, in «Quadrige», Parigi, PUF, 2009.

[140] Cfr. Jérôme Denis, David Pontille, *Travailleurs de l'écrit, matières de l'information*, in «Revue d'anthropologie des connaissances», vol. 6, n. 1, 2012, pp. 1-20.

[141] Cfr. Valérie Boussard, Didier Demaziere, Philip Milburn, *L'injonction*

tratta quindi dell'aggiunta di nuove funzioni ad attività documentarie classiche? Sarà necessario condurre le competenze di queste professioni al cuore del *savoir-faire* delle istituzioni culturali e, perciò, ripensare le diverse professioni? I lavori accademici su questi attori sono ridotti ed è importante rappresentare la realtà delle loro pratiche.

In conclusione, pensiamo che altre considerazioni, prima trascurate, intervengano ormai in ciascuno degli studi sugli utilizzi dei dispositivi d'accesso all'informazione: la questione etica (con la tracciabilità degli utenti, lo scambio, la conservazione e la commercializzazione dei dati personali), la questione economica (pubblicità online, commercializzazione accresciuta del legame sociale, monetizzazione del contesto attraverso la geo-referenziazione, ecc.), la questione pedagogica (apprendimento e cultura dell'informazione), la questione relativa all'identità (creazione di comunità online, ridefinizione delle identità, ecc.), la questione politica (la censura nell'indicizzazione dei siti da parte dei motori di ricerca, il rispetto della diversità culturale, ecc.), le sfide giuridiche (la pubblicazione online di opere protette, l'accesso a contenuti illeciti attraverso i motori specializzati, ecc.). La messa in evidenza di queste sfide socio-politiche a livello di individuo, di organizzazione o della società permette di riflettere sui rapporti di potere nati dalla generalizzazione dei dispositivi di accesso e di diffusione dell'informazione.

Conclusioni

Dall'analisi delle indagini sull'inserimento dei professionisti dell'informazione, affiora un'esigenza di ricostruzione dell'iden-

au professionnalisme. Analyses d'une dynamique plurielle, Presses Universitaires de Rennes, 2000.

tità professionale, una trasformazione delle condizioni d'esercizio, una riformulazione delle formazioni esistenti e un'incertezza sui contorni dei mestieri emergenti (webmaster, redattori di contenuti, *veilleur*[142], *knowledge manager*, *record manager*, gestori di *e-reputation*, curatori, community manager, ecc.), poiché la definizione di queste nuove professioni è instabile. Il bisogno di formazione continua e di un accesso alla ricerca scientifica nel campo è indispensabile.

La digitalizzazione crescente delle attività umane, commerciali e non, e l'evoluzione delle professioni rendono progressivamente inefficace l'opposizione tra l'informazione per il grande pubblico prodotta dagli organi di stampa e l'informazione scientifica e tecnica (Polity, 2000)[143] (Miège, 2004)[144]. Attraverso il presente capitolo abbiamo delineato un panorama delle professioni dell'informazione e delle ricerche nell'ambito delle scienze dell'informazione in Francia focalizzandoci sui lavori francesi degli ultimi dieci anni relativi a due importanti tematiche: il documento digitale e la mediazione. Abbiamo precisato che le formazioni e le ricerche sulla documentazione e sulle bi-

[142] «*Le veilleur stratégique mène des opérations de surveillance et d'action sur l'environnement de l'entreprise. Autrement dit, il anticipe les tendances du marché sur différents champs pour procurer un avantage concurrentiel à l'entreprise. En effet, les informations qu'il récolte et traite sur le web sont une base pour la définition des orientations stratégiques de sa société*».
<http://www.metiers.internet.gouv.fr/metier/veilleur-strategique>.
Il *veilleur stratégique* conduce operazioni di *sorveglianza* e di azione nell'ambiente aziendale. In altre parole, egli anticipa le tendenze del mercato su diversi fronti per procurare un vantaggio concorrenziale all'azienda. Infatti, le informazioni che raccoglie e tratta sul web costituiscono una base per la definizione degli orientamenti strategici della propria società.

[143] Cfr. YOLLA POLITY, *Information I versus Information II*, 2000.
< http://www.iut2.upmf-grenoble.fr/RI3/Information.htm>.

[144] Cfr. MIÈGE, B., *op. cit.*, 2004.

blioteche rientrano ampiamente nel campo delle scienze dell'informazione e della comunicazione. L'articolazione tra informazione e comunicazione diventa in tal modo il nostro orizzonte epistemologico (Miège, 2004)[145]. Riteniamo che sia fondamentale per un ricercatore in scienze dell'informazione coniugare una riflessione sugli artefatti e sulle loro iscrizioni culturali, sociali, giuridiche ed economiche. Concordiamo con (Metzger, 2002)[146] quando sottolinea che una ricerca in scienze dell'informazione, facendo parte delle scienze umane e sociali, debba articolarsi intorno a tre poli: il sapere formalizzato e il relativo flusso (e quindi le problematiche legate al supporto documentale, i sistemi di trattamento dell'informazione, ecc.), l'umano e il sociale (usi e pratiche) e infine la formalizzazione e il calcolo (quindi aspetti algoritmici).

I dispositivi informativi e i documenti hanno indubbiamente vincoli intrinseci, ma si inseriscono in un ambiente sociopolitico che li precede (Guyot, 2006)[147]. La particolarità della nostra disciplina scientifica sarà identificata al livello delle ibridazioni e delle articolazioni tra la tecnica e l'umano, e tra l'individuo e il gruppo. Noi ipotizziamo che lo studio del rapporto tra l'utente e il suo ambiente socio-economico sia una problematica cruciale, al pari dell'articolazione di questo rapporto con i processi cognitivi. I fattori cognitivi, quindi, non saranno analizzati solo in quanto tali, ma anche in relazione con le numerose variabili legate alla dimensione fondamentalmente sociale e culturale degli utenti.

---

145 Cfr. MIÈGE, B., *op.cit.*, 2004.

146 Cfr. JEAN-PAUL METZGER, *Les trois pôles de la science de l'information*, in Recherches récentes en sciences de l'information: convergences et dynamiques, Couzinet V., Régimbeau G. (a cura di), in «Sciences de l'information - série Recherches et documents», ADBS Éditions, 2002.

147 Cfr. BRIGITTE GUYOT, *Dynamiques informationnelles dans les organisations*, Parigi, Lavoisier, 2006.

Sciences du document – Sciences de l'information

Introduction

Les enjeux et les opportunités scientifiques, politiques, économiques et culturelles que suscite la révolution numérique sont importants. Ces évolutions ont brouillé les repères dans le champ des sciences de l'information et de la communication tant sur le plan des métiers que de la recherche. Les professionnels de l'information sont confrontées à une nouvelle écologie du savoir qui les mets en concurrence avec de nouveaux acteurs mais qui leurs offrent de nouvelles opportunités (Miège, 2004). La question des métiers et du socle de compétences se pose à nouveau. Paradoxalement, l'importance prise par le numérique n'a pas permis pour l'instant de consolider ce champ professionnel et de recherche.

Les métiers de l'information sont divers. Ils concernent aussi bien la documentation, la bibliothéconomie que la gestion des archives. Historiquement (Accart, 2000), l'enseignement de la documentation et de la bibliothéconomie est éclaté entre l'offre universitaire et celle des grandes écoles et instituts (l'Ensb devenue Enssib; l'Institut National des Techniques Documentaires et l' école des bibliothécaires-documentalistes rattachée à l'institut catholique de Paris). Les unités de formation et de recherches (UFR) délivrent des formation de niveau master. La formation à la recherche est assurée par des écoles doctorales. A ce panorama, il faut ajouter les formations courtes délivrées par les Instituts Universitaires de Technologies (Baccalauréat + 2 années) et les Instituts Universitaires Professionnalisés (bac + 3 années) sans oublier les préparations aux formations de professeur-documentaliste Ces acteurs de l'informations sont affiliés à diverses associations professionnels (Adbs[1], ABF[2], Fadben,[3] Adbu[4], AAF[5]...).

[1] Association des professionnels de l'information et de la documentation (http://www.adbs.fr/)
[2] Association des bibliothécaires français
[3] Fédération des associations des enseignants- documentalistes de l' Éducation Nationale (http://www.fadben.asso.fr)
[4] Association des directeurs & personnels de direction des bibliothèques universitaires et de la documentation (adbu.fr)
[5] Association des archivistes français (http://www.archivistes.org/)

Dans le domaine des archives, l'école nationale des chartes, crée en 1821, assure la formation des cadres de la conservation du patrimoine national. Depuis le début des années 80, la formation des archivistes s'est diversifiée. Les universités proposent désormais plus de 15 formations de niveau master dans les métiers des archives et du patrimoine (Defrance, 2002). Ces formations sont souvent assurées dans les départements d'histoire ou des SIC. Il s'agit de deux conceptions, que nous estimons complémentaires, du rôle des archivistes et de leur organisation. Pour certains, l'archiviste est d'abord un conservatoire de la mémoire historique. La connaissance de l'histoire est primordiale. La seconde vision tend à rapprocher la formation des archivistes de celles des bibliothécaires-documentalistes (Favier, 1993). Les étudiants issus de ces formations présentent les concours de la fonction publique territoriale ou intègrent le privé (Defrance, 2002).

Alors que les trois secteurs de métiers s'appuient sur des savoirs théoriques et des compétences équivalents, Palermiti et Polity (2002) montrent que les professions de l'information sont éclatées et qu'il n'y a pas en France, depuis l'après-guerre, de vision globale et unifiée de ces professions. Elles constatent la quasi-absence du monde des bibliothèques et des archives au niveau de la recherche en sciences de l'information. Ces professions, selon ces deux auteures: «*se distinguent par des discours et des images stéréotypées qui reflètent davantage une appartenance aux lieux d'exercice qu'à celle d'une même famille professionnelle...Les raisons de ces divergences et de ces oppositions s'expliquent historiquement. Elles sont liées particulièrement à des politiques publiques incohérentes, tant au niveau des statuts des personnels qu'au niveau d'un système de formation centralisé, fermé sur lui-même et peu progressif, créant notamment une relative confusion chez les employeurs et une absence de clarté dans la carte des formations*».

Si les métiers d'archivistes, de bibliothécaires et de documentalistes sont devenus disparates, la réflexion sur leur avenir ne peut qu'être commune (Melot, 2005). Ce dernier estime qu'à l'avenir, ces métiers pourraient bien être appelés à se rejoindre. Ce rapprochement est favorisé par un contexte qui ne remet pas en cause les spécificités des publics (Wiegandt, 2005). La majorité des professionnels de l'information assurent selon Fondin, et Rouault, (1998) soit des fonctions d'intermédiation documentaire ou culturelle entre des usagers et des ressources informationnelles, soit des fonctions patrimoniales de conservation de tous les do-

cuments. Ces métiers ont évolué en intégrant de nouvelles compétences liées à l'environnement de l'entreprise, à l'émergence de nouveaux besoins des usagers et à la généralisation des technologies de l'information et de la communication (TIC). Ces auteurs incluent dans cette catégorie les métiers en amont du traitement de l'information (éditorialisation, gestion de site web, gestion des bases de données documentaires) ou plutôt en aval de l'exploitation de l'information en élargissement les métiers de la documentation à ceux de la veille et l'intelligence économique mais aussi de la gestion du contenu d'entreprise et au *knowledge management* (CNE, 93).

L'évolution[6] des métiers de l'information est constamment interrogée par divers enquêtes de l'ADBS, de l'ANPE (Agence Nationale pour l'Emploi devenue Pôle Emploi) ou de cabinets de conseil (Serda, Cepid, Histen Riller). De ces études, il ressort d'abord une incertitude sur le marché de l'emploi des professionnels de l'information. Lebigre (2011) citant des statistiques de l'Insée (Institut national de la statistique et des études économiques) et les enquêtes du cabinet Serda indique qu'entre 35000 et 80000 professionnels de l'information exercent en France aujourd'hui. Un tiers d'entre eux exerce dans le secteur privé, généralement des documentalistes. Parmi les points de convergence de ces travaux, Lebigre (2011) et Stiller, (2001) citent la dilution des fonctions information-documentation au sein des organisation puisqu'un tiers des répondants travaillent d'une façon autonome, en dehors d'un centre de documentation. Prés d'un tiers des entreprises ne disposent plus de service de documentation (Lebigre, 2011). Lorsqu'il existe, on observe une diminution de la taille des structures en sachant que la situation est hétéroclite, varié selon les secteurs d'activité. Ainsi selon Lebigre (2011) «Dans certaines entreprises, la fonction information-documentation est parfaitement reconnue: en plus des fonctions classiques, on lui confie des missions de veille, on l'associe au processus de gestion des connaissances, on la fait participer au déploiement des NTIC. Dans d'autres, à l'inverse, cette fonction est diluée au sein des services, où aucune structure propre ne la prend en charge». Les évolutions économiques (crise, mondialisation) couplées à la révolution numérique sont, selon Michel, (2011 et 2003) à l'origine de

[6] Pour un panorama des qualifications professionnelles en Europe, voir Mahon, B., The disparity in professional qualifications and progress in information handling: a European perspective. «*Journal of information science*», Vol 34, n 4, 2008, pp. 567-575

la mise en cause des modalités d'interventions professionnelles de la documentation. Pour Jean Michel (2003) «c'est le modèle de centralisation à l'œuvre dans les centres de documentation» développés dans le prolongement de la première informatisation des pratiques documentaires (années 1970-1995) que les mutations actuelles rendent obsolètes. La chaîne documentaire issue de ce modèle apparaît aujourd'hui inadaptée. Ces enquêtes montrent aussi une évolution des activités au sein des fonctions. Les activités liées au catalogage, à la collecte de l'information décroît au profil de l'analyse et la valorisation des documents (Lebigre, 2011) (Stiller, 2011). Cette diversification des métiers de l'information est un des enjeux de l'insertion professionnels. Certaines fonctions éditoriales, de communication autour du web, de gestion stratégique de l'information complètent la panoplie des métiers traditionnels (Lebigre, 2011). Le défi pour les professionnels de l'information est de montrer que leurs compétences demeurent applicables au-delà du centre de documentation traditionnel (Stiller, 2011). Parmi les perspectives identifiées par les différentes enquêtes, deux en ressortent. La première est de maîtriser les outils, méthodes et organisation des métiers du web. La seconde est la compréhension des besoins propres aux métiers de l'entreprise. Les professionnels de l'information allient des compétences spécifiques en science de l'information à une connaissance souvent approfondie d'un secteur d'activité

Après avoir présenté la situation des métiers de l'information en France, nous nous intéressons dans ce qui suit à la situation de la recherche en science de l'information en questionnant les fondations théoriques et académiques de cette jeune discipline.

1. De la documentation aux sciences de l'information

1.1 *Fondations théoriques*

Alors même que l'on doit à l'Union Française des Organismes de Documentation (créée en 1932) l'idée nouvelle d'une «science de l'information» dès l'entre-deux-guerres (Fayet-Scribe 2000, note 27 p. 52), il faudra attendre les années 1970 pour qu'elle ait une «existence académique» en France au sein d'un domaine plus vaste «les sciences de l'information et de la communication». C'est aux Etats-Unis seulement que la science de l'information et des bibliothèques *(Library and Information*

Science) pourra se développer de manière autonome à l'université même si, à son tour, dans les années 1990 elle verra son importance décroître au profit d'autres dénominations éludant la référence à la bibliothèque (*School of Information, school of information sciences, Ischools etc.*).

Les fondations théoriques des sciences de l'information se sont élaborées à l'intérieur d'un projet politique. En effet, le concept de *documentation* et le sens moderne d'*information* vont s'adosser à la construction de la Société des Nations (SDN) et en son sein, à la coopération intellectuelle internationale[7], grâce au travail remarquable que firent les deux avocats belges, Paul Otlet et Henri Lafontaine. Si la Société des Nations et le pacifisme de cette période se soldèrent par un échec politique et une seconde guerre mondiale, le travail de Paul Otlet continue à interpeler les contemporains par sa modernité: la séparation du contenu et du support, l'idée de l'hypertexte et la préfiguration du Web, la notion de «langage documentaire», la médiation documentaire. Les trois premiers thèmes sont contenus dans le concept moderne de «documentation» et le dernier dans un nouveau rapport au savoir qui passe par l'*information* telle qu'elle va être définie dans ce contexte «pré-informatique».

Nous verrons ainsi comment les fondations théoriques des sciences de l'information et de la documentation se sont construites à partir du double héritage du Répertoire Bibliographique Universel de Paul Otlet et d'une alliance avec l'étude la communication propre à une conception universitaire française du domaine.

I-1.1 L'héritage du Répertoire Bibliographique Universel

Le concept de documentation

Le projet de Répertoire Bibliographique Universel (RBU) que forgèrent les deux avocats est un renouvellement de l'encyclopédisme du siècle des Lumières, en ce qu'il vise à rassembler «la science vivante»[8], tout en créant un système nouveau où il s'agit, non plus de synthétiser le

[7] En 1921 est créée une Commission internationale de coopération intellectuelle par l'assemblée de la Société des Nations et, en 1926, l'Institut international de coopération intellectuelle qui voit le jour à Paris la rend permanente. Cette commission est considérée comme les prémisses de ce qui deviendra l'UNESCO après la guerre alors que la SDN deviendra l'Organisation des Nations Unies.

[8] Sur cette expression, voir Françoise Levie, *L'homme qui voulait classer le monde*. Bruxelles, *Les impressions nouvelles*, 2006.

savoir mais de donner accès aux publications originales qui l'ont formulé. Ainsi «*L'ensemble de tous les écrits pourra, en un certain sens, être considéré comme formant un seul grand livre, un livre aux proportions formidables, aux chapitres en nombre quasi illimité*» (Otlet 1903, «Les sciences bibliographiques et la documentation», *Bulletin de l'Institut International de Bibliographie*). La documentation est à mi-chemin entre l'encyclopédie et la bibliographie: elle est à la fois ce «*grand livre*» et l'ensemble des travaux qui le compose. Mais elle ne se réduit ni à l'un ni à l'autre car elle n'est ni une entreprise de vulgarisation scientifique (de reformulation des travaux), ni un inventaire des publications existantes (la bibliographie). Elle est «*la carte immense des domaines du savoir, avec tout le complexe des divisions et des subdivisions de leurs territoires*» grâce à laquelle «*nous pourrions localiser tout naturellement chacun des travaux dans quelqu'une des circonscriptions. Nous les verrions s'y rattacher aux travaux similaires pour compléter ce qui furent écrits antérieurement, et à leur tour servir de lien entre ces données du passé et les progrès de l'avenir*» (*ibid*).

Le RBU est une innovation à plusieurs points de vue. Sur le plan matériel, il donne lieu au meuble à tiroir avec fiches mobiles que nous avons connu jusqu'à l'informatisation des catalogues. Cette manière de concevoir la documentation s'oppose à la présentation de la bibliographie sous forme de volumes cumulatifs. Sur plan organisationnel, le RBU impose la standardisation des matériels utilisés comme des méthodes de travail. La documentation rompt là encore avec la tradition bibliographique qui restait une entreprise individuelle : elle applique les méthodes industrielles du temps telles qu'elles donnent naissance aux premières agences de normalisation. Otlet et Lafontaine étaient d'ailleurs des admirateurs de Ford et du fordisme. La coopération internationale qui s'incarne dans l'Union des Associations Internationales se fait grâce à cet effort de normalisation qui restera une préoccupation majeure de la documentation. Sur le plan fonctionnel, le RBU est un système de recherche d'information et non seulement un classement raisonné des livres dans les bibliothèques. En cela il préfigure le Web comme l'ont signalé plusieurs auteurs et personnalités depuis Boyd Rayward[9] jusqu'à Vinton Cerf (inventeur du protocole

[9] Rayward Boyd, W. 1994. «Visons of Xanadu: Paul Otlet (1868-1944) and Hypertext», Journal of the American Society for Information Science, vol.45. Rayward Boyd, W.

TCP/IP) et Thierry Geerts, (directeur de Google Belgique)[10]. Le lien entre les connaissances se fait par la Classification Décimale Universelle (la CDU). La CDU n'est pas seulement une technique de classement. Elle est conçue comme une véritable langue universelle: «*Créer une classification synthétique avec notation concise des idées c'est doter l'esprit d'une véritable langue écrite universelle capable d'agir puissamment sur la forme elle-même de la Pensée*». (Otlet 1935 *Monde. Essai d'universalisme*). C'est la différence avec la Classification Décimale de Dewey (CDD) dont elle s'inspire pourtant directement. La CDU introduit un paradigme nouveau malgré la continuité avec l'entreprise de Dewey. Elle transforme le schéma strictement hiérarchique et énumératif de la CDD en une langue pouvant exprimant des liaisons d'une autre nature entre les sujets (introduction du principe des facettes par les tables auxiliaires et définition de symboles de relation notamment). La CDU devient ainsi une sorte d'*esperanto* dans lequel tout sujet doit pouvoir être codifié en lien avec le système universel des sciences.

Le schéma «Les étapes de l'organisation documentaire» résume ce nouveau concept de la documentation.

Image 1. ©*Archives Mundaneum*

2003. «Knowledge organization and a new world polity: the rise and fall of the ideas of Paul Otlet». *Transnational Associations* 1-2/2003, p. 4-15.

[10] Voir Le Monde 03/11/2012 «Le Web, une histoire belge» citant V.Cerf au World Science Festival de New York et T.Geerts.

Cette entreprise destinée à rassembler le savoir du monde nécessite de nouvelles définitions

> «*Nous entendons par le terme général information les données de toute nature, faits, théories nouvelles, qui, parvenus à l'intelligence humaine, constituent des notions, des éclaircissements, des directives pour la conduite et l'action; d'autre part, nous entendons par documentation l'ensemble des moyens propres à transmettre, à communiquer, à répandre les informations, (livres, périodiques, catalogues, textes et images, documents de toutes formes)*» (Otlet 1917 «L'information et la documentation au service de l'industrie», *Bulletin de la Société d'encouragement pour l'industrie nationale)*.

Un nouveau rapport au savoir: du RBU au Mundaneum

Le RBU permet de concevoir la mise à jour perpétuelle des connaissances, l'«éphérémographie» selon le principe posé par Charles Limousin en 1900 (voir Fayet-Scribe 2000 p. 96-97). Les bulletins de sommaires, les analyses d'articles, «la revue à découper» (*ibid*) deviennent des enjeux pour accéder rapidement à l'information et éviter la perte d'argent. Le RBU et les associations internationales qui y contribuent répondent à ces besoins de la société industrielle en plein développement. Mais, au-delà de la mise à jour, c'est l'idée que la documentation implique une médiation entre les lecteurs et les documents qui se fait jour. Cette idée est présente dans les textes d'Otlet concernant la création d'un Office de Documentation industrielle qui qualifient les agents de cet office d'«*intermédiaires vivants entre le public et les documents*»:

> «*L'Office est actif tandis que la bibliothèque est passive; il repose sur l'idée que les personnes auxquelles il est destiné doivent être amenées à agir dans une direction donnée et, pour cela, qu'elles doivent être incitées à la faire, que l'Office doit les aider, de toute manière, dans leur effort, pour se représenter les choses, non pas d'une manière quelconque mais le plus exactement possible. Pour cela, les agents de l'Office sont des intermédiaires vivants entre le public et les documents. Les questions leur sont posées sous la forme concrète des cas appliqués. Ne pouvant tout connaître ni tout retenir, ces agents doivent pouvoir se retourner vers des sources autorisées (…), des livres, des dossiers, des répertoires contenant des données antérieurement élaborées (…)*» (Otlet 1917 «L'information et la documentation au service de l'industrie», *op.cit*.).

Que la médiation documentaire soit incarnée par de nouveaux métiers tels ces agents que l'on appellerait aujourd'hui «documentalistes» ou par des moteurs de recherche et autres outils (annuaires, signets, portails…) que nous connaissons aujourd'hui, le règne de l'information et du document implique une médiation de plus en plus complexe pour accéder aux ressources et imaginer les parcours de recherche des usagers qu'Otlet appelle le «public».

Plus encore que l'accès à l'information, c'est l'éducation qui est au cœur du projet de la documentation. Il imagine l'intégration de la documentation dans un projet plus vaste, le Mundaneum, ou Palais mondial, qui rassemblerait un centre de documentation, une bibliothèque, un centre de culture scientifique, un musée, une bibliothèque internationale et se déclinerait en Mundaneum locaux (convergeant vers un Mundaneum central) grâce aux associations internationales:

> «*De même que les États ont organisé la Société des nations et les Unions officielles des gouvernements qui en dépendent, de même les Associations internationales ont à se relier entre elles en une Union des Associations internationales et à faire du Mundaneum leur œuvre capitale*» *(idib Otlet 1935). «Il est devenu nécessaire d'établir des liens entre ces organisations et d'en faire les partie d'une institution générale: a) en prenant pour base les quatre grandes institutions intellectuelles qui ont noms Bibliothèque, Musée, Université, Académie, Société scientifique (...)» (ibid).*

La convergence que nous observons aujourd'hui entre documentation, bibliothèques, archives et musées, liée à la numérisation des œuvres, se rapproche de cette conception prémonitoire. Ne pouvant imaginer un réseau virtuel comme lieu de cette convergence, il pense à une cité mondiale où se déploie le *rete mundaneum* illustré dans la figure ci-dessous (image 2).

Le RBU n'est donc qu'un élément d'une utopie plus complexe orientée vers l'éducation tout au long de la vie, thème qu'on ne peut imaginer plus contemporain:

> «*le Mundaneum repose sur le principe que l'éducation n'est pas simplement utilitaire. (...) L'éducation continue s'impose donc à chacun comme un devoir individuel et social de perfectionnement, développé et continué à tout âge*» (Otlet 1935 *op.cit.*).

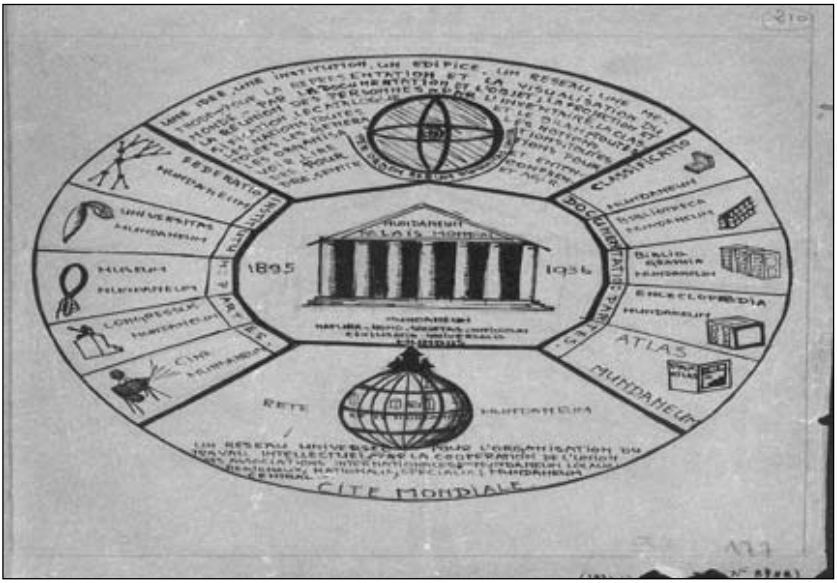


Image 2. ©Archives Mundaneum

Telle est l'ambition politique, à la fois pacifiste et universaliste, qui donna le jour aux fondations théoriques des sciences de l'information que enseignons aujourd'hui à l'université même si la révolution cybernétique, la naissance de l'informatique et la généralisation de la numérisation ont semblé en avoir été les principaux instigateurs.

La naissance académique des sciences de l'information et de la communication (SIC) en France n'ont pas revendiqué prioritairement cet héritage. Quelle que soit l'interprétation que l'on peut donner à cette réunion très particulière à la France de l'information et de la communication, c'est plutôt à l'héritage du modèle de la théorie mathématique de la communication de Shannon et Weaver (le modèle et ses critiques) que l'on se réfère pour théoriser l'unité de l'information et de la communication[11].

[11] Contrairement à ce qu'affirment plusieurs auteurs. Voir par exemple Ibekwé-San-Juan 2012 (JASIST, September 2012, «The French Conception of Information Science») In-

I-1.2 *De la documentation aux sciences de l'information et de la communication*

Selon Fondin (2006), la théorie mathématique de la communication sert de référent théorique aux sciences de l'information (SI), non par nécessité conceptuelle, mais par souci de donner une allure scientifique à la théorisation de l'«activité documentaire»:

> «*Pour la quasi-totalité des chercheurs en SI, le référent théorique, parce qu'il faut un référent théorique si l'on veut faire "scientifique», est, implicitement et même très souvent explicitement, la théorie de l'information de Claude E. Shannon. L'activité documentaire est assimilée à un codage/décodage des signes d'un message qu'il faut réaliser dans les meilleures conditions techniques. (…) On est là dans un modèle scientifique tel que le XXème siècle l'a forgé et imposé comme modèle à toutes les sciences, qu'elles soient exactes ou sociales, autrement dit quel que soit le type de phénomènes à étudier*» (Fondin, 2006).

La science de l'information anglo-saxonne n'a pas eu besoin de ce référent théorique pour trouver un modèle de scientificité. Si elle a réuni, dès les années 1950, ingénieurs et spécialistes de l'organisation des connaissances (indexation, langages documentaires) issus de divers horizons, c'est finalement pour créer un domaine spécifique: les sciences de l'information et des bibliothèques.

En France la construction des SIC se fait, selon de nombreux auteurs, au profit d'une prépondérance de la communication, dont la théorie mathématique de Shannon apparaît souvent comme le point d'ancrage naturel, fut-ce pour le critiquer. Concernant cette prépondérance, Courbières (2010) affirme même que «les sciences de la communications» jouent le rôle de «science-mère», y compris pour ceux qui, comme J. Meyriat, ont développé le champ des SI en France. Ainsi Meyriat (1993) fait-il de la communicologie la science à partir de laquelle se développent les autres branches (medialogie, informatologie, communicologie fonctionnelle). La communicologie se définit comme l'étude de «la com-

deed, there is a widespread belief among members of the academic community in France that the Anglophone conception of information science is very different from theirs, in that it is rooted mainly in Shannon's mathematical theory of communication Voir aussi (Fondin, 2006)

munication sous tous ces aspects», (cité par Courbières *op.cit.* p. 64) et n'inclut aucun champ spécifique pour les SI. La documentologie (objet de la medialogie comme les études sur la presse et la communication de masse) se distingue de la science du contenu sémantique de la communication (informatologie) et de la communication fonctionnelle (fonctionnement des systèmes de communication remplissant des fonctions particulières (dont la communication scientifique et la publicité). Bref, ce qui relève des sciences de l'information aujourd'hui dans les cursus universitaires se trouve ici éclaté dans les différentes branches de l'étude de la communication.

Un autre fondateur éminent des SIC, Robert Escarpit, affirme cette primauté de la communication. Son ouvrage *L'Information et la Communication. Théorie générale* (1991) couvre un spectre large qui s'étend de la théorie de Shannon et le «rêve cybernétique» à l'étude de la culture de masse en passant par l'analyse du langage («Language et languages», «La communication et l'événement») et celle de la documentation. Deux chapitres concernent le champ des sciences de l'information: «L'information et le document», «Les problèmes documentaires». La communication est le point d'ancrage de cette théorie. Elle est définie comme «un cas particulier du transport», celui de l'information. Les limites du modèle thermodynamique et entropique se font jour dès qu'elles sont confrontées au vivant et à la pensée (Escarpit, *op.cit.* p. 23-24). «Un nouvel outil épistémologique devient nécessaire. Cet outil est la notion d'information». «La communication est un acte et l'information est son produit»: qu'il y ait échec ou succès de cet acte, il est également productif d'information. C'est une différence avec le rendement d'un système télégraphique. Dans ce cadre général, la notion de *document* va se définir par rapport aux spécificités du vivant et du temporel. Un système de communication n'existe comme tel que s'il est perçu par une «personne» humaine, «c'est-à-dire par une entité autonome, à la fois mécanique, vivante et pensante, toujours en lutte contre le temps. Toute description humaine d'un tel système tend donc à se soustraire à l'événement (...) ou tout au moins à le prévoir, à le dominer, à le manipuler». Le document est alors défini comme un «anti-événement». Il n'est pas une inscription mémorielle qui saute fugacement d'un message à un autre mais il est la forme matérielle de la «mémoire des données», mémoire à la fois individuelle et culturelle. Un «pas décisif» a été franchi quand l'homme «a institué le document, cumulation de traces fixes et permanentes où les réponses données en *feed-*

back à travers le temps restent disponibles pour une lecture (...)». Les «problèmes documentaires» sont précisément ceux qui traiteront de ce *feedback*. Il n'y a donc pas matière à une science de l'information mais il ya bien une place pour la documentation à l'intérieur des SIC.

En créant les SIC, les français se sont davantage tournées vers l'étude de la communication sous toutes ses formes que vers l'héritage historique de la documentation et de l'idée ds science de l'information inventée par l'UFOD au début du XXème siècle.

1.2 Fondations académiques

Il n'y a pas d'équivalent des «*library and information sciences*» anglo-saxonnes en France. Les sciences de l'information constituent à l'origine un point de rencontre entre des professionnels de l'information et des chercheurs (Couzinet, 2002). Conçue, selon B. Olivier (2005), comme un phénomène sociotechnique, «l'information est désormais quelque chose qui circule et se voit traduit dans le cadre de médiations et d'interactions complexes qui dépasse largement la conception purement technique dans laquelle on l'a longtemps isolée». Les recherches issues des sciences de l'information et de la communication relèvent à l'étranger de disciplines telles des sciences de l'information, d'études des médias, des sciences de la communication, des études culturelles, de la sociologie, de sciences politiques, d'études de littérature ou de la sémiotique. (Jeanneret, 2001).

Les travaux en documentation sont intégrés dans la section 71e du CNU (Conseil national des universités): les Sciences de l'Information et de la Communication. Cette section a été officiellement créé en tant que discipline académique en Janvier 1975 par des personnalités éminentes telles Robert Escarpit, Jean Meyriat, Roland Barthes et Fernand Terrou pour donner un cadre universitaires à un ensemble de formations hétéroclites tels l'histoire du livre, la bibliothéconomie, ou le journalisme. En 1978, la Société française des sciences de l'information et de communication (Sfsic)a été fondée. Toutefois, il est important de rappeler ici, les recherche sur l'informatique documentaire et les systèmes d'organisation de connaissances menées dés les années cinquante par des pionniers tels que Gardin, Pages, Levy (Boure, 2002)

Ainsi sont considérés comme relevant des SIC, selon le CNU les travaux (CNE, 1993):

- «Les études sur les notions d'information et de communication, sur

leurs relations, sur la nature des phénomènes et des pratiques ainsi désignés, de même que les différentes approches scientifiques qui s'y appliquent».
- «L'étude, d'une part, des processus, des productions et des usages de l'information et de la communication, d'autre part, de la conception et de la réception de celles-ci. Ainsi que l'étude des processus de médiation et de médiatisation».
- «L'étude des professionnels (dont notamment les journalistes) et de leurs pratiques».
- «L'étude de l'information, de son contenu, de ses systèmes sous l'angle des représentations, des significations ou des pratiques associées».
- «L'étude des médias de communication et des industries culturelles sous leurs divers aspects».

La particularité de ces recommandations est de prendre en compte l'ensemble de la chaine d'information et de sa communication. La création de la section a permit d' institutionnaliser la recherche académique en donnant aux universités la possibilité de délivrer des doctorats de troisième cycle constituant ainsi un vivier de chercheurs (CNE, 1993).

Au niveau des publications, la section du 71e CNU a une liste des revues qui sont reconnues comme des revues de référence dans la discipline de l'information et de communication. Les revues sont en deux catégories: la première comprend les revues académiques arbitre principalement sur les sciences de la communication, et le second contient des revues professionnelles qui publient des articles scientifiques. Afin d'évaluer la production scientifique des chercheurs, l'Aeres a intégré des revues internationales en sciences de l'information ainsi que deux revues professionnelles publiées en France: Documentaliste - Sciences de l'information créé en 1964 et publié par l'association Adbs, et le BBF (Bulletin des Bibliothèques de France) créé en 1956 et publié par l'Enssib.

Analysant la nature des articles parus dans la revue Documentaliste-Sciences de l'information depuis 1964, Couzinet (1999) distingue trois périodes: la période (1964-1976), où les articles sont rédigés presque exclusivement par des professionnels, puis la période 1976-1989, lorsque les chercheurs commencent à publier dans le journal (1 à 5 articles par an), enfin la période commençant en 1990, lorsque le nombre d'articles publiés par les chercheurs ont tendance à égal à celui des articles écrits

par des professionnels. Une analyse de 318 articles publiés par le BBF (Couzinet, 1999) pendant quatre ans montre que 52 articles (16,35%) ont été publiés par les chercheurs. En ce qui concerne les thèses, Polity (2001) montre à travers l'examen de 90 thèses de doctorat soutenus en sciences de l'information pour la période 1971 à 2000 que la majorité des des thèses portent sur l'intelligence économique, la sociologie de la lecture et l' informatique documentaire. En s' appuyant sur la distinction opérée par Richard Whitley entre l'institutionnalisation cognitive et l'institutionnalisation sociale Polity (2001) constate «qu'il n'y a pas eu pour les sciences de l'information de corrélation entre une certaine institutionnalisation cognitive amorcée dès le début de ce siècle, et leur reconnaissance universitaire qui est un des facteurs académiques déterminant de l'institutionnalisation sociale. Cette dernière s'est faite en l'absence d'une fraction importante du monde de l'information, celle du secteur des bibliothèques et des archives, que les sciences de l'information n'ont pas su ou pas pu intégrer». Il est important de signaler que la mise en œuvre de reformes structurelles dans les universités françaises (création de l'Aeres[12], baisse des budgets, individualisation de l'évaluation, modification du statut des chercheurs, concurrence intra et interuniversitaire,) ont modifié les conditions d'accès et de diffusion des connaissances scientifiques (Devroey, 1999). Les récents changements institutionnelles notamment les procédures d'évaluation normées par l'Aeres peuvent aboutir à nouveaux modes d'argumentation mais aussi une évolution des modes d'évaluation «scientifique» dans notre discipline.

D'un point institutionnel et en se référant aux travaux de (Cardy et Froissart, 2002), (Cardy et Froissart, 2006) la recherche en sciences de l'information et de la communication est un domaine en plein essor et de plus en plus reconnu en France. Le nombre de recrutement et d'offres de poste est de plus en plus important. De 1977 à 2005 le nombre de postes a augmenté annuellement de 10%, à partir de 43 enseignants-chercheurs en 1977 à 663 en 2005. Toujours, selon Cardy & Froissart (2006) «le nombre de postes universitaires en Infocom a dépassé les domaines tels que la philosophie et les sciences politiques, et est proche de la sociologie et la linguistique» l'analyse des profils de poste sur la période (1995-2001) montre selon (Cardy & Froissart, 2006) que la documentation et les

[12] Agence nationale d'évaluation de l'enseignement supérieur et de la recherche

NTIC représentent (23%) des postes. Les thématiques de recherche des enseignants-chercheurs en information-documentation (Fondin& Rouault,1998) (CNE, 1993) portent sur la représentation des connaissances et l'organisation du savoir, la communication homme-machine (analyse des besoins des usagers, analyse des pratiques informationnelles...); l'instrumentalisation des processus (conception, développement et évaluation des dispositifs informationnels...).; sur l'information et sa gestion (travaux sur la veille et le management de l'information et des institutions...) et enfin l'information dans la société (économie de l'information, droit de l'information, fracture numérique...). Le succès de l'internet a suscité des travaux qui questionnent à nouveau l'information. Nous relatons dans ce qui suit, les récents travaux français sur deux thématiques importantes en relation avec les évolution des métiers et de la discipline.

2. Défis et perspectives pour les sciences de l'information

L'univers documentaire à partir duquel s'est stabilisé un corpus théorique, des outils, des méthodes et normes durant un siècle est actuellement bouleversé par le numérique. Cette r(é)volution du numérique a conduit certains à parler de redocumenatirisation (Salaun, 2007), d'aucuns disent bibliothécarisation et/ou «googlisation» du monde (Bazin, 2006) puisque de plus en plus le traitement de l'information (qui a longtemps été un domaine propre aux bibliothécaires) n'est plus l'apanage des spécialistes de l'information et devient un enjeu de société au sens le plus large. Ce nouvel ordre documentaire accompagne des organisations sociales et idéologiques différentes d'où le besoin d'une critique de cette raison numérique (Paul & Perriault, 2004)

2.1 *Penser les métamorphoses du document*

Les récents travaux menés en France sur le document, à partir des réflexions du collectif Roger T. Pédauque ont notamment souligné l'émergence de questions nouvelles concernant en particulier la pérennité, la temporalité, la métamorphose et l'hybridation du document et de l'espace documentaire. Sur le plan technologique, ces dix dernières années ont vu le développement de nouveaux modèles de recherche exploitant notamment la structuration des documents, une plus grande personnalisa-

tion des fonctionnalités, ainsi que l'intégration de la recherche d'information sur de nouveaux supports comme le PDA. Les technologies numériques associées à internet ont progressivement conduit à une déstructuration complète de la notion originelle de document mais non pas à sa suppression ou son effacement (Ihadjadene & Chaudiron, 2008). Ce processus est indéniablement lié à un contexte technologique particulier mais il est également induit par des pratiques informationnelles nouvelles (le dépôt des pré-publications dans les archives ouvertes par exemple) et une évolution des modèles économiques des grands éditeurs (la vente à l'article ou au chapitre)..

Dans le domaine de l'IST, (information scientifique et technique) une facette de ce changement est liée à la refondation des conditions de la recherche scientifique qui posent le problème de la place des spécialistes de l'information. En effet, les travaux actuels sur les e-sciences et les humanités numériques, questionnent la démarche scientifique et les moyens permettant de valider le processus de collecte, d'analyse, de traitement et de restitution des données. Au-delà du document numérique (article, monographie…) ou de la donnée brute, il s'agit dorénavant de proposer un accès aux protocoles d'expériences, aux collections de données et aux logiciels utilisés dans les expérimentations scientifiques. Se construit ainsi un continuum dans le «construit» scientifique où les données observées, les protocoles, les hypothèses, les inférences, les résultats, etc. s'articulent dans une structure en réseau. Pour le bibliothécaire ou le documentaliste, l'enjeu ne sera plus simplement de gérer la publication des résultats de la recherche mais de gérer les données de recherche elles-mêmes. Dans le domaine des SHS, (sciences humaines et sociales) des recherches ont été engagées depuis le milieu des années 1980 sur la modélisation a priori des textes scientifiques pour proposer une écriture multimédia adaptée à Internet qui tente à la fois d'assurer la manipulation aisée des constructions scientifiques et l'accès à l'ensemble des données qui les fondent (Gardin et Roux, 2004). Darnton cité par (Renoult, 2001) montre qu'Internet est en mesure de modifier en profondeur non seulement l'économie des revues mais également leur capacité d'argumentation scientifique. Il appelait à une transformation de la monographie scientifique en une hiérarchie de six niveaux (imprimé, couche pédagogique, niveau théorique ou historique, appareils critiques, les notes, les appendices, etc...). Cette analyse montre que la notion classique de monographie ou même de document est en train de se trans-

former et de se développer au profit de la notion évolutive de ressource numérique (Bachimont, 1999). Il ne s'agit pas seulement de mettre en ligne des documents mais de constituer une véritable éditorialisation des contenus numériques oeuvrant ainsi à nouvelles pratiques d'écriture dont le potentiel n'est pas toujours mis en oeuvre (Bachimont, 1999).

Enfin, se pose la question de l'appartenance d'un document à une collection au sens bibliothéconomique du terme. Nous ne nous appesantirons pas sur ce point qui a déjà été souligné à maintes reprises mais nous rappellerons seulement que, dans le champ de l'information scientifique et technique, cette dé-construction de l'«effet collection» est particulièrement forte. Wiegandt (1995) souligne la la nécessité pour les spécialistes de l'information de s'affranchir de la localisation et de la notion de possession au profit de la notion d'accès. Selon Wiegandt (95) cette évolution constitue un tournant majeur qui perturbe le rapport du professionnel à la collection. Les notions classiques traditionnelles de document, de fragment documentaire, de collection sont bouleversées par le numérique. Le support d'inscription et la forme sémiotique choisie contraignent selon (Bachimont, 1999) l'expression du contenu et ses conditions de réception, d'interprétation et de communication.

Enfin, penser les métamorphoses du document, implique une réflexion sur les possibilités de coopération et de convergence entre les archives, les bibliothèques et les musées (ALM selon l'acronyme anglo-saxon). Yarrow et al (2008) et Zorich (2008) dressent un état des coopérations et collaborations de ces institutions culturelles. Ces auteurs envisagent la collaboration comme un continuum qui varie du simple contact entre les institutions de base à une convergence et interdépendance qui permet aux institutions de transcender leurs missions d'origine et rôles.

2.2 *Approches infocommunicationelles*

Si, dans le passé, la singularité française, les sciences de l'information et de la communication fruit d'une conjoncture particulière et institutionnelle articulant deux disciplines largement isolés à l'étranger, est difficilement soutenable sur le plan scientifique, elle est au contraire devenue, appropriée pour étudier et nous éclairer des mutation sociétales, techniques issus de la numérisation croissante des activités culturelles et marchandes. Le dépassement du cadre étroit de l'étude des documents ou de l'artefact pour s'intéresser à d'autres dimensions telles que l'usage, les pratique informationnelles, les politiques d'information ou les ques-

tion éthiques ou économiques montre au contraire le besoin d'une approche infocommunicationelle (Miège, 2004). Ce dernier rappelle que les évolutions actuelles de l'information spécialisée (marchandisation internationalisation, diffusion par les nouveaux médias) rapprochent les secteurs de la documentation-bibliothèque de celui des industries culturelles. D'autres auteurs ont proposé d'enrichir les recherches en sciences de l'information par les réflexions théoriques issues de la communication (Chaudiron et Ihadjadene, 2010). Les études en sciences de la communication ont décrit et analysé depuis presque trente ans les réactions des destinataires des programmes face aux offres qui leur étaient adressées. Ces travaux (études de réception, sociologie de la culture et des pratiques culturelles, sociologie des usages, analyse des médiations, théorie critique, etc...) constituent un apport majeur pour une meilleure compréhension et modélisations des pratiques informationnelles. Pendant longtemps, les travaux sur l'information ont été pensées sans prendre en compte les fonctions de communication qui l'accompagnent (Fondin, 2006). C'est probablement l'une des raisons pour lesquelles la majorité des modèles informationnels minorent les activités de partage et de communication de l'information et privilégient souvent l'activité d'accès et de recherche.

Pourtant, les mutations techniques des dispositifs de production, de recherche, de partage et de diffusion de l'information induisent des transformations organisationnelles, en favorisant l'émergence d'un nouvel écosystème informationnel, constituent un terrain privilégié pour renouveler les approches sur les pratiques informationnelles. Ces nouveaux dispositifs (Web 2.0, réseaux sociaux, CMS, logiciels de filtrage…) banalisent les modalités de production de l'information et fusionnent les fonctionnalités de recherche et d'édition. En ce sens, les dispositifs d'accès à l'information actuels (notamment les moteurs de recherche) présentent les caractéristiques d'un nouveau média (plutôt d'un web-media). Jean-Michel Salaun (2006) et Alexander Halavais (2009) soulignent que le monde du web emprunte le modèle économique à la radio-télévision car il est fondé sur une économie de l'attention mais qu'il emprunte aussi au modèle de la bibliothèque par le partage et la mise en commun des connaissances. Ces dispositifs sont des médias c'est-à-dire des objets qui, pour reprendre Yves Jeanneret (2009), ne font pas que s'associer à du social, mais qui en produisent via notamment des médiations symboliques et représentationnelles. La jonction avec les approches communication-

nelles peut mettre en évidence des enjeux liés à la médiation des savoirs, à la gestion de l'identité numérique et à la question des pouvoirs. Dépassant leur statut de simples récepteurs, les usagers jouent désormais un rôle actif dans la production, le classement, l'évaluation de l'information. Les individus, mais aussi les organisations, mettent aussi en œuvre des stratégies ou des politiques pour faciliter l'échange d'information et sa communication via des plateformes d'intermédiation (Intranet, réseaux sociaux, etc.) (Georges, 2010).

L'une des voies de recherche qui nous semble stimulante et mobilisatrice est d'étudier les pratiques informationnelles sous l'angle de la médiation numérique. En effet, dans une institution culturelles ou dans une organisation (notamment pour la gestion de leur contenu informationnel), le choix des normes et standards, les politiques d'indexation et de classement, les opérations d'inventaire et de constitution de collection ne sont pas neutres mais sont médiées par des logiques sociales. En effet, selon Bernadette Dufrene (2007), ces opérations ne sont pas «indépendants du cadre dans lesquels ils s'élaborent; leur sens est relatif à des conditions d'énonciation propres à des types d'institutions et à des situations historiques». L'intérêt d'une approche sous l'angle de la médiation est d'amorcer une critique des conditions dans lesquelles les textes, les données, les dispositifs sont produits afin de mieux cerner leur interprétation. (Jeanneret, 2009).

Le concept de la médiation largement utilisé dans les pays francophones depuis les années 90, a fait essentiellement l'objet de deux types d'approches: le passage du paradigme système (orienté collection) vers le paradigme usagers (LE Coadic,1997) a renforcé ainsi, dans le cas des bibliothèques, la fonction accueil et le rôle du bibliothécaire comme accompagnateur des innovations notamment techniques (Chourrot cité par Ihadjadene et Dufrene, 2011) tandis que l'événementialisation de la culture qui touche toutes les institutions patrimoniales suscite une série d'études sur le rôle culturel en matière d'animation et de valorisation. Cette médiation traditionnelle implique une interaction dans des espaces particuliers, bibliothèques, centres documentation où l'utilisateur a un besoin d'information immédiat. Ces interactions sont aussi enrichies par la présence physique des collections. (Demeurisse et al, 2008). Les sources, immédiatement disponibles peuvent aider à clarifier beaucoup d'aspects dès le début de la recherche. La mission de ces derniers a évolué de la simple mise à disposition d'informations vers la construction de

liens et d'interactions entre besoin et usage de l'information (Ihadjadene&Chaudiron, 2001).

L'émergence de pratiques d'«automédiation» ne modifie pas toutes les pratiques relevant de la médiation (Miège, 2008). Avec Internet, beaucoup d'observateurs avaient prédit que l'échange direct entre les personnes ferait disparaître les intermédiaires. Or, à fur et à mesure que le temps passe, c'est le phénomène inverse qui se développe: les intermédiaires, sous une forme nouvelle, sont en plein essor. La nature de la médiation professionnelle est diverse. Elle concerne aussi bien les fonctions de représentation des connaissances, de formation et d'information (portails communautaires). Si les interactions sociales jouent un rôle essentiel dans les comportements liés à la consommation d'informations par les internautes, toute intermédiation efficace repose aujourd'hui sur une gestion et une exploitation fine des métadonnées d'un secteur donné et une mise en oeuvre d'une éditorisalisation des contenus informationnels. Les industries culturelles sont ainsi un exemple saisissant de l'impact de la numérisation sur une chaîne de valeur et de la redéfinition de la fonction d'intermédiation (Gensollen, 2009). Selon ce dernier auteur, c'est à travers la gestion des métadonnées et des plateformes d'intermédiation que s'établit la socialisation des goûts ou la décentralisation des savoirs. C'est sans doute là qu'entre en jeu les nouveaux acteurs de la médiation. Les termes employés pour désigner ces nouveaux acteurs - dont la précision même constitue un enjeu théorique au regard de la multiplicité des discours- sont divers: «community managers, accompagnateurs, infomédaiares, go-between, formateurs».

La question de la médiation numérique peut contribuer à éclairer plusieurs enjeux actuels. D'une part, elle appartient à une catégorie d'activités dans laquelle le travail n'est pas stabilisé (Aubouin, 2010) et on peut se poser la question de l'appartenance de cette activité à ce que certains sociologues de profession appellent «les métiers flous» dont lesquelles cette indétermination peut engendrer de nombreuses interprétations (Champy, 2009). (Denis, 2012). Les acteurs ont de grandes marges de manœuvre pour définir leur métier au quotidien (Boussard, 2010). S'agit-il alors d'une adjonction de nouvelles fonctions à des activités documentaire classiques Faudra t-il amener les compétences de ces professionnels au cœur du savoir-faire des instituons culturelles et, pour cela, repenser les différents métiers? Les travaux académiques sur ces acteurs sont réduites et il est important de représenter la réalité de leurs pratiques.

Enfin, nous pensons que d'autres considérations, négligées auparavant, interviennent désormais dans toute étude sur les usages des dispositifs d'accès à l'information: la question éthique (avec la traçabilité des usagers, l'échange, la conservation et la commercialisation des données personnelles), la question économique (publicité en ligne, marchandisation accrue du lien social, monétisation du contexte via le géo-référencement, etc.), la question pédagogique (apprentissage et culture de l'information), la question identitaire (création de communautés en ligne, redéfinition des identités, etc.), la question politique (la censure dans l'indexation des sites par les moteurs, le respect de la diversité culturelle, etc.), les enjeux juridiques (la mise en ligne d'œuvres protégées, l'accès à des contenus illicites via des moteurs spécialisés...). La mise en évidence de ces enjeux socio-politiques au niveau de l'individu, d'une organisation ou de la société permet de refléter des rapports de pouvoir issus de la généralisation des dispositifs d'accès et de diffusion de l'information.

Conclusion

A partir de l'analyse des enquêtes sur l'insertion des professionnels de l'information, il ressort un besoin de reconstruction de l'identité professionnelle, une transformation des conditions d'exercice, une refondation des formations existantes et une incertitude sur les contours des métiers qui émergent (webmestres, rédacteurs de contenu, veilleurs, kno*wledge managers, record manageurs, g*estionnaires d'e-réputation, *curators, community manager,* etc.) puisque la définition de ces nouveaux métiers est mouvante.. Le besoin de formation continue et d'un accès à la recherche scientifique dans le champ est indispensable.

La numérisation accrue des activités humaines, marchandes ou non, et l'évolution des métiers rendent progressivement inopérante l'opposition entre l'information grand public produite par les organes de presse et l'information scientifique et technique (Polity, 2000) (Miège, 2004). Nous avons à travers ce chapitre établi un panorama des métiers de l'information et des recherches en sciences de l'information en France en se focalisant sur les travaux français des dix dernières années sur deux thématiques importantes: le document numérique et la médiation. Nous avons précisé que les formations et les recherches sur la documentation

et les bibliothèques s'établissent massivement au sein des sciences de l'information et de la communication. L'articulation entre information et la communication devient ainsi notre horizon épistémologique (Miège, 2004). Nous considérons qu'il est primordial pour un chercheur en science de l'information de conjuguer une réflexion sur les artéfacts et leurs inscriptions culturelles, sociales, juridiques et économiques. Rejoignant Jean Paul Metzger (2002) lorsqu'il souligne qu'une recherche en science de l'information, tout en faisant partie des sciences humaines et sociales, doit s'articuler autour de trois pôles: le savoir enregistré et du flux (et donc les questions liées au support documentaire, aux systèmes de traitement de l'information etc..), l'humain et le social (usages et pratiques) et enfin la formalisation et le calcul (donc aux aspects algorithmique, etc.).Le dispositif informationnel et les documents ont certes leurs propres contraintes mais ils s'inscrivent dans un environnement socio-politique qui les précède (Guyot, 2006). C'est au niveau des hybridations et des articulations entre la technique et l'humain, l'individu et le groupe qu'on trouvera une particularité à notre discipline scientifique. Nous faisons l'hypothèse que l'étude du rapport entre l'usager et son environnement socio-économique est une question centrale, de même que celle de l'articulation de ce rapport avec les fonctionnements cognitifs. Les facteurs cognitifs ne seront donc pas seulement étudiés pour eux-mêmes, mais aussi en relation avec les nombreuses variables qui sont liées à la dimension fondamentalement sociale et culturelle des usagers.

Bibliografia

ACCART, J-P., *Bibliothécaire, documentaliste: même métier?*, in «Bulletin des bibliothèques de France», vol. 45, n. 1, 2000, pp. 88-93

AUBOUIN, N., KLETZ, F., LENAY, O., *Médiation culturelle: l'enjeu de la gestion des ressources humaines*, in «Culture études», vol. 1, n. 1, 2010, pp. 1-12

BACHIMONT, B., *Bibliothèques numériques audiovisuelles: des enjeux scientifiques et techniques*, in «Document Numérique», vol. 2, n. 3-4, 1999, pp. 219–242

BAZIN, P., *L'avenir incertain des bibliothèques, 2006*, p. 105 <http://cla.univ-fcomte.fr/gerflint/Perou2/Bazin.pdf>

BOUSSARD, V., DEMAZIERE, D., MILBURN, P., *L'injonction au professionnalisme. Analyses d'une dynamique plurielle*, Presses Universitaires de Rennes, 2000

Boure R. (a cura di), *Les origines des Sciences de l'information et de la communication. Regards croisés*, Villeneuve d'Ascq, Presses Universitaires du Septentrion, 2002

Cacaly, S., Le Coadic, Y-F., Pomart, P-D., Sutter, E., *Dictionnaire de l'Information*, ed.2, Parigi, Armand Colin, 2006

Cardy, H., Froissart, P., *Les enseignants-chercheurs en Sciences de l'information et de la communication. Portrait statistique,* in Les recherches en information et communication et leurs perpectives: Histoire, objet, pouvoir, méthode. Actes du XIII Congrès national des sciences de l'Information et de la communication, Marsiglia, 7-9 ottobre 2002, SFSIC, 2002, pp. 353-362

Cardy, H., Froissart, P., *SIC: cartographie d'une discipline*, in Sciences de l'information et de la communication. Objets, savoirs, discipline, Olivesi S. (a cura di), Grenoble, Presse Universitaire de Grenoble, 2006, pp. 259-278

Champy, F., *La sociologie des professions*, in «Quadrige», Parigi, PUF, 2009

Chaudiron, S., Ihadjadene, M., *De la recherche d'information aux pratiques informationnelles*, in «Étude de Communications», vol. 35, 2010, pp. 13-29

Comité National d'Evaluation (CNE), *Les sciences de l'information et de la communication*, Rapporto di valutazione, 1993

Courbières, C., *Femme en contextes: la conception stéréotypée du féminin au travers du langage documentaire (1958-2008)*, Mémoire pour l'habilitation à diriger des recherches en Sciences de l'Information et de la Communication, 2010, Università di Toulouse le Mirail

Couzinet, V., *Médiations hybrides: le documentaliste et le chercheur de sciences de l'information*, Parigi, ADBS éditions, 2000

Couzinet, V., *Convergences et dynamiques nationales: pour une mise en visibilité des recherches en sciences de l'information*, in Recherches récentes en Sciences de l'information. Convergences et dynamiques : Actes du colloque Mics-Lerass, Università Paul Sabatier, Toulouse, 21-22 mars 2002, Parigi, ADBS Éditions, pp. 9-14

Demeurisse, J.S., Fabre, I., Gardiès, C., *Organisation du savoir et médiation documentaire: exemple du traitement des périodiques d'histoire dans deux bibliothèques universitaires*, in Médiations et Usages des Savoirs et de l'Information: un dialogue France-Brésil, Marteleto R.M., Thiesen I. (a cura di), 2008, pp. 379-392

Defrance, J-P., *La formation archivistique en France: l'exemple du Bureau des métiers et de la formation de la Direction des Archives de France*, in «Archives», vol. 34, n. 1-2, 2002, pp. 81-99

Denis, J., Pontille, D., *Travailleurs de l'écrit, matières de l'information*, in «Revue d'anthropologie des connaissances», vol. 6, n. 1, 2012, pp. 1-20

DEVROEY, J-P., *La place de la bibliothèque dans la formation documentaire à l'université*, 1999 <http://www.fr.ch/bcu/a/pub_elec/etudes_et_recherche_info.pdf>

DIDIER, C., *La Revue à découper, note sur un mode plus rationnel de publier les articles de revue*, in «Bulletin de l'IIB», 1898, pp. 175-182

DUFRENE, B., *Intérêts d'une approche sociohistorique des questions de médiation culturelle*, in Quelles approches de la médiation culturelle?, Parigi, L'harmattan, 2007, pp. 237-244

DUJOL, L., *La médiation numérique: l'exemple des Médiathèques du Pays du Romans, France*, in «Argus», vol. 39, n. 3, 2011, pp. 17-20.

ESCARPIT, R., *L'Information et la Communication. Théorie générale*, 6 ed., Parigi, Hachette, 1991

FAYET-SCRIBE, S., *Histoire de la documentation en France. Culture, science et technologie de l'information: 1895-1937*, Parigi, CNRS Editions, 2000

Favier J. (a cura di), *La pratique archivistique française*, Parigi, Archives nationales, 1993

FONDIN, H., ROUAULT, J., *L'information: l'arlésienne de l'interdiscipline des sciences de l'information et de la communicatio*n, documento dattiloscritto, 1998

FONDIN, H., *La science de l'information ou le poids de l'histoire*, in «Les Enjeux de l'information et de la communication», 2006, pp. 1-19 <http://w3.u-grenoble3.fr/les_enjeux/2005/Fondin/home.html>

GARDIN, J-C., ROUX, V., *The Arkeotek project: a european network of knowledge bases in the archaeology of techniques*, in «Archeologia e Calcolatori», vol. 15, 2004, pp. 25-40

GEORGES, F., *Identités virtuelles. Les profils utilisateur du Web 2.0*, Questions théoriques, 2010

GENSOLLEN, M. (a), *Le web relationnel: vers une économie plus sociale?*, in Le Web relationnel, mutation de la communication?, Millerand F., Proulx S., Québec, Presses de l'Université du Québec, 2009

GENSOLLEN, M. (b), *A quoi rassemblera le monde numérique, en 2030?*, in «Réalités Industrielles», maggio 2009

GUYOT, B., *Dynamiques informationnelles dans les organisations*, Parigi, Lavoisier, 2006

HALAVAIS, A., *Search Engine Society*, Cambridge, Polity Press, 2009

IBEKWÉ-SAN-JUAN, F., *The French Conception of Information Science: Une exception française?*, in «Journal of the American Society for Information Science and Technology», vol. 63, n. 9, 2012, p. 1693-1709

IHADJADENE, M., CHAUDIRON, S., *La recherche et la diffusion d'information sur Internet: vers de nouvelles médiations?*, in Emergences et continuité dans les recherches en information et communication: Actes du XII Con-

grès national des sciences de l'information et de la communication, Unesco, 10-13 gennaio 2001, Parigi, SFSIC, pp. 163-171

IHADJADENE, M., CHAUDIRON, S., *L'Étude des dispositifs d'accès à l'information électronique: approches croisées*, in Problématiques émergentes dans les sciences de l'information, Papy F. (a cura di), Parigi, Hermès-Lavoisier, 2008, pp. 183-207

IHADJADENE, M., DUFRENE, B., *Les médiations en bibliothèque: une logique de service public?*, in «Argus», vol. 39, n. 3, 2011, pp.22-25

JEANNERET, Y., *Les sciences de l'information et de la communication: Une discipline méconnue en charge d'enjeux cruciaux*, in «La lettre d'infor-com», n. 60, 2001, pp. 3-45

JEANNERET, Y., *La relation entre médiation et usage dans les recherches en information-communication en France*, in «RECIIS Electronic Journal of Communication Information & Innovation in Health», vol. 3, n. 3, 2009 <http://www.reciis.cict.fiocruz.br/index.php/reciis/article/view/276/320>

JEANNOT, G., *Les Métiers flous. Travail et action publique*, Toulouse, Editions Octarès, 2005

LEBIGRE, L., *Communautés de l'info-doc: un équilibre subtil*, in «Documentaliste-Sciences de l'Information», vol. 48, n. 2, 2011, pp. 22-35

LEVIE, F., *L'homme qui voulait classer le monde*, Bruxelles, Les impressions nouvelles, 2006

LE COADIC, Y-F., *Usages et usagers de l'information*, Parigi, ADBS: Nathan, 1997

LIMOUSIN, M. C., *L'éphémerographie: Bibliographie des journaux et pubblications périodiques*, in «Bulletin de l'IIB», 1900, pp. 147-149

MAHON, B., *The disparity in professional qualifications and progress in information handling: a European perspective*, in «Journal of information science», vol. 34, n. 4, 2008, pp. 567-575

MELOT, M., *Archivistes, documentalistes, bibliothécaires: Compétences, missions et intérêts communs*, in «Bulletin des Bibliothèques de France», t. 50, n. 5, Parigi, 2005, pp. 9-12

MEYRIAT, J., *Un siècle de documentation: la chose et le mot*, in «Documentaliste-Sciences de l'Information», vol. 30, n. 4-5, luglio-ottobre 1993, pp. 192-198

MICHEL, J., *Les documentalistes: l'urgence d'une reconnaissance sociale*, in «Hermès», n. 35, 2003, pp. 185-193

MICHEL, J., *Crise économique, crise de la profession... constats et perspectives d'évolution*, in «Documentaliste-Sciences de l'Information», vol. 49, n. 3, 2011, pp. 4-7

MIÈGE, B., *L'information–communication, objet de connaissance*, Bruxelles, De Boeck & INA, 2004

MIÈGE, B., *Médias, médiations et médiateurs, continuités et mutations*, in «Réseaux», vol. 2, n. 148-149, 2008, pp. 117-146

METZGER, J-P., *Les trois pôles de la science de l'information*, in Recherches récentes en sciences de l'information: convergences et dynamiques, Couzinet V., Régimbeau G. (a cura di), in «Sciences de l'information - série Recherches et documents», ADBS Éditions, 2002

OLLIVIER, B., *Les sciences de la communication. Théories et acquis*, Parigi, Armand Colin, 2007

OTLET, P., *Les sciences bibliographiques et la documentation*, in «Bulletin de l'Institut International de Bibliographie», n. 8, Bruxelles, 1903, pp. 125-147

OTLET, P., *L'information et la documentation au service de l'industrie*, in Bulletin de la Société d'encouragement pour l'industrie nationale, Parigi, Renouard, 1917

OTLET, P., *Monde. Essai d'universalisme*, Bruxelles, Editiones Mundaneum, 1935

PALERMITI, R., POLITY, Y., *Dynamiques de l'institutionnalisation sociale et cognitive des sciences de l'information en France*, in Les origines des sciences de l'information et de la communication en France, regards croisés, Boure, R. (a cura di), Lille, Presses universitaires du Septentrion, 2002, pp. 95-123

Paul, V., Perriault, J. (a cura di), *Critique de la raison numériqu*e, in «Hermès», vol. 39, Parigi, CNRS Editions, 2004

POLITY, Y., *Les bibliothèques, objets de recherche universitaire*, in «Bulletin des bibliothèques de France», vol. 46, n. 4, 2001, pp. 64-70

POLITY, Y., *Information I versus Information II*, 2000 <http://www.iut2.upmf-grenoble.fr/RI3/Information.htm>

RAYWARD, W.B., *Knowledge organization and a new world polity: the rise and fall of the ideas of Paul Otlet*, in «Transnational Associations», n. 1-2, 2003, pp. 4-15

RAYWARD, W.B., *Visions of Xanadu: Paul Otlet (1868-1944) and Hypertext*, in «Journal of the American Society for Information Science», vol. 4, n. 4, 1994, pp. 235-250

RENOULT, D., *Les bibliothèques numériques*, in Des Alexandries 1. Du livre au texte, Giard L., Jacob, C. (a cura di), Parigi, Bibliothèque National de France, 2001, pp. 83-90

ROGER T. PÉDAUQUE, *Le Document à la lumière du numérique: forme, texte, médium: comprendre le rôle du document numérique dans l'émergence d'une nouvelle modernité*, C&F éditions, 2006

SALAÜN, J-M., *Les trois facettes du document numérique et le nouvel ordre documentaire*, Intervento al Seminario «Ressources numériques au CDI», 10-11 maggio 2012

<http://cdi.ac-amiens.fr/sites/cdi.ac-amiens.fr/IMG/pdf/Les_trois_facettes_du_document_numerique_et_le_nouvel_ordre_documentaire.pdf>
SALAÜN, J-M., *Web-média, synthèse*, 2006 <http://grds04.ebsi.umontreal.ca/jms/index.php/2006/11/09/116-web-media-synthese>
SALAÜN, J-M., *La redocumentarisation, un défi pour les sciences de l'information*, in «Études de Communication» n. 30, 2007, pp. 13-23
STILLER, H., *La fonction Information-Documentation dans les grandes entreprises: une enquête*, in «Documentaliste-Sciences de l'Information», vol. 38, n. 3-4, 2001, p. 222-225
STROOBANTS, J-P., *Le Web, une histoire belge*, in «Le Monde», 03/11/2012
WIEGANDT, C., *Bibliothécaires et documentalistes: deux métiers qui se rapprochent*, in «Bulletin des Bibliothèques de France», t. 50, n. 5, Parigi, 2005, p. 16-18
YARROW, A., CLUBB, B., DRAPER, J-L., *Bibliothèques publiques, archives et musées: tendances en matière de collaboration et de coopération*, Rapporto IFLA - Comité permanent de la section des bibliothèques publiques, n. 109, 2008
ZORICH, D., WAIBEL, G., ERWAY, R., *Beyond the Silos of the LAMs: Collaboration among Libraries, Archives and Museums*, Dublino, OCLC Research, 2008

Sitografia

<http://www.etudinfo.com/diplome/licence-professionnelle/>
<http://www.campusfrance.org/fr/page/les-ecoles-d%E2%80%99ingenieur>
<http://www.lesmetiers.net/orientation/p1_325333/bien-choisir-son-ecole-de-commerce>
<http://ressources.campusfrance.org/catalogues_recherche/diplomes/fr/iut_fr.pdf>
<http://www.label-iup.org/>
<http://www.iufm.fr/devenir-ens/choisir/documentaliste.html>
<http://www.adbs.fr/>
<http://www.abf.asso.fr/>
<http://www.fadben.asso.fr>
<http://abdu.fr>
<http://www.archivistes.org/>
<http://www.enc.sorbonne.fr/>

<http://www.serda.com/>
<http://www.laterza.it/bibliotecheinrete/Cap09/Cap09_16.htm>
<http://www.cpcnu.fr/>
<http://www.cpcnu.fr/web/section-71>
<http://www.e-tud.com/encyclopedie-education/?136-doctorat>
<http://www.agendadigitale.regione.lombardia.it/cs/Satellite?c=Page&childpagename=DG_01%2FMILayout&cid=1213474655678&p=1213474655678&pagename=DG_01Wrapper>
<http://hypermedia.univ-paris8.fr/weissberg/presence/3.html>
<http://www.ramau.archi.fr/spip.php?article197>
<http://www.metiers.internet.gouv.fr/metier/veilleur-strategique>
<http://abfblog.wordpress.com/2009/06/page/3/>

Il documento digitale: profili giuridici

ENRICO DE GIOVANNI*

1. Premessa

Si premette che il nostro ordinamento nel riferirsi al documento digitale usa l'espressione *documento informatico*; nel presente scritto si utilizzerà pertanto tale ultima dizione, da considerarsi sinonimo di documento digitale.

Le prime attestazioni della parola *documento* nella lingua italiana risalgono al XIV secolo; è chiara la derivazione dalla parola latina *documentum*, a sua volta da *docere*, cioè insegnare, dimostrare.

Dunque documento è ciò che insegna, dimostra, in sintesi ciò che rappresenta un evento, una manifestazione di volontà o di scienza o fornisce comunque informazioni; pertanto *«attraverso il documento si ha una conoscenza indiretta di un fatto presente o passato in esso rappresentato»* (Masucci, 2011)[1].

L'ordinamento giuridico italiano offre una definizione di documento amministrativo nell'art. 1, lett. a) del D.P.R. 28-12-2000 n. 445[2]: *«Documento Amministrativo: ogni rappresenta-*

* Avvocatura dello Stato.

1 ALFONSO MASUCCI, *La documentazione amministrativa*, in La Documentazione amministrativa. Certezze, semplificazione e informatizzazione nel d.P.R. n. 28 dicembre 2000, n. 445, Rimini, Maggioli, 2011, p. 174.

2 Decreto del Presidente della Repubblica 28 dicembre 2000, n. 445, *Testo unico delle disposizioni legislative e regolamentari in materia di do-*

zione, comunque formata, del contenuto di atti, anche interni, delle pubbliche amministrazioni o, comunque, utilizzati ai fini dell'attività amministrativa».

A prescindere da un'analisi della locuzione *rappresentazione… del contenuto degli atti*, che appare ridondante e poco corretta, si sottolinea che in tal modo una norma positiva (anche se di rango regolamentare, qual è l'art. 1 del citato D.P.R., redatto secondo i canoni all'epoca vigenti relativi ai c.d. *testi unici misti*, che raccoglievano norme primarie, contrassegnate nel testo dalla presenza della lettera L, e regolamentari, recanti la lettera R) sancisce che per documento si intende ogni strumento (*rappresentazione comunque formata*) idoneo a rappresentare; è una definizione che assume valenza generale, riferibile ad ogni tipo di documento, anche se in particolare, ma su questo tema si tornerà più avanti, la disposizione fa riferimento all'atto amministrativo; si vedrà come essa trovi sostanziale riscontro nella definizione di *documento analogico* presente in altro testo di legge.

La lett. b) della stesso art. 1 in esame fornisce poi la definizione giuridica di documento informatico; premesso che il documento informatico, in senso generale ed extragiuridico, si sostanzia in un file con un contenuto rappresentativo, va precisato che la definizione del documento informatico fornita dall'ordinamento italiano è quella, ovviamente, rilevante in ambito giuridico: peraltro va subito segnalato che la medesima definizione, in identica formulazione, si rinviene anche nell'art. 1, comma 1, lett. P) del decreto legislativo 7 marzo 2005, n. 82, recante il Codice dell'amministrazione digitale, e s.m.e i. (d'ora in poi: CAD)[3].

---

cumentazione amministrativa, in Gazzetta Ufficiale del 20 febbraio 2001, n. 42.

[3] Decreto Legislativo 7 marzo 2005, n. 82, Codice dell'amministrazione digitale, in Gazzetta Ufficiale del 16 maggio 2005, n. 112, Supplemento Ordinario n. 93.

La definizione fornita da entrambe le fonti è la seguente: il documento informatico è *«la rappresentazione informatica di atti, fatti o dati giuridicamente rilevanti»*.

Va inoltre sottolineato che il medesimo art. 1 del CAD fornisce anche una definizione di documento analogico: *«p-bis) documento analogico: la rappresentazione non informatica di atti, fatti o dati giuridicamente rilevanti»*, definizione sostanzialmente analoga a quella, sopra citata, di documento amministrativo.

Dunque l'ordinamento giuridico italiano riconosce ormai esplicitamente due tipologie di documento, inteso come rappresentazione: analogico e informatico.

Al fine di comprendere meglio il fondamento e i limiti della valutazione giuridica del documento si aggiungono due ulteriori ma essenziali considerazioni circa la necessità che esso sia idoneo a conservare nel tempo la sua capacità rappresentativa, senza modifiche contenutistiche; è evidente che il valore probatorio del documento, cioè la sua capacità di preservare nel tempo il suo contenuto originario, costituisce un valore essenziale ai fini dell'ordinamento, apparendone del tutto palese il rilievo a fini probatori e, in ultima analisi, di tutela dei rapporti giuridici e delle posizioni giuridiche soggettive; inoltre, in relazione al contenuto del documento, preminente rilevanza può assumere il profilo della sottoscrizione del documento medesimo, profilo che, con specifico riferimento al documento informatico, può ridondare sulla immodificabilità del contenuto.

Di qui l'esigenza, ampiamente raccolta dal legislatore, di predisporre norme a tutela della autenticità e immodificabilità del documento informatico e della sua corretta conservazione, nonché sulla sottoscrizione del medesimo.

Ciò premesso, l'analisi che segue verterà sulla definizione di documento informatico e si volgerà, poi, ai tratti essenziali della relativa disciplina giuridica.

2. Il codice dell'amministrazione digitale

Prima di procedere all'analisi descritta è bene indicare subito quale sia, attualmente, la fonte normativa che disciplina il documento informatico; si tratta essenzialmente del citato CAD, che fu emanato nell'esercizio della delega volta al *«riassetto in materia di società dell'informazione»* contenuta nell'art. 10 della legge 29 luglio 2003, n. 229[4].

In attuazione della delega furono promulgati 3 provvedimenti: nell'ordine il decreto legislativo 28 febbraio 2005, n. 42, recante l'istituzione del Sistema Pubblico di Connettività e della Rete internazionale delle pubbliche amministrazioni[5]; il decreto legislativo 7 marzo 2005, n. 82, recante il Codice dell'amministrazione digitale; il decreto legislativo 4 aprile 2006, n. 159, recante integrazioni e correzioni al decreto legislativo n. 82 del 2005[6], con cui, fra l'altro fu integrato nel testo originario del decreto legislativo n. 82 del 2005 anche il citato D.Lgs. 42 del 2005.

Al CAD furono successivamente apportate ulteriori puntuali modifiche e integrazioni e da ultimo una più ampia revisione con il D.lgs. 30 dicembre 2010, n. 235[7], che lo ha novellato ampia-

---

4 Legge 29 luglio 2003, n. 229, *Interventi in materia di qualità della regolazione, riassetto normativo e codificazione* - Legge di semplificazione 2001, in Gazzetta Ufficiale del 25 agosto 2003, n. 196.

5 Decreto Legislativo 28 febbraio 2005, n. 42, *Istituzione del sistema pubblico di connettività e della rete internazionale della pubblica amministrazione, a norma dell'articolo 10, della legge 29 luglio 2003, n. 229*, in Gazzetta Ufficiale del 30 marzo 2005, n. 73.

6 Decreto Legislativo 4 aprile 2006, n. 159, *Disposizioni integrative e correttive al decreto legislativo 7 marzo 2005, n. 82, recante codice dell'amministrazione digitale*, in Gazzetta Ufficiale del 29 aprile 2006, n. 99, Supplemento Ordinario n. 105.

7 Decreto Legislativo 30 dicembre 2010, n. 235, *Modifiche ed integrazioni al decreto legislativo 7 marzo 2005, n. 82, recante Codice dell'ammini-*

mente anche in materia di documento informatico; da ultimo sono state introdotte alcune modificazioni, rilevanti ai fini del tema del documento digitale, dal D.L. 18 ottobre 2012, n. 179[8], come convertito dalla legge 17 dicembre 2012, n. 221[9].

Dunque è nell'art. 1 del CAD e nei successivi artt. da 20 a 23 quater che vanno attualmente rinvenute le più rilevanti norme applicabili al documento informatico.

3. Il documento informatico: analisi della definizione

L'approccio al tema del documento informatico prenderà le mosse da una sintetica ricognizione dei concetti di forma, atto o fatto (e dato) e contenuto e quindi di documento come da intendersi ai fini del presente testo, senza alcuna pretesa di completezza o approfondimento scientifico; si tratta, tuttavia, di precisazioni essenziali, anche per poter comprendere quali concetti e principi sono stati posti a base dell'elaborazione legislativa concretizzatasi nel CAD, il che può risultare utile anche ai fini di una ricostruzione dell'intendimento del legislatore e, nei limiti del rispetto delle regole legali d'interpretazione, del significato stesso della legge.

Il termine *forma* fu acquisito nella scienza giuridica dal linguaggio filosofico; esso è certamente fra i termini che possono

strazione digitale, a norma dell'articolo 33 della legge 18 giugno 2009, n. 69, in Gazzetta Ufficiale del 10 gennaio 2011, n. 6, Supplemento Ordinario n. 8.

8 Decreto Legge 18 ottobre 2012, n. 179, *Ulteriori misure urgenti per la crescita del Paese*, in Gazzetta Ufficiale del 19 ottobre 2012, n. 245, Supplemento Ordinario n. 194.

9 Legge 17 dicembre 2012, n. 221, *Conversione in legge, con modificazioni, del decreto-legge 18 ottobre 2012, n. 179, recante ulteriori misure urgenti per la crescita del Paese*, in Gazzetta Ufficiale del 18 dicembre 2012, n. 294, Supplemento Ordinario n. 208.

vantare un uso più antico tra i giuristi, ma anche fra i termini che, nel tempo, hanno assunto significazioni diverse.

I maggiori approfondimenti di questo concetto sono stati svolti con riferimento al negozio giuridico, cosicché è al concetto di *forma del negozio giuridico* che sono state dedicate le maggiori elaborazioni della dottrina. La definizione più immediata, corretta ed utile ai fini della presente esposizione è quella che intende la forma come *modo di essere della manifestazione di volontà*; più in generale, ampliando il discorso anche di là del negozio strettamente inteso, si può ragionevolmente adottare l'idea che per forma s'intende il modo attraverso cui un atto o fatto viene rappresentato, divenendo manifesto e percepibile.

Le espressioni *atto o fatto*, riferibili al contenuto del documento e presenti nella definizione legislativa di documento informatico sopra ricordata, vanno intese, ovviamente, non in senso naturalistico ma in senso giuridico; dunque s'intenderà un atto (umano) o un fatto (un accadimento naturale), in sintesi un evento, cui l'ordinamento attribuisce rilevanza giuridica.

Anche la parola *dati*, contenuta sempre nell'art. 1 lett. P del CAD, merita un chiarimento; uno dei più recenti ed accreditati vocabolari della lingua italiana (Devoto, Oli, 2011)[10] definisce il dato come *«ciascuno degli elementi di cui si dispone per formulare un giudizio o risolvere un problema»*, ma precisa che per dato informatico si intende *«la singola informazione codificabile o codificata»*.

Dunque per *dati*, ai sensi della definizione in esame, si dovranno intendere le estrapolazioni o comunque le rielaborazioni di informazioni relative a atti o fatti realizzate da programmi informatici, o comunque contenute in un file, aventi rilevanza giuridica.

---

[10] GIACOMO DEVOTO, GIANCARLO OLI, *Vocabolario della lingua italiana*, Le Monnier, 2011.

Riassumendo il *documento* è uno strumento attraverso cui l'atto o fatto giuridico, o dati a partire da questi rielaborati, vengono rappresentati in modo duraturo nel tempo in modo da poter essere poi conosciuti da chi abbia percezione del documento medesimo; nelle modalità di rappresentazione si rinviene la *forma*.

Sul piano giuridico, parlare di un *contenuto* distinto dalla *forma* ha utilità solo laddove quest'ultima risulti espressamente disciplinata dalla vigente normativa: in tal caso particolare rilevanza assume nel nostro ordinamento la forma scritta, richiesta, a seconda dei casi, *ad substantiam* o *ad probationem*. Il contenuto non è altro che l'atto (e fra questi anche, ad esempio, la dichiarazione di volontà negoziale) o il fatto manifestatosi in una determinata modalità, modalità che (ferma restando, ovviamente, la percepibilità), per quanto concerne le proprie specificità, è rilevante sul piano giuridico, si ribadisce, solo ove l'ordinamento ne richieda una particolare caratteristica o natura ai fini espressamente indicati.

A questo punto appare utile una precisazione ulteriore: nel parlare del *documento* si è fatto riferimento ad uno *strumento* che rappresenta, e non, come normalmente si è fatto a lungo in dottrina, ad una *res*, ad un'entità materiale su cui la rappresentazione trova stabile concretizzazione; nel caso del documento informatico va subito chiarito che la rappresentazione medesima si fissa in bit, cioè in impulsi elettronici, che difficilmente possono essere definiti oggetti o *res*; la circostanza che questi impulsi trovino base su supporti informatici le cui caratteristiche possono variare (si pensi agli hard disk, ai DVD e comunque a tutti i molteplici ulteriori strumenti, in continua evoluzione, creati dagli sviluppi tecnologici) non toglie che la rappresentazione sia in realtà contenuta nei *bit* e che dunque il supporto occasionalmente utilizzato e l'ulteriore mezzo tecnologico usato per la lettura del supporto (computer, monitor ecc.) risultino semplici mezzi idonei ad ospitare nel tempo e rendere materialmente percepibile la rappresentazione contenuta negli impulsi elettronici. Perciò

nel fare riferimento al *documento* appare necessario superare il concetto di *res*, e quindi di materia tangibile e giungere al concetto di *strumento*, inteso come mezzo materiale o immateriale idoneo comunque a contenere una rappresentazione di atti o fatti ed a consentirne, anche in uso combinato con altri mezzi, la percezione.

Si condivide, quindi, la tesi esposta in dottrina (Zagami, 2000)[11] secondo la quale l'avvento degli strumenti informatici ha travolto il preesistente concetto dogmatico-giuridico di documento, che può ormai consistere anche in un quid immateriale, quali sono gli impulsi elettronici, che ben possono essere trasferiti dal supporto su cui originariamente furono formati su altro supporto, venendo meno su quello originario, ma che non per questo perdono la propria identità; proprio questa mutabilità del supporto materiale impedisce l'identificazione tra il documento informatico ed il supporto occasionalmente utilizzato.

Dunque, le oggettive caratteristiche del documento informatico devono indurre a modificare la definizione di documento, che ha perso ormai la connotazione di *res*, di cosa materiale, e che può consistere in meri bit, impulsi elettronici.

Questa circostanza non costituisce solo una significativa novità dal punto di vista dogmatico, ma ha costretto anche ad una serie di considerazioni e scelte legislative delicate, giacché il nostro ordinamento risente certamente della circostanza che finora nel riferirsi ad un documento, ad una forma in grado di rappresentare stabilmente un atto o un fatto, si è sempre essenzialmente pensato ad un documento cartaceo.

Non a caso i termini usati dai Romani per indicare il documento erano *Chirographum, epistola, scriptura, tabula*.

È stata tale l'identificazione tra carta scritta e documento che

[11] RAIMONDO ZAGAMI, *Firma digitale e sicurezza giuridica*, Padova, 2000, p. 147.

nell'accezione corrente divengono quasi sinonimi. (Carnelutti, 1957) notava che:

> *il più antico ed ancora il più diffuso tra i mezzi documentali è la scrittura e, oggi, la materia su cui si scrive è la carta… così… scrittura e carta si adoperano, per antonomasia, con significato di documento… ma il vero è che qualunque materia atta a formare una cosa rappresentativa può entrare nel documento: tela, cera, metallo, pietra e via dicendo*[12].

Sostanzialmente nello stesso senso sembra esprimersi (Luzzatto, 1960), il quale, pur osservando che per documento *«in senso più ampio e generico può essere considerato ogni mezzo di prova diretto ad accertare l'esistenza di un fatto»*, tuttavia precisava che *«in senso più specifico il termine si riferisce alla prova scritta»*[13].

Questa evidente realtà non è rimasta senza conseguenze sul piano normativo prima dell'avvento dell'era digitale; il nostro ordinamento, ed in particolare il codice civile del 1942, nel fare riferimento al documento ed alla forma scritta fanno implicito ma palese riferimento alla scrittura su carta, nel senso che può cogliersi facilmente una sorta di riserva mentale del legislatore in tal senso. Si rilegga,ad esempio, l'art. 2702 del codice civile: *«La scrittura privata fa piena prova, fino a querela di falso, della provenienza delle dichiarazioni da chi l'ha sottoscritta […]»*. Non è precisato che per scrittura si intenda un foglio di carta recante segni grafici, ma è evidente che questa era l'idea che sosteneva il legislatore nel redigere la norma.

Dunque l'intero ordinamento risente di questa impostazione,

---

[12] FRANCESCO CARNELUTTI, *Documento. Teoria moderna*, in Novissimo Digesto italiano, vol. 6, Torino, 1957, p. 86.

[13] GIUSEPPE IGNAZIO LUZZATTO, *Documento-Diritto romano*, in Novissimo Digesto Italiano, vol. 6, Torino, 1960, p. 84.

cosicché anche le norme relative al valore probatorio, alla sottoscrizione ed al valore giuridico delle copie dei documenti appaiono costruite su tale presupposto di fatto.

Come segnalato, l'evoluzione tecnologica ha fatto sì che la materia su cui si scrive non sia più solo la carta ma sia anche (e sempre più lo sarà in futuro) il computer, con il suo corredo di tastiere, schermi, dischi fissi e mobili al servizio degli impulsi elettronici in cui si esprime il contenuto. E che, quindi, il documento perdesse quel connotato di materialità e tangibilità che certamente gli impulsi elettronici non presentano.

Quindi, non solo tutto l'armamentario teorico e culturale che ha sostenuto finora la legislazione e la dottrina in materia di documento e forma andrebbero rivisti e riletti profondamente alla luce di questa nuova realtà, ma il legislatore, nel disciplinare positivamente il documento informatico, si è dovuto porre innanzi tutto il quesito se regolarne la rilevanza giuridica sulla base di una sistematica del tutto nuova, che prescindesse da un richiamo alle vigenti disposizioni concernenti il documento inteso come carta scritta, ovvero richiamarsi ai principi già correnti.

Al riguardo va considerato che alcuni concetti, come quello di forma scritta o di sottoscrizione, sono alla base di numerosi istituti sia civilistici che pubblicistici, nonché di varie norme processuali; dunque l'introduzione di una disciplina completamente nuova ed originale avrebbe creato non pochi problemi di applicazione, con la necessità di ricorrere ad un adattamento dell'ordinamento con l'emanazione di numerose norme integrative, o avrebbe costretto gli operatori del diritto a spericolate opzioni interpretative, a danno della stessa certezza della legge.

A ciò si aggiunga che la formulazione di una nutrita serie di nuove norme applicabili esclusivamente all'informatica avrebbe ulteriormente appesantito un ordinamento già pletorico, introducendo difficoltà di conoscenza sia per gli operatori del diritto sia, soprattutto, per i privati cittadini, in palese contrasto con l'obiettivo di una semplificazione normativa perseguita (almeno a pa-

role) da anni dal legislatore e riaffermata nella stessa legge delega che ha originato il CAD.

Una scelta così radicale avrebbe, in altri termini, rischiato un impatto traumatico sull'ordinamento, determinando disarmonie, discontinuità e complessità in una materia importante, nuova e delicata: l'altra opzione, peraltro già seguita dal legislatore in occasione dell'emanazione delle precedenti normative nella stessa materia, era invece quella di cercare di ricondurre la realtà informatica nell'ambito degli istituti civilistici già esistenti, ovviamente con gli aggiustamenti del caso.

Anche questa opzione ha, naturalmente, le proprie controindicazioni: non sempre gli istituti giuridici preesistenti hanno la capacità di *ospitare* in modo congruo le nuove realtà tecnologiche, pur se adattati; ma è sembrato al legislatore delegato che questo fosse un prezzo minore da pagare rispetto al rischio di una discontinuità traumatica.

D'altronde si è considerato che l'ordinamento è un corpo vivo in continuo mutare, e che questo processo diretto ad adeguare le norme alle nuove realtà fattuali, se non addirittura, talvolta, a favorire l'utilizzo delle nuove tecnologie, condurrà in modo graduale all'affermazione dei principi legislativi che risulteranno più opportuni per consentire la disciplina dell'uso dell'*Information and Communication Technology* (ICT), secondo i tempi che le esigenze di fatto e l'evoluzione della cultura giuridica detteranno alla sensibilità del legislatore.

Pertanto il legislatore delegato ha mediato, nell'emanazione del CAD, fra l'esigenza di disciplinare la rilevanza giuridica dell'uso dell'informatica, tenendone presenti le peculiarità e le novità, e quella di salvaguardare, per quanto possibile, gli istituti giuridici previgenti richiamando, nei limiti del ragionevole, norme e concetti già noti e tentando di ricondurre negli schemi preesistenti le nuove disposizioni.

Con tale premessa appare più agevole comprendere, ad esempio, perché il CAD ha fatto esplicito riferimento, trattando del

valore probatorio del documento informatico sottoscritto con firma digitale, qualificata o avanzata all'art. 2702 del codice civile e perché affermi che il documento con le medesime caratteristiche soddisfi *il requisito legale della forma scritta*: questa attribuzione di rilevanza giuridica *per relationem*, basato sostanzialmente su un'equiparazione con il documento cartaceo, consente poi di applicare al documento informatico equiparato, senza particolari dubbi interpretativi, la disciplina vigente per il documento cartaceo, senza la necessità di ricorrere a complesse ed insidiose riscritture di norme esclusivamente applicabili all'informatica e ai suoi prodotti.

4. L'evoluzione normativa

Ciò detto sul piano dogmatico appare utile ripercorrere in estrema sintesi l'evoluzione della disciplina positiva del documento informatico nel nostro ordinamento; si dirà subito che le norme sul documento informatico sono state introdotte nell'ordinamento italiano nell'ambito di discipline pubblicistiche ma assurgono a valore di diritto comune, risultando applicabili, per le ragioni che si diranno, anche ai documenti di natura privatistica.

La legge n. 241/1990[14] sul procedimento amministrativo per la prima volta ricomprese nel concetto di *documento amministrativo* anche *«ogni rappresentazione […] elettromagnetica […] del contenuto di atti»*.

Ma fondamentale riconoscimento della piena rilevanza giuridica del documento informatico si ebbe con l'art. 15, comma 2 della legge 15 marzo 1997, n. 59[15], recante *Delega al Governo*

[14] Legge 7 agosto 1990, n. 241, *Nuove norme in materia di procedimento amministrativo e di diritto di accesso ai documenti amministrativi*, in Gazzetta Ufficiale del 18 agosto 1990, n. 192.

[15] Legge 15 marzo 1997, n. 59, *Delega al Governo per il conferimento di*

per il conferimento di funzioni e compiti alle regioni ed enti locali, per la riforma della Pubblica Amministrazione e per la semplificazione amministrativa pubblicata sulla G.U. n. 63 del 17 marzo 1997, una delle più importanti fra le cosiddette *leggi Bassanini*. La disposizione, tuttora in vigore poiché mai abrogata e pienamente compatibile con la successiva produzione legislativa, così recita: *«gli atti, dati e documenti formati dalla pubblica amministrazione e dai privati con strumenti informatici […] sono validi e rilevanti a tutti gli effetti di legge»*.

Poiché nel nostro ordinamento si ritiene vigente il principio generale di libertà della forma[16], il che implica che le norme che impongono una forma particolare debbano essere considerate come eccezioni al predetto principio, probabilmente la solenne affermazione della validità e rilevanza a tutti gli effetti di legge del documento informatico fu, a stretto rigore, inutile sul piano meramente normativo; tuttavia essa fu di capitale importanza sul piano concreto e, per così dire, psicologico, poiché spazzò via ogni dubbio (pur se infondato) sull'idoneità delle nuove tecnologie a dare vita e sostanza ad un'attività giuridicamente rilevante e dunque costituì una pietra miliare nello sviluppo dell'utilizzo delle ICT nei rapporti giuridici pubblicistici e privatistici nonché momento fondamentale nel formarsi delle relative regole giuridiche.

Va sottolineata inoltre, riallacciandosi a quanto sopra segnalato, un'interessante caratteristica della norma che tuttora caratterizza la nostra produzione normativa: l'art. 15, comma 2 si occupa, insieme, dei *documenti formati dalla pubblica amministrazione e dai privati*; dunque una legge nel suo complesso

funzioni e compiti alle regioni ed enti locali, per la riforma della pubblica amministrazione e per la semplificazione amministrativa, in Gazzetta Ufficiale del 17 marzo 1997, n. 63, Supplemento Ordinario n. 56.

16 Cfr. per tutti ALBERTO TRABUCCHI, GIORGIO CIAN, *Commentario breve al Codice Civile*, Padova, CEDAM, 1984, (commento all'articolo 1350).

esplicitamente destinata ad incidere sulla Pubblica Amministrazione (P.A.) e sull'azione amministrativa nel momento in cui si occupa di informatica detta, in realtà, esplicitamente regole di diritto comune, valide per l'intero ordinamento ed applicabili anche ai rapporti tra privati.

Dunque si manifesta l'orientamento del legislatore italiano, tuttora solidamente seguito, ad affidare a normative essenzialmente destinate a regolare l'organizzazione e l'azione della Pubblica Amministrazione anche la disciplina civilistica dell'utilizzo delle nuove tecnologie dell'informazione e della comunicazione.

Siffatto orientamento, pur se discutibile sul piano meramente astratto, trova la sua ragion d'essere in considerazioni pragmatiche ed ordinamentali: in effetti non vi è ragione per prevedere un diverso regime giuridico a seconda del fatto che l'uso dell'informatica sia realizzato da un soggetto pubblico o da un soggetto privato e dunque è ovvio che le medesime regole siano generalmente applicabili. Dunque non avrebbe ragion d'essere una frantumazione della disciplina in diverse fonti, con oggetto di natura privatistica o pubblicistica, che determinerebbe una dannosa frammentarietà ed una complessa conoscibilità dell'ordinamento.

Per tali ragioni anche nella più recente produzione normativa si è seguito l'orientamento descritto, affidando ad un testo destinato in larga misura alla pubblica amministrazione (appunto il Codice dell'amministrazione digitale) la disciplina di aspetti squisitamente privatistici del documento informatico, della sua conservazione e trasmissione e delle firme elettroniche.

Altro aspetto caratterizzante dell'art. 15, comma 2 in esame è il rinvio della disciplina degli aspetti applicativi ad una norma sott'ordinata, nel caso di specie di rango regolamentare, scelta che, anch'essa, caratterizza tuttora la produzione normativa italiana nel settore. Si tratta di una scelta corretta ed opportuna, poiché la norma primaria deve contenere i principi generali, quelli

che improntano ed orientano l'ordinamento di settore, lasciando la regolazione più puntuale a strumenti che più agilmente possono essere adeguati alla rapida evoluzione tecnologica e che comunque sono, per loro stessa natura, proprio le fonti da utilizzare per le norme applicative.

In attuazione della norma in parola fu emanato il fondamentale decreto del Presidente del Consiglio dei Ministri 10 novembre 1997, n. 513[17], che introdusse per la prima volta in Italia una disciplina organica del documento informatico, rinviando, secondo il disposto dell'art. 3, ad un decreto del Presidente del Consiglio dei Ministri per la formulazione delle regole tecniche, che fu emanato con la data dell'8 febbraio 1999[18]. Il decreto disciplinava, nell'allegato tecnico, innanzi tutto, nel Titolo I, le regole tecniche per la formazione, la trasmissione, la conservazione, la duplicazione, la riproduzione e la validazione, anche temporale, dei documenti informatici. In tale parte del provvedimento era contenuta, in particolare, la disciplina delle firme digitali, soluzione tecnica a cui il legislatore italiano attribuiva rilevanza giuridica nettamente prevalente rispetto ad altri tipi di firma, profilo di essenziale importanza anche al fine di garantire l'autenticità e l'immodificabilità del documento.

Nel Titolo II si disciplinavano le *regole tecniche per la certificazione delle chiavi*, con puntuali disposizioni sui certificatori,

---

17 Decreto del Presidente della Repubblica 10 novembre 1997, n. 513, *Regolamento contenente i criteri e le modalità per la formazione, l'archiviazione e la trasmissione di documenti con strumenti informatici e telematici a norma dell'articolo 15, comma 2, della legge 15 marzo 1997, n. 59*, in Gazzetta Ufficiale del 13 marzo 1998, n. 60.

18 Decreto del Presidente del Consiglio dei Ministri 8 febbraio 1999, *Regole tecniche per la formazione, la trasmissione, la conservazione, la duplicazione, la riproduzione e la validazione, anche temporale, dei documenti informatici ai sensi dell'articolo 3, comma 1, del Decreto del Presidente della Repubblica, 10 novembre 1997, n. 513*, in Gazzetta Ufficiale del 15 aprile 1999, n. 87.

cioè sui soggetti che rilasciano gli strumenti di sottoscrizione; il Titolo III recava le *regole per la validazione temporale e per la protezione dei documenti informatici*; il Titolo IV dettava le *regole tecniche per le pubbliche amministrazioni*.

La preminente rilevanza che nella predetta regolazione tecnica è assunta dal tema della sottoscrizione elettronica dimostra quanto essa sia, in fatto e in diritto, uno dei profili essenziali in materia di documento informatico.

Le disposizioni sul documento informatico e sulle firme elettroniche confluirono poi nel *Testo Unico sulla documentazione amministrativa*, Testo Unico di cui al decreto del Presidente della Repubblica n. 445 del 2000[19]; trattandosi di Testo Unico c.d. *misto* fu possibile mantenere alle singole disposizioni *traghettate* nel corpus normativo unitario il rango, primario o secondario, che rivestivano nelle fonti di provenienza. Il totale recepimento del D.P.R. 513/1997 determinò, fra l'altro, l'alterazione della struttura del regolamento, con modifica della successione originale degli articoli, il che tuttavia, non ne modificò la portata precettiva.

Nel frattempo, tuttavia, era stata emanata la direttiva europea 1999/93/CE[20] relativa ad un *quadro comunitario per le firme elettroniche*; direttiva che impattava direttamente e fortemente sulla disciplina del documento e delle firme elettronici presente nel D.P.R. 445/2000, imponendo al legislatore italiano di attribuire una seppur graduata rilevanza giuridica anche a firme elettroniche non digitali; la direttiva è stata recepita per la parte di li-

---

19 Decreto del Presidente della Repubblica 28 dicembre 2000, n. 445, *Disposizioni legislative in materia di documentazione amministrativa. (Testo A)*, in Gazzetta Ufficiale del 20 febbraio 2001, n. 42, Supplemento Ordinario n. 30.

20 Direttiva 1999/93/CE del Parlamento europeo e del Consiglio del 13 dicembre 1999 Relativa ad un quadro comunitario per le firme elettroniche, in Gazzetta Ufficiale delle Comunità europee del 13 dicembre 1999.

vello legislativo con D.lgs. 23 gennaio 2002, n. 10[21], e per le norme regolamentari con successivo decreto del Presidente della Repubblica: entrambi i provvedimenti hanno, in sostanza, operato con la tecnica della novellazione del T.U. 445/2000, sostituendo le norme (primarie e secondarie) superate dalla direttiva europea.

Su quest'ultima fonte è intervenuto, sostanzialmente, il decreto legislativo n. 82 del 2005, cioè il Codice dell'amministrazione digitale, che ha depauperato, per quanto concerne gli aspetti informatici, il corpus normativo formato dal T.U. 445/2000, scorporando, appunto, le disposizioni sul documento informatico e sulle firme elettroniche, sulla base della considerazione che esse meglio figurano in un contesto normativo destinato a regolare l'uso, da parte delle P.A., dello strumento operativo costituito dalle nuove tecnologie dell'informazione e della comunicazione, contesto che vede siffatti strumenti non più sotto il profilo statico della *documentazione* ma li colloca in quello dinamico dell'organizzazione e del funzionamento.

Attualmente, pertanto, le disposizioni di livello legislativo concernenti il documento informatico e le firme elettroniche si trovano quasi esclusivamente nel CAD, con l'eccezione di pochissime disposizioni (come quella del più volte citato art. 15 della L. 59/1997): è peraltro prevedibile che anche in futuro le norme relative a tali oggetti confluiranno nel medesimo CAD, anche in attuazione dell'art. 73 del Codice medesimo (*Aggiornamenti*), il quale stabilisce che:

> *la Presidenza del Consiglio dei ministri adotta gli opportuni atti di indirizzo e di coordinamento per assicurare che i successi-*

[21] Decreto Legislativo 23 gennaio 2002, n. 10, *Attuazione della direttiva 1999/93/CE relativa ad un quadro comunitario per le firme elettroniche*, in Gazzetta Ufficiale del 15 febbraio 2002, n. 39.

vi interventi normativi, incidenti sulle materie oggetto di riordino siano attuati esclusivamente mediante la modifica o l'integrazione delle disposizioni contenute nel presente Codice.

5. La disciplina positiva del documento informatico

Fatte queste premesse di ordine generale si potranno ora esaminare le disposizioni contenute nel capo secondo, sezione prima, del CAD, relative al documento informatico.

Come si è già ricordato il documento informatico è definito dall'art. 1, comma 1, lett. P) del CAD *«la rappresentazione informatica di atti, fatti o dati giuridicamente rilevanti»*, dizione riprodotta, senza modifiche, dal Testo Unico sulla documentazione amministrativa di cui al decreto n. 445 del 2000.

La definizione di documento informatico è perfettamente compatibile con le considerazioni dogmatiche sopra svolte circa il concetto di documento e la sua *species* informatica; il documento è certamente *rappresentazione*, nel senso di strumento che rappresenta e rende percepibile, mentre l'aggettivo *informatica* fa riferimento, sintetico ma inequivoco, al fatto che lo strumento è costituito dai bit, ospitati su un idoneo supporto elettronico e percepiti attraverso l'uso di un computer.

Data la definizione è innanzi tutto evidente che quella ulteriore di *documento amministrativo informatico*, a cui si dedicherà più avanti specifica attenzione, deriverà dalla combinazione delle definizioni di documento informatico e di documento amministrativo, anch'essa contenuta nel D.P.R. 445 del 2000; esso sarà, pertanto la rappresentazione informatica di atti, anche interni, delle pubbliche amministrazioni o, comunque, utilizzati ai fini dell'attività amministrativa; o, secondo la definizione già proposta in dottrina, l'oggetto informatico che, attraverso un computer, dia all'utente la rappresentazione di un atto di diritto pubblico o privato anche interno, formato dalle pubbliche ammi-

nistrazioni o da terzi, purché utilizzato per l'attività amministrativa (Masucci, 2011)[22].

Una volta fissata la definizione del documento informatico diviene necessario stabilirne la disciplina, regolandone le modalità di formazione, trasmissione, conservazione, duplicazione, riproduzione e validazione temporale, sia in generale (dunque con norme civilistiche) sia con riferimento al documento amministrativo (con disposizioni pubblicistiche), nonché fissandone il valore giuridico e l'efficacia probatoria.

Il CAD è intervenuto in modo significativo su questo tema, dettando le regole generali necessarie e lasciando alla fonte regolamentare la disciplina tecnica.

Al documento informatico è dedicata la sezione prima del capo secondo, che contiene gli articoli dal 20 al 23 quater, che saranno ora esaminati.

L'art. 20 è rubricato *Documento informatico*; la disposizione si riallaccia al soppresso testo dell'art. 8, comma 1, del D.P.R. 445/2000, norma che nel contesto del T.U. sulla documentazione amministrativa aveva natura regolamentare: la disposizione pone, però, un principio generale di diritto civile di centrale importanza rispetto al documento informatico, che assume una rilevanza notevolissima nello svolgimento dei negozi giuridici privati e dell'attività amministrativa, cioè, in una parola, nello svolgimento di ogni attività giuridica, cosicché è apparso assolutamente opportuno elevarne, nel CAD, il livello a rango primario.

Per una corretta comprensione dell'art. 20 va subito evidenziato che la norma si riferisce al documento informatico a prescindere dalla sua sottoscrizione; ai documenti sottoscritti sono dedicati gli articoli successivi.

La norma, nella sua attuale formulazione, così recita per quanto qui interessa:

---

22 MASUCCI, A, *op.cit*. p. 184.

1. Il documento informatico da chiunque formato, la memorizzazione su supporto informatico e la trasmissione con strumenti telematici conformi alle regole tecniche di cui all'articolo 71 sono validi e rilevanti agli effetti di legge, ai sensi delle disposizioni del presente codice.

1-bis. L'idoneità del documento informatico a soddisfare il requisito della forma scritta e il suo valore probatorio sono liberamente valutabili in giudizio, tenuto conto delle sue caratteristiche oggettive di qualità, sicurezza, integrità ed immodificabilità, fermo restando quanto disposto dall'articolo 21.

3. Le regole tecniche per la formazione, per la trasmissione, la conservazione, la copia, la duplicazione, la riproduzione e la validazione temporale dei documenti informatici, nonché quelle in materia di generazione, apposizione e verifica di qualsiasi tipo di firma elettronica avanzata, sono stabilite ai sensi dell'articolo 71. La data e l'ora di formazione del documento informatico sono opponibili ai terzi se apposte in conformità alle regole tecniche sulla validazione temporale.

4. Con le medesime regole tecniche sono definite le misure tecniche, organizzative e gestionali volte a garantire l'integrità, la disponibilità e la riservatezza delle informazioni contenute nel documento informatico.

5. Restano ferme le disposizioni di legge in materia di protezione dei dati personali.

5-bis. Gli obblighi di conservazione e di esibizione di documenti previsti dalla legislazione vigente si intendono soddisfatti a tutti gli effetti di legge a mezzo di documenti informatici, se le procedure utilizzate sono conformi alle regole tecniche dettate ai sensi dell'articolo 71.

La disposizione trova il proprio antecedente più lontano nel tempo nel già ricordato art. 15, comma 2, della legge 15 marzo 1997, n. 59, per quanto concerne l'affermazione della rilevanza giuridica del documento informatico e della sua trasmissione, e sotto il profilo della rilevanza delle regole tecniche nella già citata norma dell'art. 8 del D.P.R. 445/2000, che conteneva formula simile ma non identica; occorre dunque chiarire perché la legge si occupi delle caratteristiche tecniche del documento informatico e non si rimetta alla realtà fattuale, come avviene per altri supporti, quale reale interpretazione e quali conseguenti limiti applicativi abbia la norma ed in cosa l'attuale formulazione si sia discostata dalle precedenti.

La ragione della rilevanza della disciplina tecnica sta nella necessità di evitare che le caratteristiche oggettive del documento riferibili alla sua formazione, conservazione o trasmissione possano metterne in dubbio l'attendibilità, l'affidabilità e l'immutabilità. Senza giungere ad un approfondimento del tema relativo alle problematiche tecniche, si dirà qui, in sintesi, che è fattualmente possibile formare documenti che non abbiano le ricordate caratteristiche, essenziali ai fini della certezza dei rapporti giuridici; si pensi, ad esempio, al fenomeno dei c.d. *campi variabili*, per cui è possibile fornire al documento delle *istruzioni* in base alla quali esso stesso può modificare automaticamente il proprio contenuto senza un nuovo intervento umano.

Dunque a tutela della certezza del diritto, dell'affidamento dei terzi e di numerosi altri interessi giuridici è necessario che il documento presenti caratteristiche tecniche che ne assicurino la conformità alla manifestazione di scienza o di volontà in esso rappresentata e l'immodificabilità nel tempo.

Perciò il documento deve essere formato, registrato e trasmesso con modalità che non compromettano quelle caratteristiche; tanto premesso si davano due diverse soluzioni per valutare la sussistenza dei descritti requisiti: o affidarsi ad un parametro legale tipizzato in una fonte normativa o rimettersi alla valuta-

zione, caso per caso, del giudice, che può avvalersi dell'ausilio del consulente tecnico.

Il CAD ha optato, in coerenza con la normativa pregressa, in prima battuta per l'imposizione di un corpo essenziale di regole tecniche, affidate appunto ad una fonte secondaria rappresentata da Decreti governativi come può leggersi nel citato art. 71 del CAD:

> *art. 71.* Regole tecniche. *1. Le regole tecniche previste nel presente codice sono dettate, con decreti del Presidente del Consiglio dei Ministri o del Ministro delegato per la pubblica amministrazione e l'innovazione, di concerto con i Ministri competenti, sentita la Conferenza unificata di cui all'articolo 8 del decreto legislativo 28 agosto 1997, n. 281, ed il Garante per la protezione dei dati personali nelle materie di competenza, previa acquisizione obbligatoria del parere tecnico di DigitPA.*

Attualmente sono rinvenibili regole tecniche applicabili al documento informatico nel D.P.C.M. 30-3-2009[23], recante *Regole tecniche in materia di generazione, apposizione e verifica delle firme digitali e validazione temporale dei documenti informatici*.

Si è però in attesa di una più completa disciplina; dal 5 agosto 2011 è pubblicata sul sito dell'Agenzia per l'Italia digitale (già DigitPA) una bozza di *Regole tecniche in materia di formazione, trasmissione, conservazione, copia, duplicazione, riproduzione e validazione temporale dei documenti informatici, nonché di formazione e conservazione dei documenti informatici delle pubbliche amministrazioni ai sensi degli articoli 20, 22, 23-bis, 23-ter, 40, comma 1, 41 e 71, comma 1, del Codice del-*

[23] Decreto del Presidente del Consiglio dei Ministri 30 marzo 2009, *Regole tecniche in materia di generazione, apposizione e verifica delle firme digitali e validazione temporale dei documenti informatici*, in Gazzetta Ufficiale del 6 giugno 2009, n.129.

l'amministrazione digitale di cui al decreto legislativo n. 82 del 2005[24]; ma tale bozza, per poco comprensibili ritardi politico-burocratici, non ha ancora ottenuto un'approvazione definitiva e non è dunque ancora normativa vigente al momento della redazione del presente scritto.

Fermo il rispetto di queste essenziali regolazioni tecniche, tuttavia, il CAD introduce ulteriori forme di valutazione del valore giuridico del documento informatico, in relazione all'idoneità del documento informatico a soddisfare il requisito della forma scritta e al suo valore probatorio.

Sotto questo profilo la scelta fatta con il CAD è orientata alla coesistenza di una valutazione rimessa al giudice e di una valutazione legale tipica: la norma in esame, infatti, dispone che *«l'idoneità del documento informatico a soddisfare il requisito della forma scritta e il suo valore probatorio sono liberamente valutabili in giudizio, tenuto conto delle sue caratteristiche oggettive di qualità, sicurezza, integrità ed immodificabilità»*; tuttavia precisa *«fermo restando quanto disposto dall'articolo 21»*, articolo che reca disposizioni che tipizzano la valutazione del documento sottoscritto con le caratteristiche tecniche ivi descritte.

In altri termini: il documento informatico a prescindere dalla sua sottoscrizione, disciplinato dall'art. 20 del CAD, per meritare di essere scrutinato dal giudice sotto il profilo della generale rilevanza giuridica deve rispettare le regole tecniche emanate ex art. 71 del CAD; superato questo primo vaglio il giudice dovrà comunque effettuare una seconda e specifica valutazione caso per caso considerando «*l'idoneità del documento informatico a soddisfare il requisito della forma scritta e il suo valore probatorio»*, avendo come oggetto di valutazione, comunque, le *«ca-*

[24] <http://www.digitpa.gov.it/sites/default/files/normativa/Elenco%20provvedimenti%20Documento%20informatico%20e%20documento%20amministrativo.pdf>.

ratteristiche oggettive di qualità, sicurezza, integrità ed immodificabilità» del singolo documento.

Come ben si vede non vi è, nella disposizione recata dal comma 2, alcun riferimento al decreto sulle regole tecniche né ad altra regolazione esplicita: evidentemente, dunque, il giudice sarà libero di valutare se il documento risponda, oggettivamente, alle esigenze di permanenza temporale e non modificabilità; va da sé che se esso sarà stato formato nel rispetto delle regole tecniche relative (ove emanate, ovviamente) la valutazione legale tipica vincolerà il giudicante. Ma qualora quelle regole non esistano ovvero non siano state rispettate sarà comunque possibile, attraverso gli opportuni accertamenti tecnici, valutarne la rispondenza ai requisiti generali di legge e dunque ritenere il documento idoneo a concretizzare la forma scritta e ad acquisire valenza probatoria.

Ovviamente l'avere attribuito al documento così valutato il requisito della forma scritta implica il riconoscimento ampio del valore legale del medesimo; la scelta del legislatore è stata dunque quella da una parte di attribuire valore legale certo ai documenti (nonché alla loro trasmissione e conservazione) rispondenti a requisiti tecnici tipizzati, ma anche quella di non limitare in tale ambito il novero dei documenti rilevanti, al fine di non ingabbiare l'uso dell'informatica in maglie troppo strette, consentendo quindi anche una libera valutazione di altri documenti, purché idonei a fornire determinate garanzie; per quanto concerne la fonte di regolazione tecnica ci si è rimessi ad un atto di normazione secondaria, giacché cristallizzare siffatte regole in una previsione di rango primario avrebbe reso difficilmente aggiornabili le regole, che rischiano di divenire presto superate e obsolete alla luce della rapidità dello sviluppo tecnologico.

Alla stessa ratio risponde il comma 3 del medesimo art. 20 del codice, che completa quanto stabilito dai commi precedenti; la disposizione ribadisce che le regole tecniche sono dettate ai sensi dell'art. 71 e declina in modo analitico varie operazioni effet-

tuabili sul documento informatico che sono suscettibili di una regolazione tecnica, cioè la formazione, la trasmissione, la conservazione, la duplicazione, la riproduzione e la validazione temporale. Quest'ultima riceve dal secondo periodo della disposizione un'ulteriore attribuzione di rilevanza giuridica, giacché si prevede che la data e l'ora di formazione del documento informatico siano opponibili ai terzi se apposte in conformità alle regole tecniche sulla validazione temporale.

A completamento delle disposizioni descritte si pone il comma 4 dell'art. 20, che prevede l'emanazione di regole tecniche anche per definire le misure tecniche, organizzative e gestionali volte a garantire l'integrità, la disponibilità e la riservatezza delle informazioni contenute nel documento informatico. Anche in questo caso, come è evidente, obiettivo del legislatore è quello di offrire un parametro normativo mirante a garantire la certezza dei rapporti giuridici, intervenendo sul documento non solo al momento della sua formazione ma anche nella fase successiva in cui il documento viene conservato o fruito. Si tratta, anche in questo caso, di una disposizione che si pone in una linea di continuità del pregresso art. 20 del D.P.R. 445/2000.

Sempre in termini di continuità con la norma pregressa si pone poi il comma 5 dell'art. 20 in esame, che ribadisce come *«restano ferme le disposizioni di legge in materia di dati personali»*.

Dunque l'art. 20, comma 1 del Codice ribadisce quanto peraltro già affermato in via generale nell'ordinamento italiano per la prima volta in modo esplicito dall'art. 15 della legge n. 59 del 1993, cioè la rilevanza e validità giuridica del documento informatico e della sua trasmissione, condizionati, tuttavia, all'attendibilità tecnica del documento medesimo, la cui valutazione è, a seconda dei casi, sia tipizzata che rimessa al giudice.

Al documento sottoscritto è invece dedicato l'art. 21 del CAD, ampiamente riformato dal D.Lgs. n. 235 del 2010, nel 2011 e da ultimo ritoccato ulteriormente dal già ricordato D.L.

18 ottobre 2012, n. 179, come convertito dalla legge 17 dicembre 2012, n. 221:

21. Documento informatico sottoscritto con firma elettronica.

1. *Il documento informatico, cui è apposta una firma elettronica, sul piano probatorio è liberamente valutabile in giudizio, tenuto conto delle sue caratteristiche oggettive di qualità, sicurezza, integrità e immodificabilità.*

2. *Il documento informatico sottoscritto con firma elettronica avanzata, qualificata o digitale, formato nel rispetto delle regole tecniche di cui all'articolo 20, comma 3, che garantiscano l'identificabilità dell'autore, l'integrità e l'immodificabilità del documento, ha l'efficacia prevista dall'articolo 2702 del codice civile. L'utilizzo del dispositivo di firma elettronica qualificata o digitale si presume riconducibile al titolare, salvo che questi dia prova contraria.*

2-bis. Salvo quanto previsto dall'articolo 25, le scritture private di cui all'articolo 1350, primo comma, numeri da 1 a 12, del codice civile, se fatte con documento informatico, sono sottoscritte, a pena di nullità, con firma elettronica qualificata o con firma digitale. Gli atti di cui all'art. 1350, numero 13), del codice civile soddisfano comunque il requisito della forma scritta se sottoscritti con firma elettronica avanzata, qualificata o digitale.

3. *L'apposizione ad un documento informatico di una firma digitale o di un altro tipo di firma elettronica qualificata basata su un certificato elettronico revocato, scaduto o sospeso equivale a mancata sottoscrizione. La revoca o la sospensione, comunque motivate, hanno effetto dal momento della pubblicazione, salvo che il revocante, o chi richiede la sospensione, non dimostri che essa era già a conoscenza di tutte le parti interessate.*

4. *omissis*

5. *Gli obblighi fiscali relativi ai documenti informatici ed alla loro riproduzione su diversi tipi di supporto sono assolti secondo le modalità definite con uno o più decreti del Ministro dell'economia e delle finanze, sentito il Ministro delegato per l'innovazione e le tecnologie.*

Come ben si vede innanzi tutto siamo di fronte non più ad un mero documento informatico ma ad un documento sottoscritto con firma elettronica; in questo caso interviene, in taluni casi, la valutazione generale ed astratta del legislatore volta ad attribuire una particolare rilevanza al documento così sottoscritto proprio in relazione alle sue caratteristiche di stabilità e certezza.

Innanzitutto, per consentire una comprensione di questa disposizione, vanno forniti sintetici ragguagli sulle firme elettroniche.

Il CAD riconosce, in armonia con la normative europea, 4 tipi di firme, descritti nell'art. 1 lett. q e ss.; si tratta di: q) firma elettronica: q-bis) firma elettronica avanzata: r) firma elettronica qualificata: s) firma digitale[25].

[25] Si riportano di seguito le definizioni complete: *«q) firma elettronica: l'insieme dei dati in forma elettronica, allegati oppure connessi tramite associazione logica ad altri dati elettronici, utilizzati come metodo di identificazione informatica; q-bis) firma elettronica avanzata: insieme di dati in forma elettronica allegati oppure connessi a un documento informatico che consentono l'identificazione del firmatario del documento e garantiscono la connessione univoca al firmatario, creati con mezzi sui quali il firmatario può conservare un controllo esclusivo, collegati ai dati ai quali detta firma si riferisce in modo da consentire di rilevare se i dati stessi siano stati successivamente modificati;r) firma elettronica qualificata: un particolare tipo di firma elettronica avanzata che sia basata su un certificato qualificato e realizzata mediante un dispositivo sicuro per la creazione della firma; s) firma digitale: un particolare tipo di firma elettroni-*

Una trattazione dei vari tipi di firma esula dalla materia del presente scritto; si dirà solo che le varie tipologie, in relazione alla caratteristiche tecniche di ciascuna, garantiscono maggiore o minore attendibilità e sicurezza, dalla meno sicura (elettronica) alla più forte (digitale).

Il comma 1 riguarda il documento con firma elettronica semplice, c.d. *debole*; in questo caso, fermo restando che per quanto concerne la forma scritta è applicabile l'art. 20, in merito all'efficacia probatoria la norma affida di nuovo la valutazione al prudente apprezzamento del giudice, da effettuarsi secondo i parametri già descritti.

Il comma 2 si riferisce invece a documenti sottoscritti con firma elettronica avanzata, qualificata o digitale, cioè con firme c.d. *forti*, più sicure, che assumono un'efficacia probatoria tipizzata attraverso un richiamo all'art. 2702 del codice civile.

Il documento informatico sottoscritto con una delle firme c.d. *forti* fa dunque piena prova, fino a querela di falso, della provenienza delle dichiarazioni ivi contenute da chi l'ha sottoscritta, come accade per il documento cartaceo sottoscritto con firma autografa.

Tuttavia il documento informatico è addirittura più efficace, in concreto, sul piano probatorio rispetto al cartaceo se sottoscritto con firma elettronica qualificata o digitale poiché, mentre l'apparente sottoscrittore con firma autografa può disconoscere la propria firma ribaltando l'onere della prova dell'autenticità della firma sulla controparte, al contrario il sottoscrittore con firma avanzata, qualificata o digitale del documento informatico

ca avanzata basata su un certificato qualificato e su un sistema di chiavi crittografiche, una pubblica e una privata, correlate tra loro, che consente al titolare tramite la chiave privata e al destinatario tramite la chiave pubblica, rispettivamente, di rendere manifesta e di verificare la provenienza e l'integrità di un documento informatico o di un insieme di documenti informatici».

vedrà gravare su di sé l'onere probatorio poiché *«l'utilizzo del dispositivo di firma si presume riconducibile al titolare, salvo che questi dia prova contraria»*. Nel caso in cui, invece, sia utilizzata la firma avanzata riprendono vigore le norme in materia di onere della prova vigenti per il documento cartaceo, giacché la recente novella del 2012 ha ritenuto di poter assicurare la maggior rilevanza probatoria alle sole due firme più *forti*, appunto la qualificata e la digitale.

Per quanto concerne la forma scritta, ricordato che essa viene richiesta dall'ordinamento (*ad substantiam* o *ad probationem*) a fronte di particolari esigenze di autenticità, certezza e stabilità nel tempo del documento, esigenze riconducibili alla particolare rilevanza sociale o economica dell'atto o fatto da rappresentare nel documento, il legislatore del 2010 aveva ritenuto di dover distinguere non solo tra firma elettronica da un lato (valuta il giudice) e firme avanzata, qualificata e digitale dall'altro (si argomenta dalle disposizioni in esame che i documenti con tali firme abbiano sempre valenza di forma scritta) ma anche all'interno di queste ultime tre tipologie.

Infatti solo con riferimento ai documenti sottoscritti con firma qualificata e digitale (i due tipi più sicuri) al legislatore era apparso opportuno riconoscere esplicitamente per legge che il documento informatico è idoneo a fornire il rilevante grado di autenticità, certezza e stabilità necessario per sottoscrivere *«le scritture private di cui all'articolo 1350, primo comma, numeri da 1 a 12, del codice civile»*, cioè, in sostanza e in sintesi, gli atti diretti a costituire, modificare o trasferire diritti su beni immobili.

Successivamente, tuttavia, con la citata L. 179/2012, il legislatore, evidentemente alla luce dell'ulteriore evoluzione tecnologica e dell'affermarsi di tipi di sottoscrizione *avanzata* sempre più sicure, pur senza giungere ad equiparare la firma avanzata alla qualificata e alla digitale anche ai fini dell'utilizzabilità della medesima per la redazione degli atti per cui è richiesta la forma scritta *ad substantiam*, ha, come può leggersi nel comma 2 bis dell'art.

21 sopra riportato, previsto che *«gli atti di cui all'art. 1350, numero 13) del codice civile»* possano firmarsi con firma avanzata.

Si tratta degli *«altri atti specialmente indicati dalla legge*», come recita il predetto numero 13), cioè di una categoria ulteriore e residuale rispetto a quelle di cui ai numeri precedenti, individuata in diverse fonti legislative primarie; si tratta, evidentemente, di categoria ritenuta dal legislatore meno *sensibile* ad esigenze di particolare certezza formale.

In via generale, comunque, affinché le esigenze di certezza giuridica siano pienamente soddisfatte, la norma richiede che il documento sottoscritto sia formato nel rispetto delle regole tecniche stabilite ai sensi dell'art. 71 e che queste garantiscano l'identificabilità dell'autore e l'integrità e l'immodificabilità del documento: pur se la norma non lo afferma in maniera esplicita, tali caratteristiche del documento possono derivare, per ragioni tecniche, dalla stessa apposizione della firma c.d. *forte*.

La previsione va intesa, quindi, nel senso che le regole tecniche possono riguardare tanto il documento quanto la sua sottoscrizione ed in secondo luogo nel senso che l'esercizio del potere regolamentare ex art. 71 del CAD è parzialmente vincolato nell'oggetto e nei fini, dovendo necessariamente assicurare sia la possibilità di identificare con certezza l'autore del documento, sia la piena conformità, al momento della formazione e sottoscrizione e poi nel tempo, del documento al contenuto che l'autore vi abbia immesso e la non modificabilità del contenuto.

Occorre precisare che per *immodificabilità* si intende l'impossibilità di modificare comunque il contenuto del documento, sia in modo automatico a prescindere dall'intervento umano, che con il voluto e diretto intervento dell'uomo; ben si comprende, quindi, perché il legislatore abbia scelto di attribuire con valutazione generale al documento sottoscritto con firma qualificata o digitale il massimo rilievo sotto il profilo del requisito della forma scritta, che viene integrata dal documento così sottoscritto anche quando è richiesta a pena di nullità, cioè *ad substantiam*;

attraverso quel tipo di firme, infatti, è tecnicamente possibile *blindare*, rendendolo immodificabile o comunque rendendo percepibile l'intervento modificativo successivo, il contenuto originario del documento.

Si ribadisce che questa scelta non fa venir meno per il giudice la possibilità di rinvenire la forma scritta anche in altri tipi di documenti informatici, sottoscritti o non, ma non potrà negarla al documento informatico con firma avanzata, qualificata o digitale, che potrà quindi essere utilizzato senza tema di errore o smentita in ogni occasione in cui sia necessario un negozio formale; dunque, alla luce del testo dell'art. 21, comma 2 bis, come di recente riformato, laddove la forma scritta sia richiesta dalla legge per la validità del negozio dovrà più necessariamente essere utilizzata la firma digitale o qualificata solo per i negozi relativi a diritti su immobili, potendosi far ricorso anche alla firma avanzata negli altri casi previsti dall'ordinamento.

Riassumendo: il documento informatico e le firme elettroniche per assumere rilevanza giuridica devono rispettare le vigenti regole tecniche; laddove si discuta in giudizio della validità di *«scritture private di cui all'articolo 1350, primo comma, numeri da 1 a 12, del codice civile»*, il giudice dovrà semplicemente accertare che la sottoscrizione del documento informatico sia digitale o qualificata, mentre per i casi di cui al numero 13 basterà anche la firma avanzata, salvo che sia necessario accertare anche se l'utilizzo del dispositivo di firma sia effettivamente riconducibile al titolare; ove sia in esame l'efficacia probatoria del documento informatico, qualora la sottoscrizione sia avanzata, qualificata o digitale il giudice non potrà negarla (con le ricordate differenze in materia di onere della prova tra qualificata e digitale da un lato e avanzata dall'altro), a fronte della valutazione legale tipica; in caso di documento non sottoscritto, o sottoscritto con firma elettronica semplice, compito del giudice sarà quello di valutare, secondo il proprio prudente apprezzamento, tali caratteristiche, per dedurne, eventualmente anche alla luce di altri

elementi, il valore giuridico; in questa operazione è ben possibile che il magistrato possa farsi assistere da un consulente tecnico, ferma restando la valutazione finale da parte del giudicante, e può anche giungere ad attribuire valore giuridico a documenti e sottoscrizioni non rispondenti alle regole tecniche.

6. Documenti amministrativi informatici

L'art. 23 ter del CAD così recita:

23 TER. Documenti amministrativi informatici 1. Gli atti formati dalle pubbliche amministrazioni con strumenti informatici, nonché i dati e i documenti informatici detenuti dalle stesse, costituiscono informazione primaria ed originale da cui è possibile effettuare, su diversi o identici tipi di supporto, duplicazioni e copie per gli usi consentiti dalla legge.

2. I documenti costituenti atti amministrativi con rilevanza interna al procedimento amministrativo sottoscritti con firma elettronica avanzata hanno l'efficacia prevista dall'art. 2702 del codice civile.

Seguono ulteriori 5 commi in materia di copie, di regole tecniche per la formazione e conservazione dei documenti informatici delle P.A., di fruibilità e accessibilità ai disabili e di richiamo alle norme generali sul documento informatico.

Come si vede non vi sono disposizioni divergenti dalle precedenti circa il valore giuridico del documento sol perché esso è definibile come *amministrativo*: vale dunque quanto finora si è detto sul documento informatico in generale.

La presenza del comma 2 appare, anzi, ultronea, poiché residuo ormai superfluo di una precedente versione preparatoria del D.Lgs. 30 dicembre 2010, n. 235 che recava una diversa disciplina della firma avanzata.

Ciò che invece interessa è sottolineare che il comma 1 prevede che gli atti formati dalla P.A. debbano nascere informatici e che, dunque, l'atto amministrativo *è* informatico.

Dunque non occorre il documento cartaceo affinché venga ad esistenza giuridica un atto o un provvedimento amministrativo; anzi, alla luce di ulteriori disposizioni del CAD (cfr. in particolare art. 40, comma 1) le P.A. *«formano gli originali dei propri documenti con mezzi informatici»*.

Il CAD prevede quindi un vero e proprio obbligo per le Amministrazioni pubbliche di formare *gli originali* dei propri atti in modalità informatica, come momento essenziale della dematerializzazione dell'intero ciclo dell'attività documentale amministrativa che si sviluppa poi nello scambio di corrispondenza tra P.A. attraverso la posta elettronica (art. 47), nella gestione dei procedimenti amministrativi per mezzo di fascicoli informatici (art. 41), nella conservazione degli atti in modalità informatica (art. 43).

Spiace dover constatare che la previsione dell'art. 40 è, come ben noto, al momento ancora ampiamente disattesa.

7. Le copie

Il CAD dedica inoltre numerosi articoli al tema delle copie ed esattamente i seguenti:

- Art. 22. Copie informatiche di documenti analogici:
  Al riguardo si dirà, in breve, che le copie per immagine su supporto informatico di documenti originali analogici hanno la stessa efficacia probatoria degli originali da cui sono estratte, se la loro conformità è attestata da un notaio o da altro pubblico ufficiale a ciò autorizzato o se la loro conformità all'originale non è espressamente disconosciuta. Dunque se non vi è copia la cui conformità è attestata

da pubblico ufficiale e che viene disconosciuta, la copia informatica perde la capacità probatoria e non può sostituire l' originale.
Inoltre con decreto del Presidente del Consiglio dei Ministri possono essere individuate particolari tipologie di documenti analogici originali unici per le quali, in ragione di esigenze di natura pubblicistica, permane l'obbligo della conservazione dell'originale analogico.

- Art. 23. Copie analogiche di documenti informatici:
  Anche in questo caso le copie su supporto analogico di documento informatico, anche sottoscritto con firma elettronica avanzata, qualificata o digitale, hanno la stessa efficacia probatoria dell'originale da cui sono tratte se la loro conformità all'originale in tutte le sue componenti è attestata da un pubblico ufficiale a ciò autorizzato o se la loro conformità non è espressamente disconosciuta.

- Art. 23-bis. Duplicati e copie informatiche di documenti informatici:
  I duplicati informatici, con tale espressione intendendosi quei documenti che sono indistinguibili dall'originale poiché riprodotti nel medesimo formato, hanno lo stesso valore giuridico, ad ogni effetto di legge, del documento informatico da cui sono tratti; le copie e gli estratti informatici del documento informatico, cioè le riproduzioni realizzate in diverso formato, hanno la stessa efficacia probatoria dell'originale da cui sono tratte se la loro conformità all'originale, in tutti le sue componenti, è attestata da un pubblico ufficiale a ciò autorizzato o se la conformità non è espressamente disconosciuta.
  Per tutti i tipi di copia la validità giuridica è condizionata dal rispetto delle regole tecniche, sempre dettate ex art. 71 del CAD.

Bibliografia

CARNELUTTI, F., *Documento. Teoria moderna*, in Novissimo Digesto italiano, vol. 6, Torino, 1957

Decreto del Presidente del Consiglio dei Ministri 30 marzo 2009, *Regole tecniche in materia di generazione, apposizione e verifica delle firme digitali e validazione temporale dei documenti informatici*, in Gazzetta Ufficiale del 6 giugno 2009, n.129

Decreto del Presidente del Consiglio dei Ministri 8 febbraio 1999, *Regole tecniche per la formazione, la trasmissione, la conservazione, la duplicazione, la riproduzione e la validazione, anche temporale, dei documenti informatici ai sensi dell'articolo 3, comma 1, del Decreto del Presidente della Repubblica, 10 novembre 1997, n. 513*, in Gazzetta Ufficiale del 15 aprile 1999, n. 87

Decreto del Presidente della Repubblica 10 novembre 1997, n. 513, *Regolamento contenente i criteri e le modalità per la formazione, l'archiviazione e la trasmissione di documenti con strumenti informatici e telematici a norma dell'articolo 15, comma 2, della legge 15 marzo 1997, n. 59*, in Gazzetta Ufficiale del 13 marzo 1998, n. 60

Decreto del Presidente della Repubblica 28 dicembre 2000, n. 445, *Testo unico delle disposizioni legislative e regolamentari in materia di documentazione amministrativa*, in Gazzetta Ufficiale del 20 febbraio 2001, n. 42

Decreto del Presidente della Repubblica 28 dicembre 2000, n. 445, *Disposizioni legislative in materia di documentazione amministrativa. (Testo A)*, in Gazzetta Ufficiale del 20 febbraio 2001, n. 42, Supplemento Ordinario n. 30

Decreto Legislativo 23 gennaio 2002, n. 10, *Attuazione della direttiva 1999/93/CE relativa ad un quadro comunitario per le firme elettroniche*, in Gazzetta Ufficiale del 15 febbraio 2002, n. 39

Decreto Legislativo 28 febbraio 2005, n. 42, *Istituzione del sistema pubblico di connettività e della rete internazionale della pubblica amministrazione, a norma dell'articolo 10, della legge 29 luglio 2003, n. 229*, in Gazzetta Ufficiale del 30 marzo 2005, n. 73

Decreto Legislativo 30 dicembre 2010, n. 235, *Modifiche ed integrazioni al decreto legislativo 7 marzo 2005, n. 82, recante Codice dell'amministrazione digitale, a norma dell'articolo 33 della legge 18 giugno 2009, n. 69*, in Gazzetta Ufficiale del 10 gennaio 2011, n. 6, Supplemento Ordinario n. 8

Decreto Legislativo 4 aprile 2006, n. 159, *Disposizioni integrative e correttive al decreto legislativo 7 marzo 2005, n. 82, recante codice dell'ammini-*

strazione digitale, in Gazzetta Ufficiale del 29 aprile 2006, n. 99, Supplemento Ordinario n. 105

Decreto Legislativo 7 marzo 2005, n. 82, Codice dell'amministrazione digitale, in Gazzetta Ufficiale del 16 maggio 2005, n. 112, Supplemento Ordinario n. 93

Decreto Legge 18 ottobre 2012, n. 179, *Ulteriori misure urgenti per la crescita del Paese*, in Gazzetta Ufficiale del 19 ottobre 2012, n. 245, Supplemento Ordinario n. 194

DEVOTO, G., OLI, G., *Vocabolario della lingua italiana*, Le Monnier, 2011

Direttiva 1999/93/CE del Parlamento europeo e del Consiglio del 13 dicembre 1999 Relativa ad un quadro comunitario per le firme elettroniche, in Gazzetta Ufficiale delle Comunità europee del 13 dicembre 1999

Legge 15 marzo 1997, n. 59, *Delega al Governo per il conferimento di funzioni e compiti alle regioni ed enti locali, per la riforma della pubblica amministrazione e per la semplificazione amministrativa*, in Gazzetta Ufficiale del 17 marzo 1997, n. 63, Supplemento Ordinario n. 56

Legge 17 dicembre 2012, n. 221, *Conversione in legge, con modificazioni, del decreto-legge 18 ottobre 2012, n. 179, recante ulteriori misure urgenti per la crescita del Paese*, in Gazzetta Ufficiale del 18 dicembre 2012, n. 294, Supplemento Ordinario n. 208

Legge 29 luglio 2003, n. 229, *Interventi in materia di qualità della regolazione, riassetto normativo e codificazione* - Legge di semplificazione 2001, in Gazzetta Ufficiale del 25 agosto 2003, n. 196

Legge 7 agosto 1990, n. 241, *Nuove norme in materia di procedimento amministrativo e di diritto di accesso ai documenti amministrativi*, in Gazzetta Ufficiale del 18 agosto 1990, n. 192

LUZZATTO,G.I., *Documento-Diritto romano*, in Novissimo Digesto Italiano, vol. 6, Torino, 1960

MASUCCI, A., *La documentazione amministrativa*, in La Documentazione amministrativa. Certezze, semplificazione e informatizzazione nel d.P.R. n. 28 dicembre 2000, n. 445, Rimini, Maggioli, 2011

TRABUCCHI, A., CIAN, G., *Commentario breve al Codice Civile*, Padova, CEDAM, 1984

ZAGAMI, R., *Firma digitale e sicurezza giuridica*, Padova, 2000

Sitografia

<http://www.digitpa.gov.it/sites/default/files/normativa/Elenco%20provvedimenti%20Documento%20informatico%20e%20documento%20amministrativo.pdf>

Sistemi informativi e dematerializzazione

STEFANO PIGLIAPOCO*

1. Evoluzione dei sistemi informativi

Una classica definizione di sistema informativo lo descrive come *un insieme di elementi interconnessi che acquisiscono, elaborano, memorizzano e distribuiscono le informazioni necessarie per l'esecuzione dei processi di un'organizzazione, sia quelli di natura operativa sia quelli a carattere gestionale o decisionale*. Il sistema informatico, cioè l'insieme delle reti e degli apparati hardware e software utilizzati per la gestione della risorsa informazione, è un componente del sistema informativo, non l'unico, ma forse quello che maggiormente lo caratterizza.

Fino agli anni '80, l'informazione gestita in forma digitale era sostanzialmente rappresentata da dati elementari opportunamente acquisiti, trattati e memorizzati in strutture informatiche che inizialmente erano basate su file ad accesso sequenziale o casuale (*sequential or indexed file*) e poi su sistemi DBMS (*Data Base Management System*). L'importanza strategica di questi dati, l'esigenza di garantire la loro sicurezza e riservatezza oltre che una gestione efficiente e affidabile, ha determinato nell'arco di pochi decenni un enorme sviluppo dei sistemi di gestione delle banche dati, che oggi presentano funzionalità avanzate sia per

* Università degli Studi di Macerata, Dipartimento di Studi Umanistici – lingue, mediazione, storia, lettere, filosofia.

applicazioni di tipo OLTP (*On Line Transaction Processing*), cioè orientate alla trattazione online di grandi quantità di dati, sia per la realizzazione di sistemi OLAP (*On Line Analytical Processing*) finalizzati all'analisi dei dati con tecniche anche basate sull'intelligenza artificiale.

Dagli anni '90 in poi, l'evoluzione dei prodotti di *office automation*, lo sviluppo delle reti di computer e l'uso generalizzato dei servizi Internet hanno permesso alle organizzazioni di arricchire il loro patrimonio documentale con oggetti digitali in forma testuale, ipertestuale e multimediale. Il Web si configura come la più grande collezione di testi esistente, con una dimensione così elevata da sfuggire a ogni stima attendibile e in continuo aumento, anche grazie allo sviluppo dei sistemi CMS (*Content Management Systems*) che consentono ai creatori di contenuti informativi di pubblicarli su Internet in modo facile e immediato. La posta elettronica è considerata dagli analisti delle ICT (*Information and Communication Technologies*) una *killer application*, nel senso che ha raggiunto una percentuale di utilizzatori così elevata da far ritenere che in ogni computer connesso alla rete ci sia almeno un indirizzo email attivo. Abnorme è anche la quantità di messaggi di posta elettronica che circolano quotidianamente in rete, molti dei quali hanno un contenuto informativo rilevante per le organizzazioni essendo memorie di attività svolte, decisioni assunte, autorizzazioni concesse, interesse mostrato dai clienti verso prodotti o servizi. Anche l'evoluzione delle strategie di marketing, da una concezione tradizionale, dove le transazioni tra i produttori dei beni e gli acquirenti erano viste come eventi occasionali di compravendita, a una più moderna, dove l'obiettivo è la continuità del rapporto con i clienti per creare un clima di fiducia verso l'azienda, ha favorito lo sviluppo dei sistemi CRM (*Customer Relationship Management*) che gestiscono le comunicazioni come entità multimediali memorizzate in grandi contenitori digitali.

Oggi, ai sistemi informativi è richiesta un'altra evoluzione. La

grave crisi economica di questi anni, infatti, sta spingendo le organizzazioni verso soluzioni che permettono di ridurre i costi di gestione ed erogare i servizi in modo più efficiente, sfruttando le potenzialità delle tecnologie ICT per accedere, in ogni momento e in qualsiasi luogo, alle informazioni necessarie per prendere decisioni o avviare azioni. Da qui il progetto, sostenuto da tutti i governi a livello internazionale, di dematerializzare i processi delle organizzazioni[1]. L'obiettivo è quello di produrre la minor quantità possibile di carta, riducendo i costi connessi e rendendo più veloci i processi inerenti alle attività decisionali, alle comunicazioni e all'erogazione dei servizi. D'altra parte, come sopra rilevato, le organizzazioni possiedono già gli strumenti necessari: sono dotate di sistemi DBMS evoluti, di siti Web interattivi, di grandi contenitori di oggetti multimediali e di dispositivi che permettono di svolgere le attività su base digitale.

La finalità di un buon progetto di dematerializzazione, però, non è soltanto quella di gestire in forma digitale ogni genere d'informazione, formando basi di dati facilmente accessibili, ma anche di produrre contenuti digitali sostitutivi, a ogni effetto di legge, dei tradizionali documenti analogici[2]. A questo scopo, i

---

1 Il termine *dematerializzazione* viene usato per identificare la progressiva perdita di consistenza fisica dei documenti per effetto della transizione dal cartaceo al digitale. La dematerializzazione di un processo si realizza svolgendo le relative attività completamente, o prevalentemente, su base informatica con l'ausilio delle tecnologie dell'informazione e della comunicazione, e quindi producendo la minor quantità possibile di documenti cartacei.

2 Ai sensi dell'art. 1 della Deliberazione CNIPA 19 febbraio 2004, n. 11, *Regole tecniche per la riproduzione e conservazione di documenti su supporto ottico idoneo a garantire la conformità dei documenti agli originali*, per documento analogico s'intende il *«documento formato utilizzando una grandezza fisica che assume valori continui, come le tracce su carta (esempio: documenti cartacei), come le immagini su film (esempio: pellicole mediche, microfiche, microfilm), come le magnetizzazioni su nastro (esempio: cassette e nastri magnetici audio e video)»*.

Paesi dell'Unione europea hanno introdotto nei loro ordinamenti giuridici norme volte a disciplinare l'uso delle firme elettroniche per la produzione di documenti informatici e a favorire lo sviluppo di servizi di comunicazione digitale che garantiscono la provenienza, la sicurezza e la riservatezza dei messaggi[3].

Ai moderni sistemi informativi, pertanto, è richiesta la capacità di sostenere questi processi di dematerializzazione, rendendo disponibili sistemi per la formazione e la conservazione della memoria digitale delle organizzazioni.

2. Sistemi di gestione informatica dei documenti

La produzione di documenti informatici in sostituzione dei documenti analogici nell'ambito di processi dematerializzati determina inevitabilmente la formazione di archivi digitali, o meglio di archivi ibridi, cioè costituiti da documenti cartacei pregressi, o comunque sia non acquisibili né producibili in formato

[3] In ambito europeo, il complesso delle norme che disciplinano la produzione dei documenti informatici attraverso l'uso delle firme elettroniche poggia essenzialmente sulla Direttiva 1999/93/CE del Parlamento europeo e del Consiglio del 13 dicembre 1999 Relativa ad un quadro comunitario per le firme elettroniche, in Gazzetta Ufficiale delle Comunità europee del 13 dicembre 1999, entrata in vigore il 19 gennaio 2000. In attuazione a questa direttiva, in Italia è stato emanato il Decreto Legislativo 7 marzo 2005, n. 82, *Codice dell'Amministrazione Digitale*, in Gazzetta Ufficiale del 16 maggio 2005, n. 112, Supplemento Ordinario n. 93, che rappresenta la norma di riferimento per i processi di dematerializzazione. Il legislatore italiano, inoltre, con il Decreto del Presidente della Repubblica 11 febbraio 2005, n.68, *Regolamento recante disposizioni per l'utilizzo della posta elettronica certificata, a norma dell'articolo 27 della legge 16 gennaio 2003, n. 3*, in Gazzetta Ufficiale del 28 aprile 2005, n. 97, ha disegnato il servizio di posta elettronica certificata che è un «*sistema di comunicazione in grado di attestare l'invio e l'avvenuta consegna di un messaggio di posta elettronica e di fornire ricevute opponibili a terzi*».

digitale nativo, e documenti informatici. Di conseguenza, un moderno sistema informativo deve presentare tra i suoi elementi un sistema per la gestione e l'archiviazione dei documenti informatici, denominato sistema di gestione informatica dei documenti o sistema ERMS (*Electronic Records Management System*).

I requisiti funzionali dei sistemi di gestione dei *record* digitali, dove per *record* s'intende *«information created, received and maintained as evidence and information by an organisation or person, in pursuance of legal obligations or in the transaction of business»*[4], sono stati definiti nel 2001 da un gruppo di lavoro costituito nell'ambito del programma IDA (*Interchange of Data between Administrations*) della Commissione europea e descritti nel documento MoReq (*Model Requirement specifications*)[5].

2.1 *Integrazione con i sistemi di comunicazione*

Allo stato attuale delle tecnologie, un documento informatico può essere generato con strumenti diversi e avere forma e consistenza differente: può essere un messaggio di posta elettronica o posta elettronica certificata con o senza allegati, un fax ricevuto e memorizzato in formato digitale, un modulo elettronico compilato *online* previa identificazione informatica dell'autore, un file generato con un applicativo di *office automation* e sottoscritto con firma digitale, una registrazione su base informatica di una comunicazione telefonica, e altro ancora.

Ai sensi delle disposizioni contenute nel D. Lgs. 7 marzo 2005, n. 82, recante il Codice dell'Amministrazione Digitale

---

4 ISO 15489-1: 2001, Information and documentation – *Records Management – Part 1: General*, par. 3.15.

5 Le specifiche MoReq sono state pubblicate nel 2001 e successivamente aggiornate nel 2007 e 2010. L'attuale versione è contenuta nel documento DLM FORUM FOUNDATION, *MoReq2010®: Modular Requirements for RecordsSystems*, vol. 1: Core Services & Plug-in Modules, 2011. <http://moreq2010.eu>.

(CAD), a ognuno di questi oggetti digitali può essere attribuita una forza probatoria diversa in funzione del livello di sicurezza associato ai relativi processi di produzione e sottoscrizione. Ad esempio, in base all'art. 21, comma 1, un documento informatico sottoscritto con una firma elettronica *«sul piano probatorio è liberamente valutabile in giudizio, tenuto conto delle sue caratteristiche oggettive di qualità, sicurezza, integrità e immodificabilità»*. Se, invece, la sottoscrizione è effettuata con una firma elettronica qualificata, o una firma digitale, al documento informatico è riconosciuta l'efficacia di cui all'art. 2702 del codice civile, cioè l'efficacia di una scrittura privata che *«fa piena prova, fino a querela di falso, della provenienza delle dichiarazioni da chi l'ha sottoscritta, se colui contro il quale la scrittura è prodotta ne riconosce la sottoscrizione, ovvero se questa è legalmente considerata riconosciuta»*[6].

Sotto il profilo archivistico tutti questi documenti informatici, se ricevuti o prodotti da un'organizzazione durante lo svolgimento delle sue funzioni, ossia nell'ambito di attività pratiche, amministrative o giuridiche, devono confluire in archivio ed essere collegati ai precedenti per consentire la ricostruzione nel tempo delle vicende accadute. Ciò significa che, indipendentemente dal loro valore giuridico e dagli strumenti tecnologici utilizzati per la loro produzione, essi devono convergere su un sistema di gestione informatica dei documenti, o sistema ERMS, perché è su di esso che si forma l'archivio digitale.

Da queste considerazioni derivaun primo, importante, requisito funzionale: la piena integrazione tra il sistema di gestione informatica dei documenti e i sistemi di comunicazione utilizzati dall'organizzazione, che normalmente sono rappresentati dai

6 Riguardo alle definizioni di firma elettronica, firma elettronica avanzata, firma elettronica qualificata e firma digitale si veda l'art.1 del Codice dell'Amministrazione Digitale, mentre nell'art.21 sono riportate le disposizioni sul valore giuridico dei documenti informatici.

servizi di posta elettronica e posta elettronica certificata, apparecchi fax, rete di telefonia fissa e mobile, canali Web e posta cartacea.

L'integrazione tra il sistema ERMS e i servizi di posta elettronica e posta elettronica certificata permette di acquisire sul sistema ERMS tutti messaggi in entrata indipendentemente dai destinatari interni e di centralizzare la trasmissione di quelli in uscita. In questo modo si evita la dispersione delle email nei vari computer collegati alla rete e si assicura l'archiviazione sia dei messaggi in arrivo prima della loro trattazione da parte degli uffici competenti, sia di quelli in uscita prima della loro spedizione.

L'utilizzo di sistemi di gestione elettronica dei fax, comunemente denominati *FAX server,* direttamente connessi a un sistema ERMS permette di trattare come documenti informatici anche quelli inviati con questo mezzo di trasmissione e di archiviarli insieme alle email e alle altre unità documentarie digitali dell'organizzazione.

Quasi tutte le applicazioni informatiche che compongono un sistema informativo e hanno come finalità una qualche forma d'interazione con l'utenza utilizzano canali Web per ricevere e trasmettere dati e documenti informatici. La possibilità di disegnare e pubblicare su Web moduli elettronici intelligenti (*electronic forms*), unita alla disponibilità di strumenti per generare firme elettroniche, lascia intravedere l'attivazione di una miriade di canali Web abilitati alla ricezione e alla trasmissione di documenti informatici. Il rischio che questi documenti siano memorizzati in sistemi diversi dall'ERMS è molto alto e può essere ridimensionato solo attraverso l'integrazione funzionale di questo sistema con i canali Web, che permette di acquisire automaticamente e di archiviare su base digitale tutti i documenti che transitano attraverso di essi.

Nonostante la spinta decisa dei governi verso la digitalizzazione dei documenti e la dematerializzazione dei processi, è ra-

gionevole ritenere che nei prossimi anni le organizzazioni continueranno a ricevere e produrre una grande quantità di documenti cartacei e quindi un progetto di gestione informatica dei documenti non può non prevedere l'integrazione del sistema di gestione informatica dei documenti con i dispositivi che realizzano l'acquisizione delle immagini con processi di scansione di tipo interattivo o *batch*[7].

2.2 *Interoperabilità e cooperazione applicativa*

Un sistema informativo si compone normalmente di un elevato numero di applicazioni attivate su sistemi diversi per supportare tutte le attività dell'organizzazione, da quelle meramente operative a quelle decisionali. Con l'affermarsi dei meccanismi propri della dematerializzazione, a queste applicazioni sono state progressivamente aggiunte funzionalità volte a realizzare la gestione informatica dei documenti prodotti nell'ambito delle attività che esse automatizzano. L'uso di queste funzionalità, però, rischia di provocare una frammentazione disorganica della memoria digitale, che finisce per essere archiviata oltre che sul sistema ERMS anche su diversi altri sistemi, dove le logiche di organizzazione dei documenti rispondono esclusivamente alle esigenze degli uffici che li producono o ricevono.

Un esempio di questo modo irrazionale di formare un archivio digitale si ha nella pubblica amministrazione, dove il sistema di gestione informatica dei documenti è utilizzato per la protocollazione di quelli in arrivo e in partenza, mentre la gestione dell'iter di approvazione degli atti deliberativi, dalla presenta-

[7] Per processo di scansione interattivo s'intende la digitalizzazione dei documenti cartacei al momento della loro registrazione nel sistema ERMS. Un processo di scansione *batch*, invece, prevede la digitalizzazione di un blocco di documenti già registrati, anche numericamente consistente, con strumenti tecnologici che permettono di collegare automaticamente le immagini alle rispettive registrazioni.

zione di una proposta fino all'archiviazione dell'atto finale, è effettuta con un altro sistema. In questo modo, i documenti di un procedimento amministrativo sono memorizzati su due sistemi diversi e ciò inibisce la corretta formazione del fascicolo procedimentale. Questa situazione, già oggi problematica sotto il profilo archivistico, diventerà molto critica nei prossimi anni quando gli enti pubblici dovranno produrre, per obbligo di legge, mandati informatici, fatture elettroniche e cedolini stipendio digitali, comunicare con le imprese esclusivamente su base informatica, svolgere gare telematiche ed erogare i servizi anagrafici via Web. Se questi documenti informatici saranno gestiti con sistemi autonomi, diversi dal sistema di gestione informatica dei documenti e non integrati con esso, sarà molto difficile per le pubbliche amministrazioni garantire la corretta formazione del loro archivio digitale.

L'impiego di sistemi ERMS dotati dei meccanismi che realizzano l'interoperabilità e la cooperazione applicativa con gli altri elementi del sistema informativo è l'unica vera soluzione al problema evidenziato[8]. Infatti, con la disponibilità di questi meccanismi si possono utilizzare gli applicativi di settore per gestire i documenti prodotti nell'ambito di processi che essi automatizzano, con la certezza che quando questi documenti saranno perfezionati o pienamente acquisiti, gli stessi applicativi apriranno una sessione di comunicazione con il sistema di gestione informatica dei documenti e gli trasferiranno automaticamente le unità documentarie digitali con i relativi metadati. Si ottiene così il duplice vantaggio di automatizzare i processi dell'organiz-

[8] Per interoperabilità s'intende la capacità di due o più sistemi informatici di scambiarsi informazioni - normalmente in linguaggio XML - da utilizzare nei rispettivi contesti applicativi. Per cooperazione applicativa, invece, s'intende la capacità dei sistemi informatici di avvalersi della facoltà di elaborazione di altri sistemi informatici per acquisire le informazioni da utilizzare nel proprio contesto applicativo.

zazione con i sistemi progettati a questo scopo, che sicuramente soddisfano al meglio le esigenze operative degli uffici e di utilizzare il sistema ERMS per la formazione e la gestione dell'archivio digitale.

2.3 *Archiviazione digitale*

L'integrazione con i sistemi di comunicazione e l'implementazione dei meccanismi che realizzano l'interoperabilità e la cooperazione applicativa tra i moduli del sistema informativo, come descritto nei precedenti paragrafi, permettono di memorizzare nel sistema ERMS l'intera produzione documentaria digitale dell'organizzazione. Tuttavia, per formare un archivio digitale occorre organizzare questo complesso documentario in modo da rendere evidenti i *vincoli archivistici*, cioè l'insieme delle relazioni logiche e formali che esistono necessariamente tra i documenti di un archivio e li collegano logicamente alle attività del soggetto produttore.

A questo fine, un sistema ERMS deve presentare idonee funzionalità per eseguire le operazioni di registrazione di protocollo, classificazione dei documenti e formazione dei fascicoli archivistici. Questi ultimi sono le unità di base, indivisibili, di un archivio e la loro corretta formazione deve essere garantita specialmente in ambiente digitale, dove i documenti informatici non hanno la consistenza della carta e le relazioni esistenti tra di essi possono essere evidenziate solo con l'ausilio del sistema di gestione informatica dei documenti[9].

[9] L'approfondimento degli aspetti archivistici e organizzativi connessi alla registrazione, classificazione e fascicolazione dei documenti informatici non rientra tra le finalità di questa relazione che vuole analizzare l'evoluzione dei sistemi informativi verso l'archiviazione e la conservazione dei documenti informatici. Pertanto, per una trattazione approfondita di questi argomenti si rinvia al documento MoReq sopra richiamato e alle pubblicazioni specifiche su queste tematiche: Cfr. GIORGETTA BONFIGLIO-DO-

2.4 *Accessibilità e riservatezza*

Una tra le fasi più complesse di un progetto di gestione informatica dei documenti è senza dubbio la configurazione del sistema ERMS in modo da garantire la riservatezza del patrimonio informativo e documentario in esso memorizzato.

Per ottenere questo risultato è necessario rappresentare nel sistema ERMS la struttura organizzativa del soggetto produttore con l'articolazione gerarchica delle diverse unità, avendo cura di definire anche gli organismi intersettoriali quali le commissioni, i gruppi di lavoro e i team di progetto, che possono essere destinatari di documenti informatici al pari degli altri uffici dell'organizzazione.

In secondo luogo, occorre associare a ogni utente un profilo di autorizzazione basato sull'appartenenza a una determinata unità organizzativa, sulla qualifica attribuita e ruolo ricoperto, sugli incarichi assunti e le responsabilità assegnate. Unitamente a queste informazioni, nel profilo di autorizzazione si devono specificare le possibili operazioni che l'utente può compiere sui documenti informatici a cui ha accesso, dalla semplice visualizzazione alla formazione dei fascicoli. Ogni utente, inoltre, deve essere dotato delle credenziali di identificazione informatica, rappresentate da dati e/o dispositivi che, *«in quanto da loro conosciuti o ad essi univocamente associati, possono essere utilizzati per la loro identificazione online»*[10]. Nella forma più semplice e più diffusa, l'autenticazione informatica è realizzata asse-

SIO, *La formazione del fascicolo archivistico in ambiente digitale*, in Una mente colorata. Studi in onore di Attilio Mauro Caproni per i suoi 65 anni, Roma, Vecchiarelli, 2007, pp. 549-553; Cfr. STEFANO PIGLIAPOCO, *Il fascicolo elettronico*, in Il fascicolo elettronico, Pigliapoco S. (a cura di), Padova, SIAV Academy**,** 2010, pp. 33-52.

10 Si veda l'art.4, comma 3, del Decreto legislativo 30 giugno 2003, n. 196, *Codice in materia di protezione dei dati personali*, in Gazzetta Ufficiale del 29 luglio 2003, n. 174, Supplemento Ordinario n. 123.

gnando a ogni utente un codice identificativo (*user-id*) e una *password* da digitare al momento della connessione al sistema (*login*).

Infine, è necessario associare a ogni tipologia di documento un profilo di accessibilità, cioè un insieme di informazioni che, combinate con i dati contenuti nel profilo di autorizzazione degli utenti, consentono di applicare le regole per l'esercizio del diritto di accesso. Nel caso di documenti che contengono dati personali, la *policy* dell'accessibilità deve tenere conto delle determinazioni assunte in merito al titolare, al responsabile e agli incaricati del trattamento[11], restringendo le abilitazioni previste per la generalità dei documenti ricevuti o prodotti dall'organizzazione. Per certi tipi di documento, inoltre, potrebbe essere necessario differire temporaneamente l'accesso per consentire una valutazione preliminare da parte dell'alta direzione, oppure limitarlo ai soggetti appartenenti agli organismi intersettoriali sopra citati.

Mantenere aggiornato un insieme di dati così complesso e articolato non è affatto semplice; una persona, infatti, può essere trasferita da un'unità organizzativa a un'altra, delegata temporaneamente a svolgere ruoli e assumere responsabilità diverse da quelle originarie, entrare a far parte di gruppi di lavoro, team di progetto, commissioni che operano al di fuori delle gerarchie proprie della struttura organizzativa. Tuttavia, per consentire l'accesso ai documenti informatici ai soli soggetti che ne hanno diritto, non si può prescindere da una gestione efficiente dei profili di autorizzazione degli utenti e di accessibilità dei documenti.

---

[11] Per titolare del trattamento si intende l'entità nel suo complesso, l'unità od organismo periferico che esercita un potere decisionale del tutto autonomo sulle finalità e sulle modalità del trattamento. Il responsabile del trattamento è il soggetto, dotato di esperienza, capacità ed affidabilità, a cui il titolare affida il compito di effettuare, appunto, il trattamento, attenendosi alle istruzioni che dovranno essere impartite per iscritto. Gli incaricati del trattamento sono le persone fisiche chiamate ad eseguire tali operazioni sotto la diretta autorità del titolare e del responsabile.

Sotto l'aspetto tecnologico, riguardo alle esigenze in tema di accessibilità e riservatezza dei documenti, un sistema ERMS deve presentare funzionalità tali da garantire:

- l'univoca identificazione ed autenticazione degli utenti;
- la garanzia di accesso alle risorse esclusivamente agli utenti abilitati;
- la registrazione, in forma protetta e immodificabile, delle attività rilevanti ai fini della sicurezza svolte da ciascun utente;
- il tracciamento, in forma protetta e immodificabile, di qualsiasi evento di modifica delle informazioni trattate e l'individuazione del suo autore;
- la protezione del patrimonio informativo e documentario da alterazioni, cancellazioni o distruzioni più o meno volontarie.

3. Sistemi di conservazione digitale

I documenti sono prodotti da un'organizzazione in modo naturale, come esigenza pratica, operativa o giuridica, e sono archiviati per avere memoria delle attività svolte o far valere i propri diritti. I documenti dell'archivio, quindi, devono essere conservati almeno finchè questi producono effetti giuridici o sono necessari per lo svolgimento di attività pratiche-amministrative, oppure hanno rilevanza sotto il profilo culturale e storico. Il tempo minimo obbligatorio di conservazione dei documenti di un archivio deve essere stabilito dai soggetti produttori sulla base del contesto normativo, procedurale e organizzativo in cui essi operano, e riportato nel piano di conservazione dell'archivio[12] che

[12] Il piano di conservazione dell'archivio è il piano contenente i criteri di selezione periodica e conservazione permanente dei documenti, nel rispetto delle vigenti disposizioni in materia di tutela dei beni culturali.

rappresenta la guida a cui fanno riferimento gli archivisti per individuare i documenti da destinare alla distruzione perché ritenuti *inutili*[13].

Durante il periodo di conservazione, i documenti dell'archivio devono essere facilmente accessibili, intellegibili, stabili – nel senso che non devono modificarsi nel contenuto e nella forma – e mantenere il valore giuridico originario. Se queste condizioni appaiono ovvie per i documenti cartacei e possono essere soddisfatte utilizzando locali adeguatamente attrezzati e un buon servizio archivistico, altrettanto non si può dire per i documenti informatici sui quali agisce inesorabilmente l'obsolescenza tecnologica.

3.1 *Obsolescenza tecnologica*

L'archiviazione dei documenti informatici su un sistema ERMS non ne garantisce la loro conservazione nel tempo perché in seguito, per effetto dell'obsolescenza tecnologica, i componenti hardware e software di questo sistema possono deteriorarsi a tal punto da impedire la corretta ricostruzione delle sequenze binarie, e quindi la rappresentazione dei documenti a livello utente.

I supporti di memorizzazione hanno un tempo di vita che dipende dai materiali utilizzati per la loro costruzione (qualsiasi materiale in natura è soggetto a deteriorarsi in tempi più o meno lunghi), dalla tecnologia impiegata per la registrazione dei bit (magnetica, ottica, magnetico-ottica, nastro, etc.) e dagli accorgimenti adottati per la loro tenuta (involucro esterno di protezio-

[13] Ai sensi dell'art.68, comma 1 del Decreto del Presidente della Repubblica 28 dicembre 2000, n. 445, *Disposizioni legislative in materia di documentazione amministrativa*, in Gazzetta Ufficiale del 20 febbraio 2001, n. 42, Supplemento Ordinario n. 30, le pubbliche amministrazioni hanno l'obbligo di predisporre un piano di conservazione dell'archivio integrato con il titolario di classificazione dei documenti.

ne, conservazione in locali deumidificati e a temperatura ambiente controllata, protezione da interferenze elettriche o elettromagnetiche, etc.). Occorre, quindi, monitorare lostato di efficienza dei supporti ed eseguire processi di riversamento su nuovi *media* quando si registrano i primi segni di obsolescenza tecnologica. Per quest'attività di monitoraggio sono necessari sistemi di *storage management* progettati per garantire la massima sicurezza delle sequenze binarie e che presentano: un'aspettativa di vita molto lunga; funzionalità avanzate per l'esecuzione di processi di migrazione; un elevato grado di standardizzazione di tutti i suoi componenti; meccanismi per la rilevazione automatica degli errori in lettura e scrittura dei dati; funzionalità avanzate per l'esecuzione delle operazioni di *backup* e *restore*.

La perfetta ricostruzione delle sequenze binarie dei documenti informatici memorizzati in un sistema di *storage management* è una condizione necessaria, ma non sufficiente per la loro rappresentazione a livello utente, per la quale, infatti, gioca un ruolo fondamentale il formato elettronico, cioè l'insieme delle regole da applicare per interpretare i bit contenuti nei file e rappresentare i documenti a video, a stampa o su altri dispositivi di output del computer. L'evoluzione tecnologica da un lato e l'interesse del mercato ad applicazioni software per il trattamento di oggetti multimediali complessi, contenenti sia i dati sia le istruzioni per la loro gestione, lasciano intravedere lo sviluppo nel breve periodo di nuovi formati elettronici e il contestuale abbandono di quelli meno diffusi e più *vecchi*. Pertanto, per conservare i documenti informatici a lungo termine, mantenendo la capacità di rappresentarli con il contenuto e la forma originari, è necessario monitorare l'evoluzione dei formati elettronici e provvedere alla migrazione dei file prima che i relativi formati siano abbandonati. Ad ogni modo, i rischi dell'obsolescenza tecnologica dei formati elettronici possono essere ridotti utilizzando formati aperti, riconosciuti come standard da organismi internazionali, indipendenti dalle piattaforme tecnologiche utilizzate per la

loro rappresentazione, usabili e capaci di includere al loro interno i metadati descrittivi del contenuto digitale.

3.2 *Sicurezza informatica*

In un processo di conservazione digitale, accanto ai rischi riconducibili all'obsolescenza tecnologica dei sistemi hardware e software ci sono altri pericoli derivanti da attacchi informatici o dalla perdita, accidentale o volontaria, del patrimonio documentario. Non bisogna dimenticare, infatti, che qualcuno potrebbe avere interesse a cancellare fisicamente alcuni documenti conservati, in particolare quelli da cui nascono rilevanti responsabilità, distruggendo il sistema dove è conservato l'archivio digitale oppure immettendovi un virus informatico che si propaga progressivamente in tutte le copie di backup.

Per contrastare questi pericoli si deve utilizzare un impianto tecnologico posto in un ambiente dotato di:

- sistemi di videosorveglianza, antintrusione e controllo degli accessi fisici;
- sistemi UPS che garantiscono il funzionamento delle apparecchiature anche in assenza di alimentazione elettrica per il tempo necessario al ripristino dell'impianto;
- sistema di monitoraggio ambientale con sensori per il rilevamento di fumi, temperatura, umidità e acqua;
- impianto antincendio;
- impianto di condizionamento ridondato;
- collegamento con le organizzazioni che assicurano il pronto intervento nei casi di effrazione, incendio, allagamento dei locali e altre condizioni di crisi ambientale.

Inoltre, è richiesto il massimo livello di sicurezza informatica per proteggere il patrimonio documentario digitale da attacchi informatici,virus o accessi online non autorizzati.

Riguardo alla sicurezza informatica e alla continuità operativa dei sistemi si è pronunciato anche il legislatore italiano il qua-

le, nell'art. 50-bis del Codice dell'Amministrazione Digitale, ha obbligato le pubbliche amministrazioni a predisporre un piano di continuità operativa, che descrive le procedure e le misure adottate per garantire il funzionamento dei sistemi 24 ore su 24, e un piano di *disaster recovery*, che specifica le misure tecniche e organizzative implementate per assicurare il ripristino dell'operatività dei sistemi e il salvataggio dei documenti anche in caso di eventi disastrosi o azioni dolose.

3.3 *Modello concettuale e funzionale*

L'esigenza di contrastare efficacemente i rischi derivanti dall'obsolescenza tecnologica e di garantire allo stesso tempo il massimo livello di sicurezza informatica spinge ad eseguire il processo di conservazione digitale in un ambiente tecnologico affidabile, protetto e controllato[14]. E proprio queste caratteristiche rendono l'ambiente di conservazione diverso da quello di produzione e gestione il quale, invece, è disegnato per favorire la più ampia condivisione in rete del patrimonio documentario, l'interoperabilità e la cooperazione applicativa tra i sistemi. Questo spiega perché nello standard ISO 14721:2003[15], relativo a un modello di sistema informativo aperto per l'archiviazione (mo-

[14] Ai sensi dell'art. 51, comma 2, del Codice dell'Amministrazione Digitale, *«i documenti informatici delle pubbliche amministrazioni devono esser custoditi e controllati con modalità tali da ridurre al minimo i rischi di distruzione, perdita, accesso non autorizzato o non consentito o non conforme alle finalità della raccolta»*. L'art. 44 della stessa norma specifica inoltre che *«il sistema di conservazione deve garantire: l'integrità del documento; la leggibilità e l'agevole reperibilità dei documenti e delle informazioni identificative, inclusi i dati di registrazione e classificazione; il rispetto delle misure di sicurezza previste dagli articoli da 31 a 36 del d. lgs. n. 196/2003, recante il codice in materia di protezione dei dati personali»*.

[15] ISO 14721:2003, Space data and information transfer systems - *Open archival information system - Reference model.*

dello OAIS: *Reference Model for an Open Archival Information System)*[16], è previsto il versamento periodico delle unità documentarie e archivistiche, con i relativi metadati, dal soggetto produttore dell'archivio, cioè dall'ambiente di produzione, gestione e archiviazione alla struttura di conservazione.

Anche il legislatore italiano, nella Deliberazione CNIPA n. 11/2004, contenente le regole tecniche per la riproduzione e la conservazione di documenti su supporto ottico idoneo a garantire la conformità dei documenti agli originali[17], ha disegnato un processo conservativo che prevede:

a) l'estrazione dei documenti informatici dal sistema di produzione, gestione e archiviazione;
b) la loro memorizzazione su supporti ottici, o su altri supporti *«comunque idonei a garantire la conformità dei documenti agli originali»* (art. 8);
c) la sottoscrizione da parte del Responsabile delle conservazione, con la propria firma digitale e l'attribuzione di un riferimento temporale opponibile a terzi, di un file denominato *Indice di conservazione* contenente i dati identificativi e le impronte digitali dei documenti conservati[18].

Il modello OAIS prevede esplicitamente la conservazione dei documenti informatici insieme ai metadati che li identificano, li qualificano sotto il profilo dell'integrità e li collocano nel contesto di provenienza, soddisfando così l'esigenza di rendere evidenti i *vincoli archivistici*, cioè l'insieme delle relazioni logiche

16 Il documento progettuale del modello OAIS, nella forma di raccomandazione CCSDS (*Consultative Committee for Space Data System*), è disponibile all'indirizzo <http://public.ccsds.org/publications/refmodel.aspx>.

17 Si segnala che la Deliberazione CNIPA n. 11/2004 è in via di sostituzione con nuove regole tecniche che sono state predisposte dagli organi competenti, ma non ancora pubblicate in Gazzetta Ufficiale.

18 Per impronta digitale s'intende una sequenza binaria di 256 bit generata applicando al documento informatico la funzione crittografica di HASH.

e formali che esistono necessariamente tra i documenti di un archivio. Tutte le operazioni descritte in questo modello, infatti, prevedono la formazione o la modifica di *pacchetti informativi* composti di quattro elementi:

- il contenuto informativo (*content information*) con l'insieme delle informazioni che ne permettono la rappresentazione e la comprensione a livello utente (*representation information*);
- le informazioni sulla conservazione (*preservation information*), che comprendono: le informazioni d'identificazione (*reference information*), le informazioni sul contesto (*context information*), le informazioni sulla provenienza (*provenance information*) e le informazioni sull'integrità (*fixed information*);
- le informazioni sull'impacchettamento (*packaging information*), che sono i dati che indirizzano alla posizione fisica del pacchetto informativo nel sistema di *storage management*;
- le informazioni descrittive del pacchetto (*package description*), rappresentate dai dati necessari per la ricerca e l'acquisizione del pacchetto informativo nel sistema di conservazione.

Sotto il profilo pratico, un processo conservativo digitale conforme allo standard ISO 14721:2003 (modello OAIS) si articola in cinque fasi:

1. Fase preparatoria. Si tratta di una fase preliminare nella quale il soggetto produttore dei contenuti digitali e la struttura di conservazione devono concordare tutti gli aspetti tecnici, procedurali, archivistici e giuridici connessi all'erogazione del servizio di conservazione;
2. Fase della formazione della memoria digitale. Il soggetto produttore deve garantire la formazione dell'archivio digitale sul proprio sistema ERMS, registrando su di esso le

unità documentarie e archivistiche unitamente al set di metadati concordato nella fase preparatoria;

3. Fase del versamento. Il versamento delle unità documentarie e archivistiche digitali, dal soggetto produttore alla struttura di conservazione deve avvenire attraverso la formazione e lo scambio per via telematica di un Pacchetto Informativo di Versamento (SIP), con le modalità e i tempi concordati nella fase preparatoria. Si sottilinea l'esigenza di avviare il processo conservativo prima che i contenuti digitali possano essere corrotti, o perdano la forza probatoria originaria, e di garantire la conservazione anche delle unità archivistiche che si completano molto tempo dopo l'avvio delle relative attività;
4. Fase della conservazione. Qui sono ricomprese tutte le attività che riguardano la verifica dei pacchetti informativi di versamento trasmessi dai soggetti produttori, il rilascio di una ricevuta di presa in carico, la formazione dei Pacchetti Informativi di Archiviazione (AIP) e l'esecuzione dei processi di migrazione per l'aggiornamento tecnologico. A questo livello, il legislatore italiano, nella Deliberazione CNIPA n. 11/2004, ha previsto il monitoraggio continuo della leggibilità dei documenti informatici memorizzati nell'ambiente di conservazione e l'esecuzione, quando necessario, dei processi di *riversamento diretto* per contrastare l'obsolescenza tecnologica dei supporti di memorizzazione e del sistema di *storage management* e di *riversamento sostitutivo* per fronteggiare l'obsolescenza dei formati elettronici. Ai sensi della normativa vigente in Italia, il riversamento sostitutivo consiste nel trasferire i documenti conservati da un supporto di memorizzazione a un altro, modificando la loro rappresentazione informatica (conversione di formato) e apponendo, su un'evidenza informatica contenente le impronte digitali dei documenti trattati, il riferimento temporale e la firma digitale del Re-

sponsabile della conservazione. Inoltre, se tale processo è applicato a documenti informatici sottoscritti digitalmente, o a quelli sostitutivi di documenti analogici originali unici, è richiesta anche l'apposizione del riferimento temporale e della firma digitale di un pubblico ufficiale;

5. Fase dell'accesso e fruizione. La funzione del conservatore digitale non è soltanto quella di mantenere inalterate nel tempo le sequenze binarie dei contenuti digitali, ma anche di permetterne l'accesso e la fruizione per esigenze pratiche o amministrative, oltre che per finalità di studio. Il sistema di conservazione digitale, pertanto, deve essere dotato di funzionalità avanzate per la ricerca, l'acquisizione e la riproduzione del materiale documentario conservato.

3.4 *Figure professionali e Responsabile della conservazione*

Non è difficile convincersi che lo svolgimento delle attività che realizzano un processo conservativo digitale conforme al modello OAIS richiede, oltre alla disponibilità di un impianto tecnologico affidabile e posto in sicurezza fisica e logica, anche l'impiego di personale qualificato, con competenze in materia di:

- archivistica, perché l'obiettivo è conservare archivi digitali permettendone l'accesso e la fruizione nel tempo;
- informatica, perché per conservare il digitale occorre conoscere i punti di forza e di debolezza delle tecnologie;
- diritto e diplomatica del documento contemporaneo, perché è necessario conservare i documenti informatici con il valore giuridico e la forza probatoria originari;
- organizzazione, perché è richiesta la sinergia con i soggetti produttori al fine di garantire la formazione – presso di loro – di archivi digitali realmente conservabili.

Spicca in modo particolare la figura del Responsabile della conservazione che deve possedere competenze e capacità adeguate per sovrintendere a tutto il processo conservativo digitale,

archiviando la documentazione necessaria per attestare, in sede legale, l'autenticità degli oggetti digitali esibiti[19].

Queste figure di archivisti informatici sono indispensabili, come rileva anche il legislatore italiano quando ai conservatori digitali che, sensi dell'art. 44-bis, comma 1, del Codice dell'Amministrazione Digitale, chiedono di essere accreditati come fornitori in possesso dei requisiti di livello più elevato in termini di qualità e sicurezza, impone il vincolo di *«utilizzare personale dotato delle conoscenze specifiche, dell'esperienza e delle competenze necessarie per i servizi forniti, in particolare della competenza a livello gestionale, della conoscenza specifica nel settore della gestione documentale e conservazione di documenti informatici»* e di dimostrare il possesso di questo requisito fornendo il curriculum vitae delle risorse impiegate nel quale deve essere *«attestato, mediante il possesso di specifici percorsi di studio universitari, ovvero mediante congrui periodi di specifica attività in contesti specialistici, il possesso di conoscenze peculiari e documentate nel campo della gestione documentale, dell'informatica applicata alla gestione dei documenti, dei metodi e sistemi di classificazione dei documenti digitali»*[20].

[19] La figura del Responsabile della conservazione è prevista dall'art. 5 della Deliberazione CNIPA n. 11/2004, nel quale sono specificati anche i suoi compiti: definire le caratteristiche e i requisiti del sistema di conservazione in funzione della tipologia di documenti da conservare; organizzare il contenuto dei supporti ottici e gestire le procedure di sicurezza e di tracciabilità che ne garantiscono la corretta conservazione, anche per consentire l'esibizione di ciascun documento conservato; garantire la corretta e puntuale esecuzione delle procedure di conservazione di cui alla normativa vigente; verificare periodicamente l'effettiva leggibilità dei documenti conservati ed eseguire, se necessario, le operazioni di riversamento diretto o sostitutivo; stabilire le necessarie misure di sicurezza logica e fisica del sistema; definire e documentare quanto previsto dalla normativa vigente per il riferimento temporale.

[20] Si veda la Circolare DigitPA 29 dicembre 2011, n. 59, *Modalità per pre-*

Se da un lato si registra una significativa richiesta di figure professionali qualificate da impiegare per la conservazione a lungo termine di documenti informatici e archivi digitali o per la gestione informatica dei documenti (*record management*), dall'altro, purtroppo, si rileva un'offerta limitata di percorsi formativi adeguati, che ad oggi sono rappresentati quasi esclusivamente da Master universitari post lauream[21]. È senza dubbio auspicabile che le università italiane provvedano rapidamente ad aggiornare la loro offerta formativa, attivando percorsi di studio che permettano di acquisire una specializzazione nel settore della formazione, gestione e conservazione di archivi digitali.

Bibliografia

BONFIGLIO-DOSIO, G., *La formazione del fascicolo archivistico in ambiente digitale*, in Una mente colorata. Studi in onore di Attilio Mauro Caproni per i suoi 65 anni, Roma, Vecchiarelli, 2007, pp. 549-553.

Circolare DigitPA 29 dicembre 2011, n. 59, *Modalità per presentare la domanda di accreditamento da parte dei soggetti pubblici e privati che svolgono attività di conservazione dei documenti informatici di cui all'articolo 44-bis, comma 1, del decreto legislativo 7 marzo 2005, n. 82.*

Decreto del Presidente della Repubblica 11 febbraio 2005, n. 68, *Regolamento recante disposizioni per l'utilizzo della posta elettronica certificata, a norma dell'articolo 27 della legge 16 gennaio 2003, n. 3,* in Gazzetta Ufficiale del 28 aprile 2005, n. 97.

Decreto del Presidente della Repubblica 28 dicembre 2000, n. 445, *Disposizioni legislative in materia di documentazione amministrativa*, in Gazzetta Ufficiale del 20 febbraio 2001, n. 42, Supplemento Ordinario n. 30.

sentare la domanda di accreditamento da parte dei soggetti pubblici e privati che svolgono attività di conservazione dei documenti informatici di cui all'art. 44-bis, comma 1, del decreto legislativo 7 marzo 2005, n. 82.

[21] Un esempio è rappresentato dal Master universitario di I° livello in *Formazione, gestione e conservazione di archivi digitali in ambito pubblico e privato* attivato dall'Università degli Studi di Macerata e ormai giunto alla VI edizione. <http://www.masterarchividigitali.it>.

Decreto legislativo 30 giugno 2003, n. 196, *Codice in materia di protezione dei dati personali*, in Gazzetta Ufficiale del 29 luglio 2003, n. 174, Supplemento Ordinario n. 123.

Decreto Legislativo 7 marzo 2005, n. 82, *Codice dell'Amministrazione Digitale*, in Gazzetta Ufficiale del 16 maggio 2005, n. 112, Supplemento Ordinario n. 93.

Deliberazione CNIPA n. 11/2004 del 19 febbraio 2004, *Regole tecniche per la riproduzione e conservazione di documenti su supporto ottico idoneo a garantire la conformità dei documenti agli originali.*

Direttiva 1999/93/CE del Parlamento europeo e del Consiglio del 13 dicembre 1999 Relativa ad un quadro comunitario per le firme elettroniche, in Gazzetta Ufficiale delle Comunità europee del 13 dicembre 1999.

DLM FORUM FOUNDATION, *MoReq2010®: Modular Requirements for RecordsSystems*, Vol. 1: Core Services & Plug-in Modules, 2011. <http://moreq2010.eu>

ISO 14721:2003, Space data and information transfer systems – *Open archival information system – Reference model.*

ISO 15489-1: 2001, Information and documentation – *Records Management – Part 1: General.*

PIGLIAPOCO, S., *Il fascicolo elettronico*, in Il fascicolo elettronico, Pigliapoco S. (a cura di), Padova, SIAV Academy**,** 2010, pp. 33-52.

Sitografia

<http://public.ccsds.org/publications/refmodel.aspx>
<http://www.masterarchividigitali.it>

I linguaggi di descrizione documentale

EDUARDO DE FRANCESCO*

Introduzione

Esistono diversi formalismi per rappresentare informaticamente i documenti. I principali sono:

- I sistemi di codifica dei caratteri;
- I sistemi di marcatura dei testi.

I caratteri sono il livello elementare della rappresentazione di un testo su supporto digitale, infatti ogni documento è costituito da un insieme di caratteri (stringa) che può descrivere sia una sequenza di elementi lessicali, sia una sequenza di elementi di controllo (es. ritorno carrello, tabulazioni, evidenziazioni, ecc.)

Nel corso degli anni sono stati codificati molti sistemi per gestire le sequenze di controllo: alcuni di essi si limitavano a dialogare con il computer per fornire informazioni di tale natura, altri si sono specializzati per potere interpretare il significato strutturale dei testi. Questi ultimi hanno preso il nome di linguaggi di marcatura o di tagging e si sono specializzati in due categorie:

- Linguaggi procedurali indicati anche come *specific markup language;*
- Linguaggi dichiarativi detti anche *generic markup language.*

* Settori Ricerca&Sviluppo ed Information&Communication Technology del Gruppo Se.Te.L.

Alla prima categoria fanno capo linguaggi come Script, TROFF, TEX o RTF (*Rich Text Format*). Essi permettono di indicare la struttura visuale (talvolta detta topografica) del testo e riportano informazioni relative al corpo del carattere, alla sua spaziatura, alla sua posizione rispetto alla riga, alla sua dimensione, al tipo di font utilizzato, alla strutturazione delle pagine, ecc.

Della seconda categoria fanno parte, invece, linguaggi come SGML (*Standard Generalized Markup Language),* XML (*eXtensible Markup Language*) e parzialmente HTML (*HyperText Markup Language*); essi permettono di descrivere la struttura astratta del documento, dichiarando gli elementi che lo costituiscono (ad esempio: sezione, capitolo, paragrafo, nota, enfasi, ecc.).

Nel seguito, verranno approfonditi questi concetti, sia attraverso la loro storia, sia attraverso l'indicazione delle tendenze future.

L'evoluzione dei sistemi di codifica dei caratteri

I caratteri vengono rappresentati mediante una codifica numerica binaria che stabilisce un'associazione biunivoca tra gli elementi di una collezione di simboli (*character repertoire*) e un insieme di codici numerici (*code set*). L'insieme risultante viene denominato tecnicamente *coded character set*. Per ciascun set, poi, si definisce una codifica dei caratteri (*character encoding*) basata su un algoritmo che mappa una o più sequenze di bit al numero intero che rappresenta un dato carattere in un set.

Alcuni set sono stati definiti da enti di standardizzazione nazionali e internazionali (ISO – *International Organization for Standardization* – e ANSI – *American National Standards Institute* – in primo luogo) e si differenziano per il numero di cifre binarie che utilizzano e dunque per il numero di caratteri che possono codificare. Tra questi il più diffuso è il cosiddetto codice ASCII (*American Standard Code for Information Interchange*),

la cui versione internazionale corrisponde alla ISO 646 IRV[1].

Nato negli Stati Uniti, utilizza solo 7 bit; il set è composto da 128 caratteri, tra cui i simboli dell'alfabeto anglosassone e alcuni segni di punteggiatura.

Può sembrare strano l'uso di soli 7 bit (ormai tutti ragionano in byte, ossia su base 8) ma l'inglese americano non aveva (e non ha) la necessità di gestire caratteri particolari o accentati come negli alfabeti europei e a quel tempo ogni bit era prezioso[2].

La diffusione dei computer ha naturalmente determinato l'esigenza di rappresentare i caratteri di altri alfabeti. Sono stati così sviluppati molteplici *code set* che utilizzano 8 bit (un intero ottetto e quindi 256 posizioni) e che hanno di volta in volta accolto i simboli dei vari alfabeti latini. Tra di essi ricordiamo la famiglia ISO 8859, nel cui ambito è particolarmente diffuso il *code set* ISO 8859-1[3], meglio conosciuto come ISO Latin 1. Esso contiene i caratteri principali delle lingue occidentali con alfabeti latini ed è usato da molte applicazioni su Internet (ad esempio World Wide Web) e da molti sistemi operativi.

Tra gli altri set è da notare l'EBCDIC (*Extended Binary Coded Decimal Interchange Code*): esso indica un sistema di codifica dell'informazione a 8 bit usato in numerosi sistemi operativi di produzione IBM sia per elaboratori di classe mainframe (ad es. z/OS, OS/390, VM e VSE) che per minicomputer[4].

---

1 ISO/IEC 646:1991, *Information technology — ISO 7-bit coded character set for information interchange*.

2 Pochi sanno che a tutt'oggi il WEB usa ancora la codifica a 7 bit, con significativi problemi di conversione nei confronti di set di caratteri particolari.

3 ISO/IEC 8859-1:1998, *Information technology — 8-bit single-byte coded graphic character sets — Part 1: Latin alphabet No. 1*.

4 Nel seguito di questo capitolo è stato inserito un piccolo approfondimento su questo code set poiché utilizzato in modo significativo da IBM, in associazione al codice di marcatura GML, negli anni 70 ed 80.

Il più completo ed evoluto standard per la codifica di caratteri attualmente disponibile è l'ISO 10646-1, rilasciato nel 1993 e aggiornato nel 2000[5]. Esso definisce lo *Universal Character Set*, un *coded character set* basato su una codifica a 31 bit e coincide praticamente con l'omologo set a 16 bit (65.536 combinazioni) Unicode, sviluppato autonomamente da una organizzazione privata, lo Unicode Consortium.

La Figura 1 illustra quanto riportato sul sito ufficiale dello standard[6]:

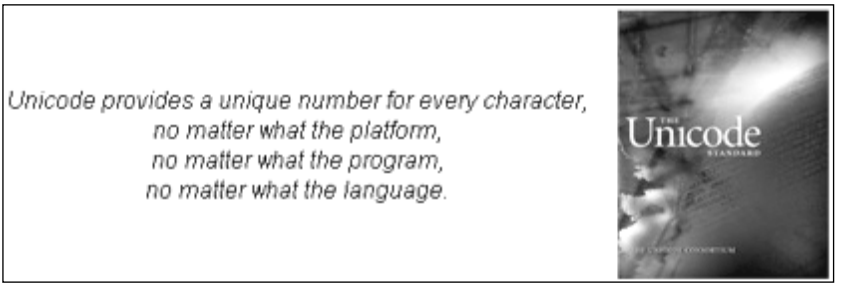


Figura 1. Unicode.

Lo standard Unicode è stato adottato da leader di mercato del calibro di Apple, HP, IBM, JustSystem, Microsoft, Oracle, SAP, Sun, Sybase, Unisys e molti altri. È alla base di molti moderni standard, come XML, Java, ECMAScript (JavaScript), LDAP (*Lightweight Directory Access Protocol)*, CORBA (*Common Object Request Broker Architecture*) 3.0, WML (*Vector Markup*

5 ISO/IEC 10646-1:2000, *Information technology -- Universal Multiple-Octet Coded Character Set (UCS) -- Part 1: Architecture and Basic Multilingual Plane*.

6 <http://www.unicode.org/standard/WhatIsUnicode.html>.
Unicode assegna un numero univoco a ogni carattere, indipendentemente dalla piattaforma, indipendentemente dall'applicazione, indipendentemente dalla lingua.

Language) ecc. e costituisce l'implementazione ufficiale dello standard internazionale ISO/IEC 10646. L'emergere dello standard Unicode, unito alla recente disponibilità di strumenti che lo supportano, è fra i più significativi sviluppi della tecnologia della globalizzazione del software.

La sua adozione sui siti Web e nelle applicazioni client/server o *multi-tiered*[7], rispetto all'utilizzo dei set di caratteri tradizionali, permette un significativo abbattimento dei costi di gestione. Unicode consente che un'unica versione di un software o di un sito Web siano fruibili con piattaforme e lingue differenti e in paesi diversi, evitando la necessità di reingegnerizzare il prodotto per ogni situazione specifica. Permette, inoltre, il trasporto del testo fra sistemi differenti senza che abbia luogo alcuna corruzione dei dati.

Si riporta di seguito (Tabella 1) una possibile cronologia dell'evoluzione dei caratteri:

Storia dei *coded character set*
Ufficialmente, il primo fu il codice Morse, nato nel 1840, seguito dal linguaggio delle bandiere usato in marina.
L'introduzione delle telescriventi porta al codice Baudot del 1930, a 5 bit.
Nel 1963 nasce lo standard ASCII, a 7 bit, che è utilizzato per l'avvento di Internet e dei protocolli connessi. Nel 1965 viene approvato lo US-ASCII, che nel 1972 diventa lo standard ISO 646.
Tra il 1963 e il 1964 è stata presentata da IBM la codifica EBCDIC. Si tratta di una codifica a 8 bit in grado di estendere in modo rilevante il sistema BCD a 6 bit allora in uso; il suo sviluppo è stato indipendente rispetto alla codifica ASCII, che è a 7 bit.

[7] Un'applicazione Web che si sviluppa su più livelli logico-funzionali.

Nel 1981 le estensioni al codice ASCII per i caratteri dal 128 al 255 vengono identificate dai *codepage* PC-DOS e traslate poi per retrocompatibilità nello MS-DOS. Nel 1985 l'ISO approva gli standard *codepage* come ISO 8859-n, dove n è un numero che identifica il particolare *codepage*.
L'affermarsi di Windows, anche in Asia, porta alle estensioni alle lingue orientali nel 1990 dei *codepage* di Windows.
La comprensibile babele risultante dallo scambio di email e documenti tra paesi a *codepage* diverso fu oggetto di attenzione dell'ISO prima con lo standard del 1986 ISO 2022[8], in vigore ma scarsamente utilizzato, e poi con la proposta del 1991 dell'Unicode 1.0, la cui versione 1.1 del 1993 diviene lo standard ISO 10646, lo *Universal Character Set* (UCS).
Lo UCS-2, che usa due byte per ogni carattere, fu utilizzato dalla Microsoft in Windows NT sin dal 1995 e poi esteso a tutte le altre versioni.
Le definizioni dei formati UTF-8 e UTF-16 datano al 1996, con la versione 2.0 di Unicode. Lo UTF (*Unicode Transformation Format*) divenne lo standard POSIX de facto, ed essendo ratificato dalla RFC 3629, è anche riconosciuto dal W3C (World Wide Web Consortium). Esistono anche lo UTF-7 e l'UCS 4. Lo UTF-16 è un'estensione dello UCS 2. La successiva versione Unicode 3.0 del 1999 introduce la bidirezionalità e la composizione di tabelle, mentre la 4.0 del 2001 include anche le lingue antiche. La versione attualmente in uso è la 6.2 di settembre 2012.

Tabella 1. Storia dei *coded character set.*

L'evoluzione dei sistemi di marcatura

La funzionalità della rappresentazione digitale del testo dipende dalla qualità con cui l'informazione può essere modellizzata e dalle operazioni ad essa applicabili. L'informazione testuale viene rappresentata mediante sequenze lineari, o stringhe,

[8] ISO/IEC 2022:1994, *Information technology–Character code structure and extension techniques*.

di caratteri codificati in forma binaria. Il markup, ossia l'inserimento di marcatori (tag), permette di assegnare una struttura alla rappresentazione del testo distinguendo, nella sequenza dei caratteri codificati, parti diverse con funzioni diverse. Il modello assegnato all'informazione testuale dipende dal sistema di markup.

L'espressione *markup* deriva dall'analogia tra questi linguaggi e le annotazioni inserite da autori, curatori editoriali e correttori nei manoscritti e nella bozze di stampa di un testo, al fine di indicare correzioni e trattamenti editoriali, chiamati appunto in inglese *mark up*. In modo simile, i linguaggi di marcatura sono costituiti da un insieme di istruzioni dette, marcatori, che servono a descrivere la struttura e la formalizzazione mediale del documento[9].

I marcatori sono sequenze di normali caratteri ASCII e vengono introdotti, secondo una determinata sintassi, all'interno del documento, accanto alla porzione di testo cui si riferiscono.

Attraverso i vari linguaggi di marcatura si possono rappresentare o il modello strutturale dei contenuti del documento o il modello di rappresentazione mediale.

Il testimone più illustre dei linguaggi di markup è SGML. Ideato da Charles Goldfarb, esso è divenuto lo standard ufficiale adottato dall'ISO per la creazione e l'interscambio di documenti elettronici. La pubblicazione dello standard risale al 1986, è stato utilizzato per lunghi anni ed è poi evoluto verso altri linguaggi di marcatura (HTML, XML ed altri) approfonditi nel seguito di questo documento.

Quanto segue è la storia dell'SGML (e quindi di fatto dei linguaggi di marcatura) così come raccontata dal suo fondatore e

[9] Un documento si può infatti esprimere non solo in forma visuale (ossia tramite il media video o la carta) ma anche sotto forma di altri media (parlato, sensazioni tattili, ecc.).

nume tutelare Charles Goldfarb sulle pagine dell'SGML User Group nel giugno del 1990[10].

Storicamente i manoscritti elettronici contenevano dei codici di controllo o delle macro che permettevano a un documento di essere formattato in un modo particolare (*Specific coding*).

Al contrario, le metodologie di codifica generalizzata, che iniziano nel tardo 1960, usano marcature descrittive (ad esempio la marcatura *titolo* al posto di un *formato-17*). Molta della spinta verso la codifica generica è dovuta ad una presentazione effettuata da William Tunnicliffe, *chairman* della *Graphic Communications Association* (GCA), durante un meeting presso il *Canadian Government Printing Office* nel settembre del 1967. In questa occasione viene infatti affermato per la prima volta uno dei concetti fondamentali dei linguaggi di marcatura moderni, ossia la separazione completa fra il concetto di contenuto informativo del documento e la sua formalizzazione mediale.

Sempre durante l'ultimo periodo del 1960 Stanley Rice (progettista di libri a New York) propose l'idea di un catalogo universale di elementi di marcatura parametrici (tag). L'idea fu raccolta da Norman Scharpf, *Director* della GCA, che propose un *commettee* sull'argomento. Il *commettee* concluse che era necessario utilizzare diversi sistemi di marcatura per differenti tipi di documenti e che documenti più piccoli potevano essere incorporati come elementi di documenti più grandi.

Il progetto evolse nel *GenCode Committee* che fu di fatto la base operativa per la creazione dello standard SGML.

Nel 1969 Charles Goldfarb dirigeva un progetto di ricerca IBM su un sistema informativo per la gestione delle leggi. In quel contesto inventò il *Generalized Markup Language* (GML) come metodo per permettere la scrittura, la formattazione, il recupero e la condivisione di documenti. Questo linguaggio era ba-

10 <http://www.sgmlsource.com/history/sgmlhist.htm>.

sato sull'idea originale di Rice e Tunnicliffe, ma introduceva anche il concetto di strutture ricorsive esplicite all'interno del documento. Esso era utilizzato dai mainframe IBM e raggiunse una larga diffusione, tanto che la stessa IBM diventò in quegli anni il secondo *publisher* a livello mondiale.

Nel 1978, l'ANSI decise la nascita di un comitato per lo studio di *Computer Languages for the Processing of Text*, capeggiato da Charles Card.

Goldfarb fu invitato a partecipare al progetto per favorire la nascita di uno standard basato su GML. La prima bozza vide la luce nel 1980 e si concluse con la pubblicazione dello stesso nell'ottobre del 1985 come ISO 8879[11]. È da rimarcare come durante questa fase si instaurò una cooperazione con lo *European Particle Physics Laboratory* (CERN) di Ginevra dove di fatto verrà poi generato il linguaggio HTML.

Nei primi anni successivi alla nascita dello standard ci furono delle importanti evoluzioni dovute all'adozione dell'SGML da parte della *Association of American Publishers* (AAP), tramite il progetto *Electronic Manuscript Project*, e del Ministero della Difesa Americano, tramite l'iniziativa *Computer-aided Acquisition and Logistic Support* (CALS)[12]. Questi due eventi di fatto permisero la diffusione dello standard nel mondo degli editori e nel mondo della difesa/spazio.

Strettamente legato allo standard HTML, a cui si è accennato, è il concetto di hyperlink. Il termine fu coniato nel 1965 da Theodore Nelson all'inizio del progetto *Xanadu*. Nelson fu ispirato da una storia in cui si descriveva una macchina per microfilm; in questo sistema ogni fotogramma mostrava una pagina ed i vari fotogrammi erano rigidamente sequenziati dalle relazioni realiz-

11 ISO 8879:2005, *Standard Generalized Markup Language*.

12 All'epoca (1985) l'acronimo CALS significava *Computer Aided Logistic Support*, rinominato successivamente negli anni '90 come sopra citato per sottolineare il cambio di missione dell'iniziativa.

zate dal supporto fisico (la pellicola di celluloide). In questo contesto, l'innovazione portata dall'hyperlink era la possibilità di collegare due fotogrammi qualsiasi (anche non contigui) realizzando una sequenza di informazioni collegabili virtualmente e non più solo fisicamente.

In una sequenza di libri pubblicati tra il 1964 e il 1980, Nelson trasferì questi concetti in un contesto informatizzato rendendo applicabile l'idea di collegamento non più alle pagine intere ma a singoli elementi (stringhe) del testo e poco dopo a differenti paragrafi di documenti diversi. Di fatto era nato il concetto di ipertesto.

Per quanto riguarda i linguaggi di codifica procedurali, invece, uno dei più comuni è l'RTF (*formato di testo ricco*), realizzato dalla Microsoft che ne è proprietaria, al fine di agevolare lo scambio di documenti fra diverse applicazioni.

L'RTF è un tipo di codifica in base al quale il documento viene rappresentato come puro testo e alcuni *identificatori* (tag) specificano quale tipo di strutture di formattazione (allineamento, impaginazione, tipo di carattere) applicare al testo stesso. Per mezzo della marcatura, quindi, il testo del documento RTF (definito sfruttando soltanto set di caratteri molto semplici e diffusi, tipo l'ASCII) viene arricchito di nuove informazioni riguardanti la formattazione, di cui i formati di puro testo sono privi: da ciò il nome *formato di testo ricco*.

I tag RTF sono in genere definiti dal carattere *back slash* \ seguito da alcune lettere che specificano il tag stesso. Un esempio di testo marcato in RTF è:

```
{\b Questo testo è in grassetto, {\i e questo è anche in corsivo.}}
```

Microsoft ha a lungo sostenuto (fino circa al 1994) questo standard in alternativa all'SGML. L'RTF ha perso la sua battaglia contro l'SGML quando la diffusione del Web ha imposto di fatto i linguaggi di marcatura generica. È da notare comunque

che oggi tutti i *word processor* più diffusi hanno la possibilità di importare e di esportare dei file in RTF e che i manuali ipertestuali realizzati con la tecnologia dell'*help* di Windows (formato *.HLP) sono basati sul suddetto formato.

La standardizzazione del metodo per descrivere gli elementi costituenti il documento

Negli anni tra il 1960 e il 1985 si sviluppa quindi un movimento che porta alla definizione del primo standard di un linguaggio di marcatura, SGML.

Vediamo ora in concreto che cosa significa scrivere tramite questo tipo di linguaggio; nell'esempio che segue (Figura 2), utilizzando un documento a tutti noto quale la *Divina Commedia* di Dante, si è provato a rappresentarlo in termini strutturali e tramite elementi di marcatura, da qui in seguito chiamati, secondo l'uso internazionale, tag.

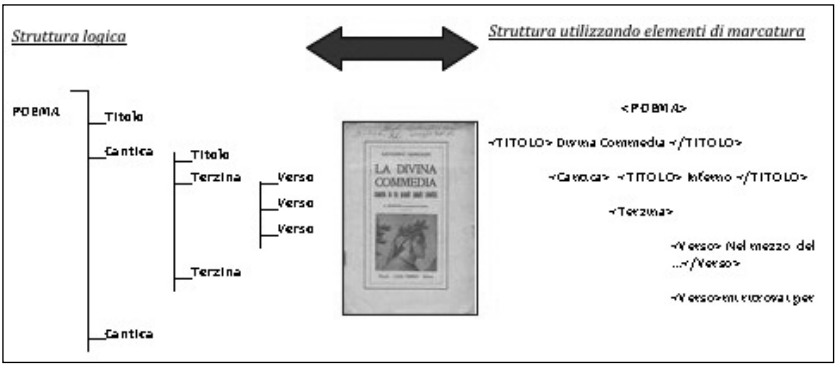


Figura 2. Esempio di marcatura di un testo.

La struttura logica riportata nella parte sinistra della figura rappresenta una delle possibili modellizzazioni con cui si può strutturare il documento in esame.

Nell'esempio che segue (Figura 3), tratto da una normativa

per la modellizzazione di documenti tecnici in ambito aeronautico (S1000D™ - *International Specification for Technical Publications Utilizing a Common Source Database*), si vede la strutturazione di un documento complesso attraverso la sua rappresentazione grafica.

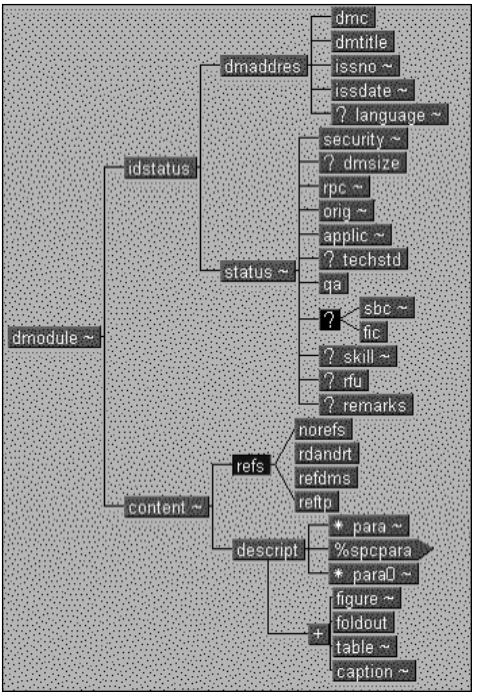


Figura 3. Esempio di strutturazione di un documento.

Ogni elemento riportato in figura, tipicamente composto da tag o da altri elementi, quali gli attributi dei tag, si descrive attraverso un linguaggio specifico di marcatura.

Il progettista documentale analizza il documento nelle sue componenti e ne definisce una struttura o modello, descrivendolo quindi attraverso un linguaggio formale specifico; nel nostro caso appunto l'SGML. Alcune caratteristiche:

- Ogni documento SGML deve contenere un solo elemento radice, cioè un elemento che racchiude tutti gli altri secondo una struttura di relazione padre-figlio, quindi strettamente gerarchica;
- L'elemento radice è il primo elemento di un documento SGML;
- Le relazioni gerarchiche esistenti fra i vari elementi danno vita al cosiddetto albero SGML.

Più che un linguaggio, SGML è un *metalinguaggio*: esso prescrive precise regole sintattiche per definire un insieme di marcatori e di relazioni tra marcatori, ma non dice nulla per quanto riguarda la loro tipologia, quantità e nomenclatura. Questa astrazione ne costituisce il nucleo e la potenza: in sostanza SGML serve non già a marcare direttamente documenti, ma a costruire, rispettando standard comuni e rigorosi, specifici linguaggi di marcatura adatti per esigenze particolari. Un linguaggio che rispetti le specifiche SGML viene definito *applicazione SGML* (*SGML application*).

Una applicazione SGML, a sua volta, descrive la struttura logica di una classe di documenti e non la loro forma fisica. Tale struttura astratta viene specificata dichiarando gli elementi che la costituiscono (ad esempio: *capitolo, titolo, paragrafo, nota, citazione*, ecc.) e le relazioni che tra questi intercorrono e che possono essere *gerarchiche* o *ordinali*.

Le dichiarazioni sono contenute in un file, denominato *Document Type Definition* (DTD)[13], che costituisce una sorta di grammatica dei documenti che ad essa si riferiscono e rispetto alla quale debbono essere convalidati. Più precisamente nella DTD

[13] Negli anni '90 un nuovo modello dichiarativo .denominato *Schema* si è affiancato alla *DTD*. Schema è associato allo standard XML ed entrambi I modelli sono attualmente in uso.

sono elencati e definiti tutti gli oggetti necessari all'elaborazione di un determinato tipo di documento:

- *Elementi*: i nomi dei marcatori adottati per identificare i blocchi strutturali costituenti l'albero dei dati;
- *Content model*: il modello di contenuto per ciascun elemento indica quali altri elementi, quando ed in che numero possono comparire al suo interno. Indica inoltre quali attributi possono essere associati ad un elemento. Gli attributi permettono l'aggiunta di informazioni addizionali, che possono essere utili per descrivere più accuratamente la struttura dei dati rappresentata. Essi possono essere paragonati a degli aggettivi;
- *Entità*: le entità permettono di creare dei riferimenti a dati esterni non necessariamente testuali, ad altri documenti, a file grafici e ad altri media. Molto usate sono le entità per rappresentare caratteri che non sono presenti nella tabella codici del documento o che si teme potrebbero essere rappresentati non correttamente nel passaggio da un sistema all'altro.

Riprendendo le parole di Yury Rubinsky, «*an SGML document should always be No Surprises*»[14], si può dire che un documento realizzato in SGML non presenta sorprese poiché:

- Deve iniziare con una *SGML declaration*. Questa dichiarazione formale standard permette al ricevente del documento di sapere esattamente quale *character set* si aspetta di ricevere e quali sono le regole specifiche di marcatura applicabili;

[14] *SoftQuad SGML Primer*, <http://www.lulu.com/shop/david-jones/the-xml-primer/paperback/product-282753.html>.
Yury Rubinsky è stato il fondatore di SoftQuad, nonché il creatore del primo editor commerciale SGML.

- Deve indicare (contenuto o riferito) qual è il modello (DTD) applicabile all'istanza documentale che segue;
- L'istanza documentale deve seguire le dichiarazioni suindicate, sapendo quale set di caratteri e quale set di regole sono applicabili.

Un'ultima annotazione può essere interessante: si è parlato di GML come precursore dell'SGML. Esiste una precisa correlazione tra il linguaggio GML e il *character set* EBCDIC, in quanto entrambi espressione del grande sviluppo commerciale dei mainframe. L'SGML invece, pur potendo accettare qualsiasi *character set*, è più legato alla codifica ASCII ed alle sue evoluzioni.

Le codifiche EBCDIC e ASCII non sono compatibili tra loro. Poiché i computer sono in grado di elaborare solo dati numerici, entrambe le convenzioni assegnano specifici caratteri a tali numeri, ma identici valori numerici sono interpretati come caratteri diversi in funzione della *codepage* impiegata. Per utilizzare dei dati memorizzati in EBCDIC è necessaria una conversione tra *codepage* per visualizzare correttamente le informazioni su elaboratori basati sulla codifica ASCII.

Nell'esempio di seguito riportato (Figura 4) si è analizzato il documento *ricette gastronomiche*; questa formalizzazione ed i relativi impatti a livello informatico saranno spesso usati come esempio nelle successive pagine.

Nella parte superiore viene presentata l'istanza del documento, nel caso specifico la ricetta, nella parte inferiore la sua modellizzazione formale (DTD).

Una volta definito un determinato linguaggio, esso può essere utilizzato per rappresentare infiniti documenti in base ad una sintassi rigorosa. A ciascun elemento corrisponde una coppia di marcatori. La sintassi standard prevede che i marcatori siano racchiusi tra i simboli di maggiore e minore. Ogni elemento viene identificato da un marcatore iniziale e uno finale (costruito pre-

```
<ricettario>
    <ricette>
        <ricetta idref="Alici Marinate">
                <ingredienti>
                    <ingrediente idref="Alici fresche" quantita="600 g"/>
                    <ingrediente idref="Aglio" quantita="2 spicchi"/>
                    <ingrediente idref="Prezzemolo" quantita="qb"/>
                    <ingrediente idref="Sale marino" quantita="qb"/>
                    <ingrediente idref="Origano" quantita="qb"/>
                    <ingrediente idref="Olio extra vergine" quantita="qb"/>
                    <ingrediente idref="Peperoncino rosso" quantita="qb"/>
                    <ingrediente idref="Limoni" quantita="2"/>
                    <ingrediente idref="Aceto Bianco" quantita="2 bicchieri"/>

                </ingredienti>

                <preparazione>
                    <fase idref="1" testo="Pulire le alici,togliendo .."/>
                    <fase idref="2" testo="Sopra ogni strato mettere sale.."/>
                    <fase idref="3" testo="Aggiungere a ogni strato il succo..."/>
                    <fase idref="4" testo="All'ultimo strato a tutti gli in...."/>
                    <fase idref="5" testo="Far riposare per mezz'ora e poi ...."/>
                    <fase idref="6" testo="Servire aggiungendo un filo d'olio.."/>
                </preparazione>
        </ricetta>
    </ricette>
</ricettario>



                        <!DOCTYPE Ricettario>
                        <!ELEMENT Ricettario (Glossario,Ricette)>
                                    <!ELEMENT Glossario ........>
                                    <!ELEMENT Ricette (Ricetta+)>
                                    <!ELEMENT Ricetta (Titolo, Ingredienti+, Preparazione)>
                                    <!ELEMENT Titolo (#PCDATA)
                                    <!ELEMENT Ingredienti EMPTY>
                                                <!ATTLIST Ingrediente IDREF #REQUIRED>
                                    <!ELEMENT Preparazione (Fasi+)>
                                    <!ELEMENT Fasi (Testo, Ingredienti)
                                    <!ELEMENT Testo (#PCDATA)>
```

Figura 4. Strutturazione documento *ricette gastronomiche.*

mettendo uno slash al nome del marcatore iniziale), a meno che non sia un elemento vuoto (nel qual caso è identificato solo dal marcatore iniziale seguito da uno slash).

I tag (SGML e suoi successori), oltre alla loro potenza espressiva, offrono una serie di vantaggi dal punto di vista del trattamento informatico dei documenti. Un documento marcato è composto esclusivamente da una sequenza di caratteri in un dato *code set* ed è quindi facilmente portabile su ogni tipo di computer e di sistema operativo. Lo stesso testo marcato può essere utilizzato per scopi differenti: stampa su carta o video, rappresentazione su media diversi (audio, tatto, ecc.), analisi tramite software specifici, elaborazione con database, correlazione automatica di corpora linguistici. Ciò anche in tempi diversi, senza dovere pagare i costi di complesse, costose e spesso inaffidabili conversioni tra formati spesso incompatibili.

Ma forse l'elemento primario che deve essere considerato quando si parla di documentazione rappresentata attraverso linguaggi di marcatura é il fatto che, nella sostanza, il documento

diventa un dato e quindi si presta allo sviluppo di applicazioni complesse. Diventa lecita l'interazione con i database, la creazione di strumenti di information retrieval contestuali, l'utilizzo per corsi interattivi, per l'insegnamento a distanza, ecc.

Particolare rilevanza assumono i linguaggi di marcatura per la produzione e la manutenzione di pubblicazioni articolate, come quelle relative alla documentazione tecnica, dove la sicurezza del dato è elemento spesso determinante per la sicurezza dell'uomo, o al settore legale, dove ogni singola parola ha spesso valenza specifica.

Per queste sue caratteristiche, SGML ha trovato impiego soprattutto in contesti industriali e militari, nei quali la gestione efficiente e sicura dei documenti tecnici ha una funzione critica. Ma non mancano applicazioni SGML in ambito scientifico o anche nel dominio umanistico, dove una applicazione SGML denominata *Text Encoding Iniziative* (TEI) è divenuta lo standard per la codifica e l'archiviazione dei testi su supporto digitale e per la creazione di biblioteche digitali[15] .

Senza dubbio l'applicazione SGML che gode della diffusione maggiore, anche se molti non la percepiscono come tale, è il linguaggio attuale del Web, ossia HTML.

L'evoluzione dei tag: dall'identificazione degli elementi costituenti la singola istanza al collegamento tra più istanze (hyperlink)

Il modello documentale legato all'SGML è relativo a una singola istanza del documento; ciò significa in pratica che é possibile modellizzare un libro, un documento tecnico, la descrizione di un farmaco, un ricettario, ma di per sé non è possibile collegare dei libri, dei documenti tecnici, delle descrizione di farmaci, dei ricettari. Ossia non esiste originariamente, all'interno del-

15 <http://www.tei-c.org/index.xml>.

la codifica di marcatura SGML, un metodo per il collegamento tra più istanze. Questa enorme limitazione che di fatto non permette la creazione di ipertesti fra più istanze, viene rilevata intorno alla fine degli anni ‘80 e trova una sua esplicitazione in una serie di studi che culminano con la creazione del linguaggio HTML, il cui padre fondatore viene considerato Tim Berners-Lee.

Egli era un ricercatore del CERN e, come si è accennato precedentemente, esiste una forte correlazione culturale tra i comitati ISO che hanno portato alla nascita dello standard SGML e l’istituto di Ginevra. In tale contesto emerge, come progetto interno, la necessità di correlare più documenti tra loro.

Quanto di seguito riportato è tratto proprio dal documento originale[16] che ha dato avvio al progetto del CERN e quindi di fatto ad HTML, diventato poi lo standard mondiale del Web.

Il CERN era una magnifica organizzazione, coinvolgeva svariate migliaia di persone, molte delle quali veramente creative, tutte al lavoro verso obiettivi comuni. Per quanto esse fossero nominalmente organizzate in una struttura gerarchica, non c’erano limiti tra le varie persone per comunicare, scambiare informazioni, apparati e software trasversalmente ai gruppi. La struttura era come una ragnatela multipla, le cui interconnessioni variavano in continuazione. Il personale infatti cambiava ruolo ogni due anni circa. Ciò creava un ambiente estremamente stimolante ma inevitabilmente le varie riunioni finivano con la seguente considerazione: *«OK è tutto posto, ma come possiamo tenere traccia di un progetto così grande?»*

Se il CERN fosse stato un esperimento statico, si sarebbe potuto provare a mettere tutte le informazioni in un *grande libro*, ma esso variava continuamente e sempre nuove idee venivano generate man mano che nuove tecnologie diventavano disponi-

16 <http://www.w3.org/History/1989/proposal.html>.

bili. Ciò rendeva impossibile depositare le informazioni in un solo posto centrale.

La figura che segue (Figura 5) è tratta dal documento originario e può rendere l'idea del problema.

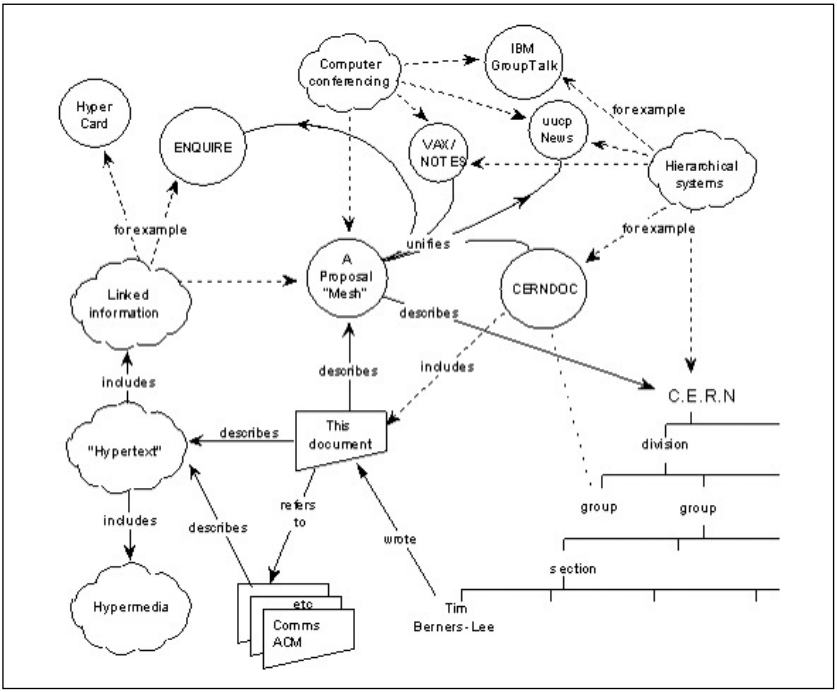


Figura 5. Complessità organizzativa al CERN.

Le informazioni che servivano durante le riunioni erano di questo genere:

- Dove è utilizzato questo modulo?
- Chi ha scritto questo codice e come funziona?
- Che documenti ci sono su questo argomento?
- Quali laboratori sono coinvolti con questo progetto?
- A chi si riferisce questo documento?

Come si può notare il problema della perdita di informazioni era particolarmente grave, ma in effetti il CERN di allora rappresentava solo in miniatura il problema attuale dell'informazione nel Web. Si trattava della correlazione tra più elementi informativi tra loro fisicamente separati.

Oggi lo definiremmo un tipico problema di ipertesto o addirittura di ipermedia. È in questo contesto che Tim Berners Lee usa e di fatto battezza la parola *Web: «The organization is a multiple connected web»*[17].

In questo contesto egli propone un *linked information system* che possa evolvere con l'organizzazione ed il progetto che descrive. Per rendere possibile ciò, il metodo di memorizzazione delle informazioni non può porre restrizioni all'informazione stessa. Ciò perché una ragnatela *Web* di note, ognuna con le proprie connessioni (*link*), può essere molto più utile di un sistema gerarchico centrale.

Era nato il concetto di informazione diffusa nel Web, che sarebbe diventato il modello attuale del mondo moderno.

Una seconda considerazione di Lee era quella che non si poteva parlare di modelli predefiniti di documento, ma di volta in volta occorreva riferirsi a documenti molto differenziati tra di loro. Arriva quindi alla conclusione che occorre sviluppare un modello generico e non un uno specializzato di documento, mentre doveva essere esaltato al massimo il concetto della correlazione fra più tipi di informazioni e quindi di documenti.

Non era stato ancora però definito il modello informatico che poteva supportare questa architettura, ma occorre ricordare che il CERN era stato fortemente coinvolto durante il progetto SGML e quindi aveva piena competenza dei linguaggi di marcatura.

Da tutte queste considerazioni emerse quindi il modello informatico, basato ovviamente sull'SGML, tramite il quale si realizzò uno schema di documento assolutamente generale, dotato

[17] <http://www.w3.org/History/1989/proposal.html>.

di una testatina, contenente le parole chiave, e di un corpo generico. Ma la grande novità è la formalizzazione e quindi la normalizzazione dei link fra più documenti. Un grande passo avanti è stato effettuato: non si concepisce più soltanto il grande documento monolitico (il grande manuale tecnico, la grande enciclopedia) ma anche una serie di piccoli documenti, diversi tra loro ma collegati nel modo più flessibile possibile.

È nato l'HTML in cui l'accento è posto sulla parola *Hyper Text* (HT) ed ovviamente su *Mark-up Language* (ML).

I linguaggi di marcatura

SGML ed HTML

Anche in questo caso occorre fare un po' di storia.

Abbiamo visto come nasce SGML, ne abbiamo capito la forza ma anche le limitazioni ed abbiamo seguito la successiva nascita di HTML, quasi 10 anni dopo la definizione dello standard ISO 8879.

Riassumendo, l'SGML si pone come linguaggio di marcatura generica standard pensato per la realizzazione e gestione di grandi moli documentali, composte da singole istanze di significativa entità. La sua forza risiede nella standardizzazione, nella capacità di interscambio tra sistemi diversi, nella possibilità di ipotizzare processi molto più complessi, in quanto il documento diventa dato. Il suo limite sta nella complessità del processo che porta alla realizzazione di una simile architettura documentale; infatti, al di là di enti che hanno una grande capacità di spesa (mondo della difesa, mondo aeronautico, mondo automobilistico, grandi editori e pochi altri), tutti gli altri trovano tale modello troppo complesso da affrontare, troppo rigido, troppo costoso.

Anche un certo approccio, forse troppo accademico o elitario, della comunità SGML contribuisce a rendere questo mondo distante dall'utenza normale e quindi a limitarlo solo ad ambienti

molto specifici. A ciò si somma un importante elemento commerciale: Microsoft sta tentando di affermare nel frattempo il suo standard RTF e quindi fa una guerra feroce alla comunità SGML. In questo scenario l'arrivo di HTML è dirompente, in quanto l'utente della strada scopre improvvisamente che questi linguaggi complessi permettono di generare delle semplici pagine e di correlarle rapidamente tra di loro: è nato lo sviluppatore di pagine Web. La tecnologia viene in soccorso e si sviluppano prontamente dei visualizzatori di codice HTML capaci non solo di visualizzare il semplice documento, ma anche di spostarsi su un altro documento semplicemente cliccando sulla parola ipertestuale: è nato il browser Web.

Nel giro di poco tempo si affermano sul mercato due browser: Mosaic e Netscape. Fino a questo punto Microsoft ha continuato a proporre il suo formato RTF ed il suo browser per la visualizzazione della sua versione ipertestuale, appunto l'Help di Windows.

Ma il mondo *parla ormai Web* e Microsoft capisce che la partita è persa e che rischia di perdere l'appuntamento con Internet ormai diventata una realtà. Netscape si è intanto affermata e sta diventando per certi aspetti concorrenziale con la stessa Microsoft.

A questo punto Microsoft compie due mosse: compra il rivale di Netscape (NSCA Mosaic), gli cambia il nome in Explorer e, contemporaneamente (spalleggiata in questo dalla SUN), entra con forza nei comitati di sviluppo degli standard ed impone la nascita di un nuovo standard, denominato XML (*eXtensible Mark-up Language*).

XML

Cerchiamo quindi di capire che cosa è XML: è solo un'operazione commerciale? o c'è qualcosa di più tecnico alle spalle?

Ci sono entrambi: da un lato, si cancella la dizione SGML (oggi addirittura non più prevista dalla lista del W3C), dall'altro

si integrano all'interno della struttura SGML due nuovi componenti fondamentali.

Il primo è il recupero dell'hyperlink, così come modellizzato, sviluppato e fatto crescere dall'ambiente HTML.

Il secondo è l'eliminazione dell'obbligo della DTD: essa infatti diventa opzionale mentre si inserisce un nuovo concetto, quello di un documento *well formed* ossia *ben costruito*. In pratica questo concetto è molto semplice: esso consiste nel fatto che ogni tag aperto debba essere anche correttamente chiuso.

A latere di queste varianti fondamentali vengono anche eliminate molte *incrostazioni* che negli anni si erano aggiunte (o non erano state eliminate) allo standard SGML, quali i concetti di *omittag*[18], di ricorsività illimitata di entità SGML e varie altre. Molte di queste erano già state di fatto eliminate dalla prassi applicativa dello standard.

In sostanza quindi SGML ed XML sono estremamente vicini, ma la mancanza dell'obbligo del modello formale (DTD) apre la strada all'uso integrato dell'SGML con i DBMS *(Database Management Systems)*, come verrà meglio chiarito di seguito, e quindi una migliore interfacciabilità dei documenti con i database tradizionali a tabelle. Questa proprietà era di specifico interesse dei grandi gestori dei database e quindi dei fornitori hardware a loro collegati (nel caso specifico SUN) che non a caso erano entrati nei comitati degli standard.

Ovviamente, nel giro di poco tempo, i fornitori di browser dichiararono la loro compatibilità al nuovo standard anche se si pose il grosso problema della visualizzazione dell'informazione, a tutt'oggi solo parzialmente risolto.

Occorre fare alcune considerazioni di base: il linguaggio SGML dichiarava esplicitamente una totale separazione tra contenuto e forma. In Figura 6 viene indicato un esempio del concetto di contenuto e forma.

---

[18] Possibilità di omettere interamente alcuni tag.

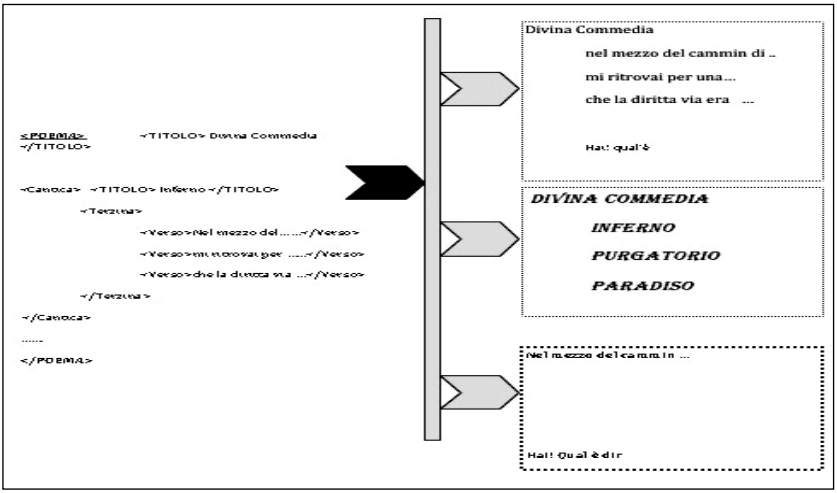


Figura 6. Forma Vs Contenuto.

Nell'esempio all'istanza *divina commedia* del modello *poema* vengono applicati tre stili: ognuno di essi ne evidenzia caratteristiche particolari. Non soltanto possono essere variati gli stili dei caratteri e il loro corpo, ma si può decidere se mostrare o meno parti coerenti del documento.

I browser SGML puri, del periodo tra il 1986 ed il 1996, erano dotati di uno *style editor* necessario per creare la forma da applicare al contenuto.

Gli stili, che di fatto erano dei file in SGML che descrivevano la formalizzazione della visualizzazione/medializzazione non erano standard ma venivano realizzati tramite tecnologia proprietaria degli sviluppatori dei browser. Era nata una iniziativa di standardizzazione dello stile chiamata DSSSL[19] (in gergo Diesel) ma non aveva avuto un grande successo. Questa iniziativa era poi di fatto morta a causa di HTML.

19 *Document Style Semantics and Specification Language.*

L'HTML infatti, per non complicare la vita al *popolo delle pagine Web*, non aveva mantenuto questa separazione rigida tra forma e contenuto, ma i tag trattavano indifferentemente l'uno e/o l'altro. Ciò aveva semplificato notevolmente anche lo sviluppo dei browser HTML, che non doveva occuparsi di questa complessa gestione della esplicitazione di un qualsiasi modello di contenuto tramite una qualsiasi forma di stile. In questo senso i browser SGML erano estremamente più complessi e di fatto ad oggi mai raggiunti dalla tecnologia attuale. L'HTML può essere infatti considerato quasi un linguaggio procedurale (*specific markup language*) e come tale è infatti molto più semplice da trattare.

Tornando ora all'XML, i browser HTML si trovarono a dover trattare di nuovo un problema di separazione tra contenuto e forma, a cui non erano pronti a reagire, proprio perché progettati a fronte di un linguaggio procedurale.

Si è quindi riesumato il concetto di stile, che, attraverso varie evoluzioni (CSS[20], XSL[21], XSLT[22]), ha riproposto il problema della standardizzazione della formalizzazione. Come il suo predecessore DSSSL, ha avuto vita dura, in quanto i produttori di browser hanno tentato di risolvere questo problema riconvertendo in HTML il sorgente XML, interpretato alla luce del file di stile. Di fatto, ad oggi, non è ancora stato riproposto un browser che risolva in modo nativo (ossia non passando attraverso HTML o suoi derivati) lo stile.

Le varie iniziative che descrivono gli stili, gestiti dall'ente che ha standardizzato il Web, ossia il W3C, verranno descritte nell'ultima parte di questo capitolo.

---

20 *Cascading Style Sheets.*
21 e*Xtensible Stylesheet Language.*
22 *eXtensible Stylesheet Language Transformations.*

XML come elemento di congiunzione tra linguaggi di descrizione documentale e metodi di descrizione dei contenuti dei DBMS

Nel tentativo di approfondire alcune caratteristiche di XML, è importante comprendere il motivo per cui viene considerato un vantaggio la non obbligatorietà della DTD. Per capirne il significato occorre risalire all'ultima fase di sviluppo di SGML all'inizio degli anni '90. In quel periodo si tentava l'integrazione tra basi dati documentali e basi dati tabellari.

Il problema che è subito emerso è che i database avevano ovviamente un loro modello di riferimento, formulato dall'analista durante la relativa fase di progettazione. Quando si tentava di integrarlo con il documento, occorreva che entrambi ragionassero secondo un modello comune. Per ottenere ciò, si ricorreva a complicate tabelle di correlazione il cui risultato non era sempre efficace.

L'XML, non avendo la necessità di esporre il suo modello, può tranquillamente accettare e ospitare il modello del database integrandolo all'interno dalla sua struttura.

Formuleremo in termini teorici questo concetto nel successivo paragrafo, dopo aver preso confidenza con le relazioni tra documento, modello del documento e base dati giocando un po' con le parole.

Per far ciò ci riferiremo al modello di ricetta già utilizzato all'inizio di questo percorso (Figura 7).

Poniamoci alcune domande: la ricetta è un documento o un database? Forse nessuno dei due: questi dati possono essere espressi sotto forma di documento o possono essere classificati tramite un database; sostanzialmente è un insieme di dati.

Al di là del gioco di parole, la realtà è che i tre concetti (documento, database, dato) sono confondibili poiché di fatto sono la stessa cosa.

Introduciamo un terzo concetto: la ricetta è un processo? La risposta non può che essere positiva.

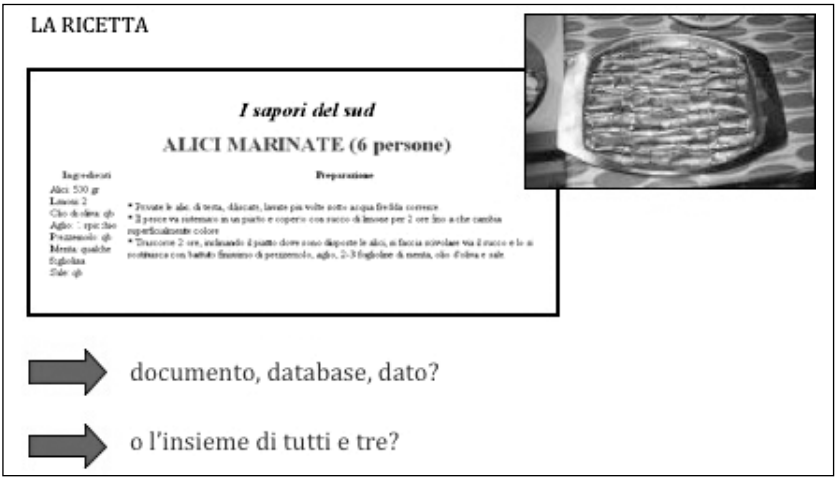
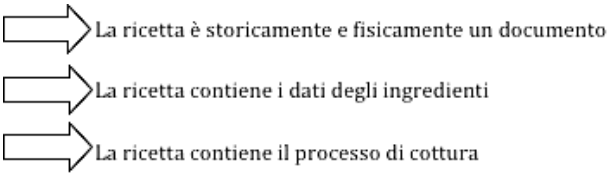


Figura 7. Documenti, modelli, database.

Proviamo a fare un po' di ordine. La moderna civiltà industriale si basa sui processi, i processi hanno bisogno dei dati, in molti casi i dati sono rappresentati tramite documenti. Tutto questo mette in ordine le parole con cui abbiamo liberamente giocato, ma pone il problema concreto di realizzare e di modellizzare questa interazione tra processo, dati e documenti.

Quindi riassumendo graficamente:



Riesaminando quindi la ricetta, e guardandola come una base dati otteniamo le informazioni indicate nella figura sottostante (Figura 8), nella quale siamo riusciti ad estrarre i titoli delle ricette, il procedimento con cui si realizzano, gli ingredienti necessari.

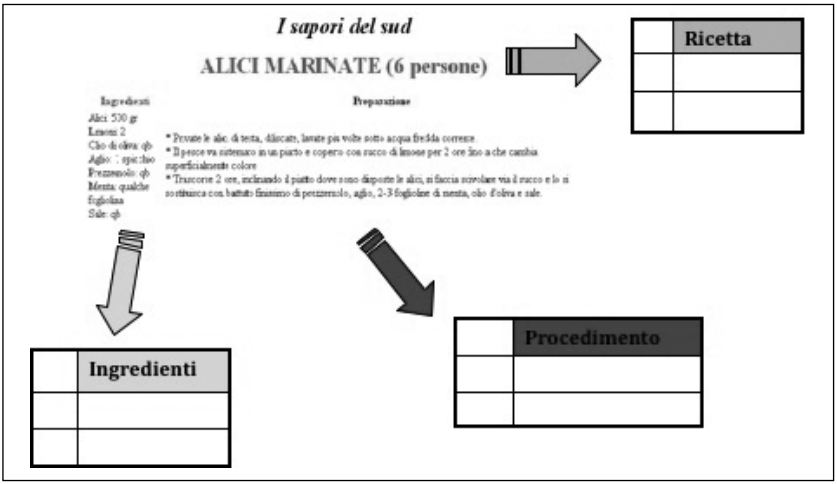


Figura 8. Ricetta sottoforma di database.

Effettuiamo ora la stessa operazione sul modello documentale in XML della ricetta. riportato nella figura sottostante (Figura 9).

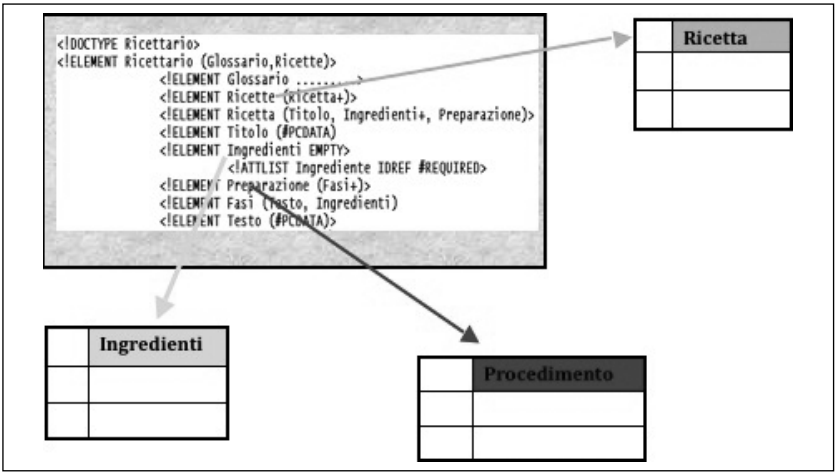


Figura 9. Ricetta strutturata in XML.

Come si può notare, si possono estrarre esattamente le stesse informazioni della precedente figura, ma questa volta si sta lavorando sulla modellistica documentale e non sul database.

Modelli arborescenti, modelli piatti, modelli misti

Un modello arborescente è un modello dati che, partendo da un elemento radice, genera una serie di contenitori, ognuno dei quali ne contiene altri in una serie potenzialmente infinita.

Tale modello prende ovviamente nome dal concetto di albero e dalla sua ramificazione dal tronco verso i rami e quindi verso le foglie. Nella Figura 10 è rappresentato il modello dei rami principali della ricetta, in cui peraltro non tutti sono stati esplosi né in profondità né nella loro molteplicità. Ad esempio sono state mostrate soltanto due ricette e di una sola di queste sono stati mostrati i livelli inferiori (livello tre e livello quattro). Se avessimo dovuto presentare il modello reale, l'albero sarebbe stato molto più complesso ed articolato.

È da notare che volutamente abbiamo lasciato la voce glossario separata e non sviluppata per preparare i successivi argomenti.

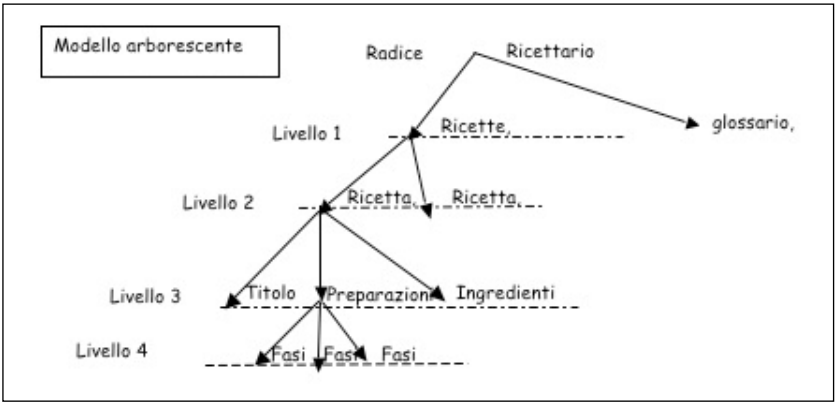


Figura 10. Modello arborescente.

Questo modello è sviluppato a fronte di una DTD (ossia quelle riportata al precedente paragrafo) ed è possibile effettuare un controllo tra l'istanza del documento e il suo modello.

La possibilità di un controllo è un concetto assolutamente lecito (anzi obbligatorio) in SGML mentre, come abbiamo spiegato, è solo opzionale in XML. Cerchiamo di capire graficamente perché questo è un vantaggio.

Osserviamo ora una struttura che apparentemente ha poche varianti, ma vedremo che queste sono sostanziali.

La struttura, definita a modello piatto, è tipica di una tabella in cui il nodo padre (radice) è il nome della tabella ed il primo elemento della struttura piatta è tipicamente il titolo della colonna. Tutte le successive ripetizioni della stessa struttura rappresentano le righe della tabella. Possiamo ripetere lo stesso concetto parlando di campi e di record, ossia parlando di database.

Ipotizziamo ora di utilizzare questo tipo di struttura per realizzare la base dati di un glossario in cui ogni riga sia composta da una voce e dalla sua spiegazione (Figura 11).

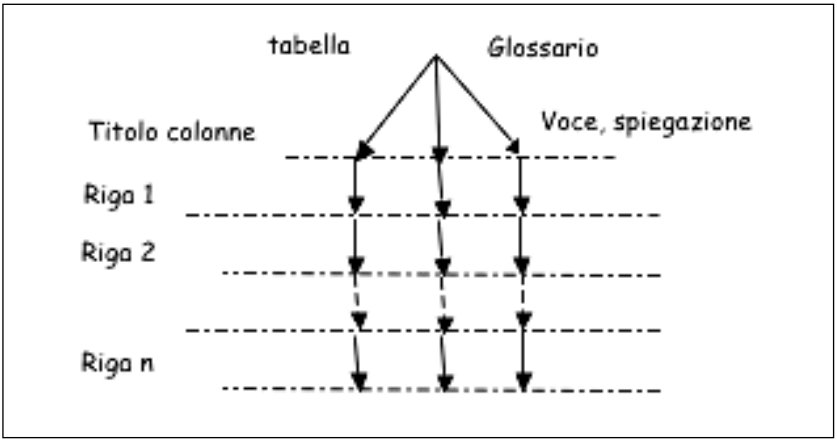


Figura 11. Modello piatto.

Come si può notare, e peraltro come era facile intuire, c'è una correlazione precisa tra una tabella ed un potenziale database. Ovviamente il database ha il suo modello.

Inseriamo ora detta tabella all'interno del mondo documentale arborescente precedentemente definito. Come possiamo vedere in Figura 12, abbiamo applicato il modello tabella all'elemento glossario, lasciato precedentemente appositamente libero.

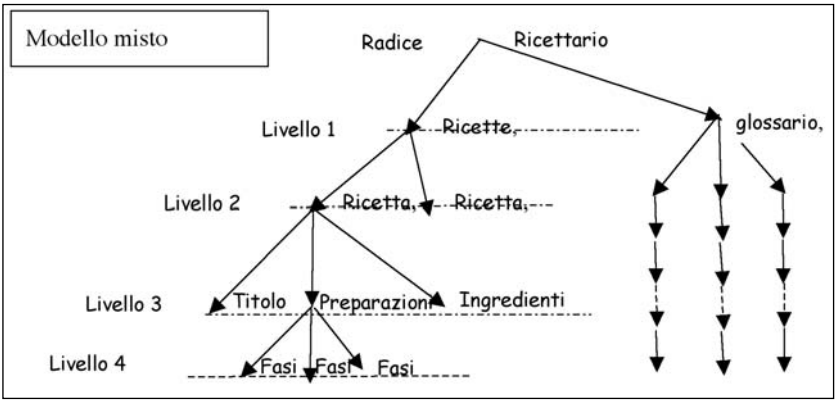


Figura 12. Modello misto.

Siamo quindi riusciti a realizzare un modello misto composto da una struttura documentale e da una struttura tabellare.

Per realizzare ciò in SGML, il modello del database avrebbe dovuto essere assolutamente identico al modello della tabella descritta documentalmente.

In XML, invece, possiamo usare il modello documentale per la parte documentale ed il modello proveniente dal database per la parte tabellare.

Quando il gioco si fa più complicato ed i modelli documentali devono integrare modelli di tabelle provenienti da database estremamente complessi e sofisticati, questo approccio diventa vincente.

Non si sottovaluti questa problematica. Per rendersi conto di quanto questo sia importante, basta guardare con occhio critico qualsiasi pagina di un sito Web che fornisca elenchi di voli, alberghi, orari, conti bancari, informazioni assicurative, ecc. Senza questo approccio il Web avrebbe dovuto fare a meno (o almeno avrebbe avuto difficoltà a gestire) di una delle informazioni più importanti disponibili oggi: l'enorme quantità di dati accumulati negli ultimi cinquant'anni da database di ogni tipo e dimensione.

I linguaggi di marcatura ed il WEB 2.0

Le prime righe della definizione di Web 2.0 estratta da Wikipedia[23] possono essere così sintetizzate: il termine Web 2.0 è nato dalla Web 2.0 conference di O'Reilly alla fine del 2004. Il termine 2.0 è derivato dalla notazione usata nello sviluppo software che indica la *release* di un software.

È opinione personale che la sintesi sia assolutamente corretta: il Web 2.0 è una necessità commerciale per coloro che sviluppano software.

Ciò non vuol dire che non ci siano state evoluzioni nel settore del *social networking*, ma oggetti come le BBS (*Bulletin Board System*) e le ontologie esistevano molto prima del Web *taggato* e svolgevano lo stesso ruolo; la poderosità tecnica ed economica dell'impianto attuale del Web hanno fatto il resto fornendo uomini e mezzi a vecchie idee.

Questo ragionamento non è valido in assoluto poiché fenomeni quali *wiki* o *google maps* non avevano precedenti ma, per quanto attiene i contenuti a cui questo volume può essere interessato, i due fenomeni che emergono sono lo stesso *wiki* e le

[23] <http://it.wikipedia.org/wiki/Web_2.0>.

metodologie di ricerca semantica che si sono particolarmente sviluppate in questo periodo. Per converso, i tag, in quanto identificatori di contenuti e di strutture, hanno probabilmente perso qualche punto a favore della ricerca dei significati e degli strumenti SW (software) puri. L'XML, nato per i documenti, serve sempre di più come comodo vettore di dati non documentali piuttosto che come organizzatore di strutture documentali.

Non a caso le aziende che producono motori di ricerca hanno avuto il sopravvento (economico) su quelle che producono computer e basi dati tradizionali e sono anche le stesse che cavalcano il tema delle mappe e della geolocalizzazione.

Guardando come persone di tutte le età e di tutti i ceti si avvicinano al Web risulta chiaro che la variante fondamentale è quella sociale. Se si prova a valutare (in termini di uso reale) la somma di applicazioni per viaggi, sesso e *chiacchiere di cortile* si raggiunge probabilmente oltre il 90%.

D'altra parte le civiltà moderne avevano perso la *piazza del paese* ed in un certo senso ne è stata trovata una che permette di partecipare senza muoversi e (apparentemente) uscire dal gioco quando non si vuole condividere. Molti non si rendono conto della sua potenziale pericolosità.

Sempre da Wikipedia: *«Gli scettici replicano che il termine Web 2.0 non ha un vero e proprio significato e dipenderebbe principalmente dal tentativo di convincere media e investitori sulle opportunità legate ad alcune piattaforme e tecnologie»*.

Io sono fra quegli scettici.

Da quanto annunciato nei precedenti paragrafi, dovrebbe risultare evidente l'importanza che i linguaggi di marcatura hanno oggi. La dimensione assunta dal Web ha portato al fatto che la loro standardizzazione sia di fatto compito dell'ente che oggi si occupa di promuovere, gestire e custodirne gli standard, il W3C.

È quindi nel W3C che oggi ha luogo la continuazione delle iniziative prese nell'ambito dell'ISO.

In questo senso il sito istituzionale del consorzio[24] è il punto di riferimento assoluto per chi intenda operare in questo settore e per chi voglia cogliere le tendenze in atto per quanto attiene ai linguaggi di marcatura.

Prima di commentare il lungo elenco di iniziative ricavate appunto dall'esame del sito, occorre però fare alcune considerazioni di metodo.

I linguaggi di marcatura oggi non sono più connessi ai problemi documentali; l'evoluzione di XML, che ha portato alla integrazione con le basi di dati esistenti, ha anche dimostrato che il linguaggio di marcatura si prestava alla codifica di molte altre sorgenti informativ. In questo modo esse fruiscono gratuitamente di tutta la metodologia costruita e di tutte le tecnologie relative.

È quindi nata una serie di nuovi linguaggi per la descrizione formale di grafiche, musiche, voce, sincronizzazione multimediali, ecc.

Esaminiamo ora alcuni di questi standard consolidati o emergenti, per come essi appaiono, ad oggi, sul sito istituzionale del W3C[25]. Dato l'elevato numero di standard gestiti dal Consorzio ci si limita ad esaminare quelli che hanno un senso in questo contesto:

[24] <http://www.w3.org/>.
[25] <http://www.w3.org/standards/>.

FEBBRAIO 2013	SIGNIFICATO
Accessible Rich Internet Applications (WAI-ARIA)	La *Web Accessibility Initiative* (WAI) è pensata per avvicinare al Web coloro che attualmente non ne sono utenti.
WCAG	*Web Content Accessibility Guidelines*: esiste concettualmente la possibilità di poter lavorare sugli stili per facilitare l'accesso ai disabili. In tal senso offre delle grandi possibilità, in particolare ai non vedenti, ai non udenti, a coloro che presentano difficoltà di apprendimento o deficienze cognitive, che presentano limitazioni nei movimenti o difficoltà di parola, ecc.
Authoring Tool Accessibility Guidelines (ATAG)	Oggi è diviso in due parti: • Parte A: *tool di authoring* per autori disabili; • Parte B: tool per la creazione di contenuti per disabili.
Amaya	Amaya mira a sviluppare un ambiente neutro (non proprietario) per quanto attiene agli editori ed ai browser XML. In questo senso probabilmente tenta di ricostituire un ambiente di modellistica documentale di classe (complessità) SGML. Infatti include capacità di gestire DTD multiple, documenti multipli ed entità differenziate (SVG, MathML, ecc.) in un contesto unitario. Allo stesso tempo integra le capacità di hyperlink di HTML. Come i vecchi browser di classe SGML, include vari *layer* collaborativi, in cui i vari autori possono condividere le cosiddette *annotation*.
CC/PP	Il *Composite Capabilities/Preferences Profile* è un insieme di regole suggerite dal W3C per la creazione di schemi di descrizione dei dispositivi (preferibilmente adottando il linguaggio RDF[26]), tali da semplificare i meccanismi di negoziazione tra server di origine e dispositivo di fruizione. Un profilo CC/PP, quando presente nel dispositivo, contiene una descrizione delle caratteristiche del dispositivo stesso e delle preferenze dell'utente.
Compound Document Formats (CDF)	Questa iniziativa tenta di risolvere il problema di far convivere vari linguaggi di marcatura in un contesto unitario. Le problematiche individuate sono:

[26] *Resource Description Framework.*

	• nell'area della propagazione degli eventi, trasversalmente ai *namespace* propri di ogni linguaggio; • nel *rendering* combinato sullo schermo in presenza di più linguaggi; • nel modello di interazione con il computer in presenza di più linguaggi.
CSS	*Cascading Style Sheets* è il meccanismo più elementare individuato sul Web per gestire lo stile in XML. Infatti CSS è il primo linguaggio di stile che può essere associato ad un'istanza XML per permetterne una visualizzazione definita da un protocollo. Al CSS sono associate iniziative quali *CSS Validator* per permettere la validazione di file di stile.
CSS Mobile	È indirizzata ai dispositivi mobili e soprattutto ai cellulari.
DOM	Questa è un'importante iniziativa del Web, che attiene molto di più al dominio del software che non a quello della documentazione. Data la sua importanza tuttavia è opportuno darne qualche cenno. Il DOM (*Document Object Model*) è una piattaforma ed un linguaggio di interfacciamento neutrale che permette ai programmi ed agli *script* di accedere dinamicamente e di aggiornare il contenuto, la struttura e lo stile dei documenti. Ci sono molte iniziative legate al DOM.
DOM event	Il *DOM event* gestisce il collegamento tra gli elementi di un documento ed un'occorrenza quale il click del mouse o la gestione di un errore.
Efficient XML Interchange	Anche questo gruppo si occupa della problematica di ricollegare le varie iniziative e i vari linguaggi di marcatura. Nello specifico deve sviluppare una sintassi RDF/XML che contenga dei meccanismi per legarsi ad altre sintassi XML (specificatamente l'XHTML[27] o altri microformati). Come si può notare, lo sviluppo commerciale dei linguaggi di marcatura sta generando una notevole quantità di dialetti, così quasi da contrapporsi ai principi di interscambio che erano alla base degli standard di partenza. Si è evoluto come ottimizzatore dell'allocazione delle risorse computazionali.

[27] *eXtensible HyperText Markup Language.*

Electronic Commerce	Il documento descrive le attività che, nel prossimo futuro, potrebbero coinvolgere il W3C sul tema del Commercio Elettronico.
eGovernement	La tecnologia WEB permette ai governi di scambiare bidirezionalmente con i cittadini una grande quantità di informazioni. Questa disponibilità deve fronteggiare vari problemi relativi all'ambiente, agli aspetti legali, politici e culturali.
Evaluation and Report Language (EARL)	L'*Evaluation and Report Language* è un vocabolario standardizzato per esprimere i risultati di test. L'obiettivo è quello di realizzare una piattaforma indipendente per il loro scambio.
Geospatial	Questa specifica definisce un'API[28] per l'accesso ad informazioni di localizzazione geografica.
Government Linked Data	Questa attività comprende varie iniziative legate alla standardizzazione dei dati di interesse governativo. Tra questi vanno segnalati: • Termini per descrivere le persone quali nomi, indirizzi o riferimenti ad altri elementi tra cui organizzazioni e progetti; • Una *core ontology* per strutture organizzative.
GRDDL	*Gleaning Resource Descriptions from Dialects of Languages.* Questo gruppo lavora su problematiche simili a quelle sopra descritte per l'*Efficient XML Interchange*, infatti esiste un problema di riduzione e di descrizione dei *Dialects of Languages*. Il problema è legato al fatto che, quando si usa uno standard, al di là della sua definizione formale, gli utenti tendono a deformarlo durante lo sviluppo delle applicazione pratiche. Occorre infatti una particolare mentalità formale per mantenersi aderenti agli standard, anche durante l'uso pratico degli stessi. In molti casi, alcuni *utenti della strada* tendono a trovare delle *scorciatoie* di fronte a problemi di interpretazione dello standard.

[28] *Application Programming Interface.*

Health Care and Life Sciences (Semantic Web)	Una *knowledge base* per il settore biomedicale usa le seguenti ontologie: • *Ontology of Rhetorical Blocks*; • *Neuromedicine (SWAN) Ontology.* Fornisce anche indicazioni per l'allineamento tra le ontologie *SWAN - Semantic Web Applications in Neuromedicine* e *SIOC - Semantically-Interlinked Online Communities*.
HTML	Dato quanto detto precedentemente, non occorre ridescrivere che cosa sia HTML, ma deve ovviamente esistere una sezione del sito istituzionale dove è depositato tutto quanto attiene alla formalizzazione del linguaggio. Naturalmente in questo stesso sito sono presenti tutte le varianti o microvarianti di argomenti collegati allo standard, alcune delle quali sono: • *HTML Tidy;* • *HTML Validator;* • *XHTML 2;* • *XHTML For Mobile;* • *XHTML Modularization.*
InkML	*Ink Markup Language*  Questo gruppo si sta occupando di scrivere un formato dati in XML che rappresenti un *digital ink* quale può essere generato da una penna o da uno stilo elettronico. Lo studio include problematiche quali: il riconoscimento della scrittura di testi, formule matematiche e formule chimiche.
Internationalization	Questo gruppo si occupa di rendere sempre più globale tutto quanto attiene al Web. *Internationalization of:* • *Web Design and Applications;* • *Web Services;* • *XML;* • *Web Architecture*.
MathML	Questo gruppo si occupa della definizione formale di un linguaggio che descriva in maniera univoca le formule matematiche, così da permettere un passaggio tra più sistemi dello stesso algoritmo. Si occupa anche della visualizzazione corretta dell'algoritmo su una pagina Web.

Media Access	Norma l'accesso ai *media* attraverso due strumenti: • Definizione della sintassi per costruire le URI di elementi mediali ed il metodo per gestirle quando si usa il protocollo http; • Definizione di un *core vocabulary* delle risorse mediali.
Mobile WEB Application *Mobile WEB for social development* *XHTML For Mobile*	Questa iniziativa si occupa di rendere disponibile, efficace e standardizzato l'approccio al Web tramite dispositivi mobili. È di notevole interesse in quanto il proliferare di piccoli dispositivi mobili (telefonini, palmari, ecc.) rende disponibile ad un'utenza estremamente diffusa la possibilità di fruire di informazioni. Ovviamente anche in questo caso si usa un linguaggio di marcatura.
Multimodal Web Applications	Questo gruppo studia nuovi metodi per l'interazione con l'informazione; in particolare si pone l'obiettivo di interagire attraverso un metodo *multi modo*; per esempio integrando i dispositivi di ingresso al computer quali lo stilo con i comandi vocali. Ciò perché molti dispositivi di nuova generazione sono piuttosto miniaturizzati e quindi la mancanza di tastiere classiche, associate all'opportunità di una maggiore diffusione dei dispositivi, rende interessante la ricerca di nuovi modelli di interazione sia in ingresso che in uscita.
OWL	Il gruppo si occupa dello sviluppo del *Web Ontology Language*. Esiste un capitolo specifico sull'ontologia all'interno di questo libro.
DCCI	La DCCI (*Delivery Context: Client Interfaces*) usa una ontologia per fornire un modello formale delle caratteristiche dell'ambiente in cui i dispositivi interagiscono con il Web o con altri dispositivi.
P3P	Usando questo linguaggio, un utente può esprimere le sue preferenze permettendo al suo *user agent* di prendere decisioni automatiche o semiautomatiche sulle proprie politiche di privacy.

PNG	Il documento descrive il file grafico PNG (*Portable Network Graphics*) che può essere usato per immagini *raster* e che può sostituire il TIFF in molte applicazioni.
POWDER *Provenance*	Protocollo per le WEB *Description Resources*. Questo set di specifiche mira a definire i vari aspetti che sono necessari in un ambiente eterogeneo come il WEB, per acquisire la visione dell'interoperabilità delle informazioni di *provenienza*.
RDB2RDF	Questo documento definisce un metodo per la mappatura diretta tra dati relazionali e RDF.
Semantic Annotation for WSDL and XML Schema	Questo documento definisce una serie estesa di attributi per i *Web Services Description Language and XML Schema* che permette la descrizione di semantiche addizionali dei componenti WSDL. La specifica definisce come fare annotazioni semantiche facendo riferimento a modelli semantici esterni (ontologie).
RDFa	Il *Resource Description Framework* integra una pletora di applicazioni quali: cataloghi librari; indici a livello del Web globale, aggregazioni di notizie, software e contenuti vari, collezioni personali di musica, foto o eventi, usando l'XML come sintassi di interscambio. Fa parte di questa attività il *Semantic Web* che fornisce un ambiente comune che permette ai dati di essere scambiati e riutilizzati trasversalmente alle applicazioni, alle aziende ed alle comunità.
RIF	Le *Rule Interchange Format* (RIF) sono state create per essere uno standard per lo scambio di regole tra sistemi (in particolare nel Web).
Security for User Agents *Security for Web Applications*	Linee guida per i contesti legati alla *Web security* verso gli *end-user.*
SKOS	Questo documento definisce il *Simple Knowledge Organization System* che è un modello di dati comuni per condividere e collegare *Knowledge Organization Systems* tramite il Web

SML	*Service Modeling Language*: questo gruppo di lavoro opera nella definizione di un modello adeguato alla manutenzione dei sistemi. In questo risente molto delle attività del gruppo di lavoro dell'industria aeronautica S1000D (ex AECMA100D). È uno dei gruppi più importanti come modello applicativo.
SMIL	*Synchronized Multimedia Integration Language*. Il linguaggio SMIL permette *l'authoring* di presentazioni audiovisive e interattive. È comunque da molto tempo che i gruppi di lavoro SGML ed XML stanno tentando di trovare soluzioni per la gestione standardizzata della componente multimediale ed in particolare delle problematiche temporali di sincronizzazione che essa comporta. Negli anni '90 aveva raggiunto un notevole livello di diffusione lo standard *HyTime* di cui in qualche modo lo SMIL è l'erede.
SPARQL	SPARQL è un *query language* per RDF che permette di esprimere query che coinvolgano più data source.
SVG *SVG Tiny*	*Scalable Vector Graphics* L'SVG è un linguaggio per descrivere grafiche bidimensionali ed in generale applicazioni grafiche in XML. SVG permette concetti nuovi come l'integrazione tra l'informazione contenuta all'interno del testo con quelle contenute all'interno della grafica. Si stanno differenziando delle varianti per i diversi tipi di dispositivi di fruizioni: *SVG Tiny* è definito per i cellulari, il secondo profilo (*SVG Basic*) è adeguato ai PDA (*Personal Digit Assistant*).
Timed Text	Questo gruppo di lavoro si occupa del problema di standardizzare il modo di gestire testi sincronizzati (quali i sottotitoli sincronizzati con gli audio).
URI	Le URI (*Uniform Resource Identifiers*) rappresentano sia l'elemento centrale del *framework* che il link tra l'RDF ed il Web.
UAAG	Questo documento fornisce delle linee guida per progettare *user agents* che abbassino le barriere per l'accesso al Web per i disabili. Gli *user agents* includono i browser HTML ed altri tipi di software che collegano e presentano i contenuti del Web.

Audio *Voice*	*Voice Browser Activity* Fanno parte di questa categoria le seguenti attività: • *Speech Synthesis Markup Language (SSML);* • *Voice Extensible Markup Language (VoiceXML).* Entrambi gli standard mirano alla normalizzazione delle informazioni che generano lo *stile* dell'audio sintetico.
WebCGM	Questo è uno standard non XML, ma è lo standard grafico per il mondo della documentazione tecnica. Pertanto è in atto un'attività per normalizzare il suo equivalente (WebCGM) per il mondo Web.
WEB AND TV	Questo documento indirizza le applicazioni di tipo televisivo e differenzia i requisiti in due categorie: • *Home Networking Scenarios;* • *TV Broadcast URI Schemes*
Web Fonts	Questa iniziativa propone nuovi approcci ai *font* nel Web. Sono in corso due azioni: • WOFF che definisce una compressione leggera e di facile uso di *font data* da usare con i CSS *@font-face;* • CSS3 che definisce come specificare i font e come le *font resource* sono caricate dinamicamente.
Web IDL	Questo documento definisce un linguaggio per la definizione delle interfacce (Web IDL) e può essere usato per descrivere le interfacce da implementare nei Web Browser.
WebRTC	Questo documento fornisce specificazioni per le Comunicazione in Tempo Reale tra i Browser nel Web
Web Services	Vari standard collegati ai *Web Services*: *Addressing*: descrive le collaborazioni *peer-to-peer* tra i partecipanti; *Choreography*: vede il sistema da un punto di vista globale ed i presupposti, comuni e complementari, osservabili. *Architecture*: Varie note legate alla Internazionalizzazione, *Life Cycle, Glossary, Scenarios*, ecc.
WSDL	*Web Services Description Language*: questo documento definisce un set esteso di attributi per gli XML Schema per la descrizione di semantiche addizionali (facendo riferimento a modelli semantici) per i componenti WDSL.

XBL	*Xenogamous Binding Language*: forniva un metodo per aggregare vari elementi in un documento con script, *event handlers*, CSS ecc. L'iniziativa è dismessa.
XInclude	Questa specifica introduce un meccanismo generico per fondere documenti XML nei casi in cui l'applicazione richieda questa capacità.
XKMS *XML Signature*	Questi documenti forniscono protocolli per distribuire e registrare *chiavi pubbliche* adeguate per l'uso con *XML Signature* e *XML Encription*. Le *XML Key Management Specification* sono formate da due parti: le *XML Key Information Service Specification* (X-KISS) e le *XML Key Registration Service Specification* (X-KRSS).
XLink	*XML Linking Language*. Questo linguaggio permette di inserire in modo standardizzato dei link tra le risorse, all'interno dei documenti XML. Il linguaggio può descrivere sia semplici hyperlink monodirezionali sia link più complessi.
XML- XML Base	Questa è la specifica base XML, ovviamente ospitata all'interno del sito istituzionale. Si vanno sommando una pletora di specificazioni di dettaglio: • *XML Canonicalization;* • *XML Design Techniques;* • *XML Events;* • *XML Fragments;* • *XML Pipeline (XProc);* • *XML Relationship to other formats;* • *XML-binary Optimized Packaging.*
	La missione di questo *working group* è di sviluppare un processo per la crittografia/decrittografia di dati digitali, ovviamente con particolare interesse per i file XML o per porzioni di esso.
XQuery	Questo linguaggio usa l'XML come linguaggio strutturato per potere effettuare interrogazioni *intelligenti* su vari tipi di dati. Le query sono particolarmente efficaci su dati rappresentati a loro volta tramite file XML.

XML Schema	L'XML schema è l'equivalente della DTD dell'SGML. Questa specifica standardizza come costruire il modello documentale in XML.
XPath	Questo linguaggio permette di accedere in modo standard ad una parte (interna) di un file XML. Il linguaggio è usato dall'interno del linguaggio XSLT che viene descritto di seguito.
XPointer	*L'XML Pointer Language* (XPointer), e un linguaggio usato per puntare qualsiasi *URI-reference* al'interno del Web.
XSL and XSLT	Questi linguaggi sono utilizzati per la gestione avanzata dello stile in ambito XML. Ne sono previsti due tipi: • *XSL Formatting Objects* (XSL-FO) Questo è un vocabolario XML per specificare la semantica della formattazione; • *XSL Transformations* (XSLT) Un linguaggio per effettuare trasformazioni dell'istanza XML. *Content Trasformation* L'attività è riferita al nuovo ambiente *mobile* e si riferisce alla gestione delle richieste e delle risposte, realizzata tipicamente dai proxy ed ottimizzata per fornire all'utente una vista che possa risultare soddisfacente sul dispositivo in uso.
Xforms	XForms rappresenta la prossima generazione di *forma* del WEB. Una *XForm based WEB form* processa i dati in XML separando presentazione, scopo e contenuto attraverso tre meccanismi: • Un modello dichiarativo composto da formule, limiti, tipi di dati ecc. per il processo dei dati; • Un *layer* di interfacciamento con l'utente; • Un *controller* per orchestrare le manipolazioni e le interazioni tra i restanti due meccanismi.

Bibliografia

ISO 8879:2005, *Standard Generalized Markup Language*
ISO/IEC 10646-1:2000, *Information technology -- Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane*
ISO/IEC 2022:1994, *Information technology – Character code structure and extension techniques*
ISO/IEC 646:1991, *Information technology – ISO 7-bit coded character set for information interchange*
ISO/IEC 8859-1:1998, *Information technology — 8-bit single-byte coded graphic character sets — Part 1: Latin alphabet No. 1*

Sitografia

<http://it.wikipedia.org/wiki/Web_2.0>
<http://www.lulu.com/shop/david-jones/the-xml-primer/paperback/product-282753.html>
<http://www.sgmlsource.com/history/sgmlhist.htm>
<http://www.tei-c.org/index.xml>
<http://www.unicode.org/standard/WhatIsUnicode.html>
<http://www.w3.org/>
<http://www.w3.org/History/1989/proposal.html>
<http://www.w3.org/standards/>

La Terminologia

GIOVANNI ADAMO*

L'origine e il progresso delle nostre conoscenze
dipendono interamente dalla maniera
con la quale ci serviamo dei segni.
Condillac

Terminologia è voce registrata – fino agli ultimi anni del XX secolo e in tutta la produzione lessicografica europea – per fare riferimento all'*«insieme dei termini propri di una scienza, di un'arte, di un autore»*. Questa è senza dubbio l'accezione più diffusa, nonostante si sia in presenza di un termine composto con l'elemento formante *-(o)logia*, che esprime il significato di *studio sistematico, trattazione, teoria, dottrina*. Solo da alcuni anni i dizionari hanno accolto anche il riferimento più specialistico al *«settore degli studi linguistici che si occupa degli aspetti teorici della formazione e dell'uso sistematico dei termini nell'ambito di una scienza o di una disciplina»*. Peraltro, già nei primi decenni del XIX secolo, un repertorio dedicato al lessico tecnico e alle sue implicazioni etimo-filologiche definiva la *terminologia* come la *«dottrina dei termini o dei vocaboli propri di una data*

* Istituto per il Lessico Intellettuale Europeo e Storia delle Idee del Consiglio Nazionale delle Ricerche.

arte o scienza» (Marchi, 1828-1829)[1]. Altrettanto significativa è una segnalazione di (Rey, 1992):

> *L'uso moderno del concetto oggettivo [di terminologia] sembra determinarsi in Inghilterra. La definizione di William Whewell [epistemologo e moralista inglese, 1794-1866], nel 1837, attribuisce alla parola il suo valore scientifico, ancora limitato a un gruppo di scienze: 'sistema di termini usati nella descrizione degli oggetti della storia naturale'. La connessione dei concetti di 'sistema', di 'oggetto' e di 'scienza' a quello di 'termine' dà a questa definizione troppo dimenticata un'attualità sorprendente*[2].

Il periodo più ricco di fermenti determinanti per il concepimento della moderna teoria terminologica è, però, da collocare nei primi decenni del XX secolo, connotati da una forte espansione dell'attività industriale e da una vivace attività culturale. Sono gli anni in cui nascono e si affermano il Circolo di Vienna[3]

---

1 MARCO AURELIO MARCHI, *Dizionario tecnico-etimologico-filologico*, Milano, Giacomo Pirola, 2 voll., 1828-1829.

2 *«L'emploi moderne de la notion objective s'élabore, semble-t-il, en Angleterre. La définition de William Whewell, en 1837, donne au mot sa valeur scientifique, encore restreinte à un groupe de sciences: 'système des termes employés dans la description des objets de l'histoire naturelle'. L'articulation des notions de 'système', d''objet' et de 'science' à celle de 'terme' donne à cette définition trop oubliée une actualité surprenante»*. ALAIN REY, *La terminologie. Noms et notions*, ed. 2, Parigi, Presses Universitaires de France, pp. 6-7.

3 Il Circolo, attivo tra il 1922 e la fine degli anni Trenta, raccolse attorno a Moritz Schlick (filosofo e fisico) molti tra i più prestigiosi intellettuali e studiosi dell'epoca, di formazione culturale diversa. Tra essi: Rudolf Carnap (filosofo), Philipp Frank (fisico teorico), Kurt Gödel (logico), Hans Hahn (matematico), Otto Neurath (economista e sociologo), Friedrich Waismann (filosofo). Anche altri insigni studiosi, tra i quali Karl R. Popper, John Von Neumann, Carl G. Hempel, Federigo Enriques e Hans Rei-

e il Circolo linguistico di Praga[4]. Uno degli elementi di maggiore interesse consiste nell'ambizioso progetto del Circolo viennese di rifondare la conoscenza umana su basi esclusivamente logiche ed empiriche, attraverso un linguaggio unificato della scienza[5]. È difficile pensare che la formazione e il pensiero dell'ingegnere austriaco Eugen Wüster (1898-1977) non abbiano risentito di influssi tanto profondi e innovativi. Ma occorre attendere il 1979, anno in cui viene pubblicata postuma la sua *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*[6], per vedere definitivamente elaborati il nucleo

chenbach, ebbero contatti o parteciparono occasionalmente alle riunioni del Circolo, che si fece conoscere dall'opinione pubblica internazionale con un manifesto intitolato *Wissenschaftliche Weltauffassung. Der Wiener Kreis*, pubblicato da Hahn, Neurath e Carnap nel 1929. La produzione più significativa del Circolo fu affidata alle pagine del periodico «Erkenntnis» (nelle ultime annate «The Journal of Unified Science»), pubblicato tra il 1930 e il 1940, in collaborazione con un gruppo di studiosi berlinesi organizzato da Reichenbach.

4 La sua prima riunione si tenne il 6 ottobre 1926, sotto la presidenza di Vilem Mathesius. Tra i membri fondatori: Roman Jakobson, Nikolaj S. Trubeckoj e René Wellek. Da ricordare il grande impulso dato dal Circolo praghese agli studi di linguistica sincronica e descrittiva.

5 Cfr. *Foundations of the Unity of Science. Toward an International Encyclopedia of Unified Science*, Neurath O., Carnap R., Morris C., (Eds.), ed. 3, 2 voll., Chicago-London, The University of Chicago Press, 1971. Nei *Travaux du IX Congrès International de Philosophie. Congrès Descartes. L'Unité de la Science: la Méthode et les méthodes*, Bayer R.,(Ed.), Parigi, Hermann, 1937, sono raccolte, tra le altre, due comunicazioni di grande interesse per il nostro argomento: RUDOLF CARNAP, *Einheit der Wissenschaft durch Einheit der Sprache*, pp. 51-57; OTTO NEURATH, *Prognosen und Terminologie in Physik, Biologie, Soziologie*, pp. 77-85.

6 L'opera, pubblicata in due volumi a Vienna-New York per i tipi di Springer e subito tradotta in francese (*Introduction à la théorie générale de la terminologie et à la lexicographie terminologique*, Québec, Università di Laval, 1979), era stata anticipata da EUGEN WÜSTER, *Die allgemeine Terminologielehre, ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften*, in «Linguistics. An In-

teorico dello studio delle terminologie specialistiche, di cui lo stesso Wüster è il codificatore. Ma è opportuno sottolineare come egli parlasse ancora di *terminologische Lexikographie*, quasi a preservare il legame che, per alcuni secoli, si era instaurato in via pressoché esclusiva tra la prassi lessicografica, più o meno attenta – occorre riconoscerlo –, e le nomenclature dei settori specialistici.

Prospettive teoriche della terminologia

A partire dagli anni Settanta del XX secolo, in concomitanza con la riflessione teorica elaborata da Wüster, si è andata consolidando una *tripartizione di orientamenti e tendenze* che riflette prospettive e finalità diverse e che ha avuto conseguenze significative anche dal punto di vista del riconoscimento dello statuto scientifico-disciplinare della terminologia.

(Felber, 1984)[7] individua i *tre orientamenti*, considerandoli come indirizzati rispettivamente ai settori specialistici, alla filosofia, e alla linguistica. Il *primo*, caratterizzato dall'interdisciplinarità, è quello espresso nella *Teoria generale della terminologia* elaborata da Wüster. Vi si pongono al centro dell'attenzione: il concetto e le sue relazioni con le altre unità del medesimo siste-

ternational Review», n. 119, 1974, pp. 61-106. Ma già EUGEN WÜSTER, *Die vier Dimensionen der Terminologiearbeit*, in «Mitteilungsblatt für Dolmetscher und Übersetzer», vol. 15, n. 2, 1969, pp. 1-6, n. 5, pp. 1-5, riproponeva la relazione presentata a un Colloquio tenutosi presso l'Istituto per traduttori e interpreti di Germersheim (Università di Mainz) nell'ottobre dell'anno precedente. Vi si definivano le quattro dimensioni della prassi terminologica: il settore specialistico, le lingue, l'obiettivo (trattare documenti, utilizzare una terminologia, condurre ricerche su un campo concettuale), e il grado di astrazione.

7 HELMUT FELBER, *Terminology Manual*, Parigi, Unesco-Infoterm, 1984, p. 47.

ma concettuale, la corrispondenza tra concetto e termine, l'assegnazione dei termini ai concetti. Il *secondo* orientamento, piuttosto vicino al primo, preferisce porre l'accento sulle classificazioni dei concetti in base a categorie filosofiche. Esso si propone, quindi, di studiare gli aspetti connessi con le teorie della classificazione, che lo accomunano alle problematiche della documentazione (Dahlberg, 1974)[8]. Il *terzo*, orientato agli studi linguistici, si fonda sul principio che le terminologie, essendo sottocomponenti del lessico di un linguaggio speciale, rientrano nei sottocodici delle varie lingue.

Qualche anno più tardi, (Cabré, 1993)[9] riprende uno schema già proposto da Pierre Auger e delinea un quadro più complesso e ricco di sfumature, articolato anch'esso in *tre grandi tendenze*, che considerano la terminologia dal punto di vista del sistema linguistico, da quello della traduzione, e, infine, da quello della pianificazione linguistica. La *prima* tendenza, linguistico-terminologica, si caratterizza per la propensione alla standardizzazione concettuale e denominazionale ed è rappresentata da tre scuole, ritenute *classiche*: la scuola di Vienna, che ha come figura di riferimento Eugen Wüster, considera il lavoro terminologico in vista della normalizzazione di concetti e termini, e vede negli specialisti dei differenti settori i responsabili delle diverse terminologie; la scuola di Praga, rappresentata da Lubomír Drozd e ispirata agli studi di linguistica funzionale del noto Circolo linguistico, si occupa della descrizione strutturale e funzionale dei linguaggi settoriali, considerati come uno *stile professionale*; la scuola di Mosca (Lotte, 1981)[10], molto attenta alle posizioni di

---

8 Cfr. INGETRAUT DAHLBERG, *Grundlagen universaler Wissensordnung. Probleme und Möglichkeiten eines universalen Klassificationssystem des Wissens*, Monaco, Verlag Dokumentation Saur KG, 1974.

9 MARIA TERESA CABRÉ, *La terminología. Teoría, metodología, aplicaciones*, Barcellona, Editorial Antártida/Empúries, 1993, pp. 39-41.

10 DMITRIĬ SEMENOVICH LOTTE, *Principes d'établissement d'une terminolo-*

Wüster, si interessa alla normalizzazione di concetti e termini nella cornice dei problemi connessi con la situazione di plurilinguismo dell'ex Unione Sovietica. La *seconda* tendenza, orientata alla traduzione, è molto sviluppata nei Paesi francofoni ed è alla base dei lavori terminologici prodotti dai grandi organismi internazionali multilingui, con l'intento di stabilire equivalenze terminologiche tra le diverse lingue, per orientare i traduttori nel loro lavoro. La *terza* tendenza, quella *aménagiste* (secondo la denominazione coniata nel Québec), considera la terminologia all'interno di un più ampio processo di pianificazione linguistica[11]. La terminologia, ritenuta uno degli elementi fondamentali di ogni codice comunicativo, diviene strumento per la rivitalizzazione delle lingue minoritarie, ma anche per il rilancio della funzione comunicativa internazionale di lingue stabili e consolidate, che si trovano però a dover fronteggiare la concorrenza pervasiva delle lingue di Paesi in situazione di forte predominio culturale, economico o tecnologico. Per raggiungere questo obiettivo, si ritiene indispensabile un progetto d'azione sistematico e strategico, sostenuto da adeguati interventi legislativi.

Su posizioni diverse si trova (Sager, 1990)[12] che non riconosce alla terminologia lo statuto di disciplina autonoma, pur sottolineando l'importanza e il valore dell'attività terminologica e la necessità della sua presenza nei percorsi formativi.

---

gie scientifique et technique, in Textes choisis de terminologie, Rondeau G., Felber H. (a cura di), Québec, Girsterm, 1981, pp. 3-53.

11 Per una comprensione più agevole, il concetto di *pianificazione* può essere ritenuto equivalente a quello di *progettazione*. Un approfondimento di tale concetto si trova in CABRÉ, M.T., op. cit, 1993, p. 108: *«La normalizzazione di una lingua deve consistere in un processo pianificato che parta da una situazione esplicita, che si proponga alcuni obiettivi concreti da perseguire in un periodo di tempo determinato, che si basi su canali di diffusione e su risorse di 'disseminazione' adeguate e che disponga di una legislazione che favorisca questo processo di scambio»*.

12 JUAN C. SAGER, *A practical course in terminology processing*, Amsterdam-Philadelphia, John Benjamins, 1990, p. 1.

D'altro canto, Teresa Cabré affermava di recente la necessità di una rivisitazione dei fondamenti teorici della terminologia, sottolineando l'inadeguatezza alla realtà attuale dell'istanza riduzionista e uniformatrice, già presente nella teoria di Wüster e amplificata da alcune successive prese di posizione di studiosi a lui vicini. Inoltre, ribadendo l'esigenza di considerare più attentamente i riflessi sociali della comunicazione specialistica, (Cabré, 1998)[13] ha fornito nuovi contributi per la formulazione di una proposta teorica e ha evidenziato il ruolo e l'influsso dell'organizzazione della società in rapporto alle due dimensioni fondamentali della terminologia: la rappresentazione e la trasmissione delle conoscenze trattate.

Il quadro complessivo appare, dunque, piuttosto frammentato, fino a far pensare che il solo elemento di unitarietà possa consistere in una prassi di lavoro comune, più che in un'improbabile conciliazione delle differenti impostazioni teoriche. A questo si aggiungano alcuni tentativi di portare alle estreme conseguenze il ruolo di scienza e disciplina autonoma attribuito alla terminologia (Picht, 1996)[14], ruolo che Wüster sembra avere comunque molto temperato, elaborando le *Teorie speciali della terminologia*, con l'intento di andare incontro alle peculiarità di ciascuno dei diversi settori specialistici e di ogni singola espressione linguistica (Felber, 1984)[15] e dichiarando esplicitamente l'appartenenza della terminologia alla linguistica applicata (Wüster, 1979)[16].

---

13 Cfr. MARIA TERESA CABRÉ, *Elementos para una teoría de la terminología: hacia un paradigma alternativo*, in «El lenguaraz», vol. 1, n. 1, Buenos Aires, 1998, pp. 59-78.

14 HERIBERT PICHT, *En record d'E. Wüster. La multidisciplinarietat de la terminologia*, in Terminologia. Selecció de textos d'E. Wüster, Cabré M.T (a cura di), Barcellona, Università di Barcellona, Servei de la llengua catalana, 1996, pp. 253-287, in particolare il paragrafo *6. Resum i perspectives*, pp. 277-278.

15 FELBER, H., *op. cit.*, p. 47.

16 EUGEN WÜSTER, *Einführung in die allgemeine Terminologielehre und ter-*

Obiettivi della terminologia

La terminologia ha come finalità quelle di individuare e determinare le unità concettuali che costituiscono il sistema strutturato di conoscenze proprio di un settore specialistico e, successivamente, di occuparsi della denominazione di quelle unità concettuali mediante i termini. In altre parole, la terminologia si propone di esaminare e descrivere la struttura e l'organizzazione delle unità di conoscenza elaborate all'interno dei diversi settori del sapere specialistico: da quelli che, per il rigore metodologico e l'ambito di interesse più nettamente circoscritto, hanno raggiunto un livello di codificazione tradizionalmente consolidato e molto evoluto, come avviene nel caso delle scienze, a quelli delle tecnologie, delle arti e delle tecniche, ai settori di attività professionale, artigianale e pratica. Si tratta, per usare un'espressione di carattere onnicomprensivo, dei più diversi campi dell'esperienza umana, che tendono a distinguersi per un interesse omogeneo e condiviso e che sviluppano la necessità di esprimere le conoscenze e le esperienze maturate al loro interno attraverso un *linguaggio speciale*[17], che si diversifica via via dalla

minologische Lexikographie, Vienna, Springer, 1979, pp. 64-65.

[17] Si registrano anche altre espressioni affini (tecnoletto, linguaggio settoriale, linguaggio della scienza e della tecnica, linguaggio scientifico, linguaggio specialistico), che si è soliti ricondurre all'ambito dei sottocodici linguistici. Rispetto al codice di base, questi ultimi presentano tratti distintivi peculiari. Occorre tuttavia notare il favore crescente che incontra la posizione teorica di quanti considerano lo studio della lingua da un punto di vista unitario: *«La lengua es única, los llamados 'lenguajes de especialidad', o tecnolectos, no son más que especializaciones de determinados elementos de la única lengua general, y, por lo tanto, forman parte de ella. Por tal razón considero que es una inconsecuencia con los postulados más generales de la lingüística hablar de 'lengua especializada', 'lengua de especialidad' y expressiones similares. [...] El término es un signo lingüístico como cualquier unidad léxica de las áreas no especializadas, y los hablantes científicos y técnicos son tan hablantes como el resto. Las*

lingua dell'uso comune, producendo quello che (De Mauro, 1995) ha definito *«un uso speciale della lingua»*[18]. Quest'uso tende a essere codificato sulla base di tratti particolari che lo caratterizzano nei tre piani: formale, funzionale e del significato. In particolare, e soprattutto per i linguaggi scientifici, da un punto di vista *formale*, sono assai numerosi i termini formati mediante il ricorso a affissi e confissi, molti dei quali di origine greca e latina: risulta in tal modo favorita la comprensione e la circolazione interlinguistica dei termini. Su un altro piano, si riscontra un'inclinazione all'essenzialità dei moduli *sintattici*, riducendo l'uso dei tempi e dei modi verbali e privilegiando una struttura lineare e nominale, che raramente ricorre a costruzioni ipotattiche. Per quanto riguarda l'aspetto *semantico*, si cerca di stabilire un rapporto diretto e univoco tra termine e designato, evitando che i termini si carichino di uno spettro ampio di significati (*polisemia*), che vengano usati, cioè, per designare più oggetti o concetti diversi tra loro. Si osserva, inoltre, che il significato è trasmesso senza ricorrere alle situazioni proprie dell'espressività comunicativa[19]. Tali caratteristiche nascono fondamentalmente

---

diferencias entre 'término' y 'no término' son de tipo pragmático, no formal, funcional ni semántico. Esto hace que muchos de los postulados tradicionales de la terminología sean desmentidos, o al menos puestos en entredicho, por la práctica cotidiana», RODOLFO ALPÍZAR CASTILLO, *¿Cómo hacer un diccionario científico-técnico?*, Buenos Aires, Editorial Memphis, 1997, p. 8.

18 *«Quando un cospicuo numero di accezioni di parole diverse e, eventualmente, di parole tecniche e neologismi siano in nesso tra loro e usati da gruppi specifici di parlanti per trattare di argomenti determinati, nasce ciò che la linguistica storica e sociologica chiama una 'lingua speciale' o 'linguaggio speciale' (o 'settoriale') e che meglio si dirà un 'uso speciale della lingua'»*.
TULLIO DE MAURO, *Minisemantica. Dei linguaggi non verbali e delle lingue*, ed. 3, Roma-Bari, Laterza, 1995, p. 131.

19 Per un esame più completo delle caratteristiche dei linguaggi scientifici si veda MAURIZIO DARDANO, *I linguaggi scientifici*, in Storia della lingua ita-

dal bisogno di rappresentare e di esprimere – nel modo più chiaro, sintetico e preciso possibile – la struttura concettuale del settore specialistico: si pensi alle grandi classificazioni tassonomiche degli esseri viventi elaborate da Carl von Linné (1707-1778) e alla riforma della nomenclatura chimica, condotta sul finire del XVIII secolo da Guyton de Morveau, Lavoisier, Berthollet e Fourcroy. Nello stesso periodo, Diderot pensava a una *grammatica delle arti*, come fattore di regolazione sistematica delle terminologie e dei linguaggi delle scienze e delle tecniche, e nella sua voce *Encyclopédie* notava:

> *La conoscenza della lingua è il fondamento di tutte queste grandi speranze; esse rimarranno incerte, se la lingua non è stabilmente definita e trasmessa alla posterità in tutta la sua perfezione; e questo è lo scopo primario tra quelli di cui gli Enciclopedisti dovrebbero occuparsi profondamente. Noi ce ne siamo accorti troppo tardi; e tale inavvertenza ha proiettato un'imperfezione su tutta la nostra opera. Il versante della lingua è rimasto debole (dico della lingua, e non della Grammatica); e perciò questo dev'essere l'argomento principale in una voce in cui si esamina imparzialmente il proprio lavoro, e in cui si cercano i mezzi per correggerne i difetti*[20].

liana, Serianni L., Trifone P. (a cura di), vol. 2, Torino, Einaudi, 1994, pp. 497-551, in particolare alle pp. 497-498. Sui *contrassegni di qualità*, che rendono conto del carattere specifico dei linguaggi scientifici, si veda MAURIZIO DARDANO, *I linguaggi scientifici nell'italiano di oggi*, in La terminologia tecnica e scientifica. Attualità e prospettive, Ministero dell'Università e della Ricerca Scientifica e Tecnologica, Supplemento al Notiziario «Università e Ricerca», Roma, Istituto Poligrafico e Zecca dello Stato, 1996, p.12.

20 *«[…] la connaissance de la langue est le fondement de toutes ces grandes espérances; elles resteront incertaines, si la langue n'est fixée et transmise à la postérité dans toute sa perfection; et cet objet est le premier de ceux dont il convenait à des Encyclopédistes de s'occuper profondément.*

Quanto più elevato è il rigore metodologico e specificamente circoscritto l'ambito di interesse di un settore specialistico, come avviene nel caso delle scienze, tanto più le caratteristiche di quest'uso speciale della lingua si manifestano in modo evidente e peculiare. Al punto che alcuni studiosi arrivano a distinguere nettamente i linguaggi scientifici da quelli settoriali, osservando che *«le argomentazioni della scienza, i metodi di indagine, le nuove prospettive del pensiero, modificando il campo della percezione, creano nuove forme enunciative e linguistiche»* (Dardano, 1994)[21]. Anche altre ragioni contribuiscono a determinare il lento processo di diversificazione dei linguaggi specialistici dalla lingua d'uso: in primo luogo, la necessità di una comunicazione rapida e efficace, in grado di superare agevolmente le barriere delle espressioni linguistiche nazionali, ma anche il bisogno di trasmettere a nuovi cultori le nozioni che formano il patrimonio di conoscenze sviluppate nel settore al quale intendono dedicarsi.

Occorre, a questo proposito, almeno ricordare l'impegnativo tema della varietà dei discorsi scientifici, ripartiti tra lingua della teoria (*Theoriesprache*), lingua del laboratorio (*Werkstattsprache*) e lingua della divulgazione (*Verteilersprache*) (Pöckl, 1990)[22]. Considerata da un punto di vista terminologico, questa

---

Nous nous en sommes aperçus trop tard; et cette inadvertance a jeté de l'imperfection sur tout notre ouvrage. Le côté de la langue est resté faible (je dis de la langue, et non de la Grammaire); et par cette raison ce doit être le sujet principal, dans un article où l'on examine impartialement son travail, et où l'on cherche les moyens d'en corriger les défauts», *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers. Nouvelle impression en facsimilé de la première édition de 1751-1780*, vol. 5, Stuttgart-Bad Cannstatt, Friedrich Frommann Verlag (Günther Holzboog), 1966, p. 637r.

21 DARDANO, M., *op. cit*., 1994, p. 499.

22 WOLFGANG PÖCKL, *Französisch: Fachsprachen. Langues de spécialité*, in Lexikon der Romanistischen Linguistik (LRL), G. Holtus - M. Metzeltin - Ch. Schmitt (Eds.), Tübingen, Max Niemeyer Verlag, 1990, pp. 268-269.

diversificazione del discorso scientifico, che si basa sul pubblico dei destinatari e quindi sulla rilevazione del diverso grado di specializzazione dei testi, si riflette profondamente anche nell'uso dei termini. Per fare un solo esempio, si pensi alla denominazione dei prodotti chimici: il composto individuato e determinato dalla formula H_2O può essere anche denominato *monossido di diidrogeno* o, più semplicemente, *acqua*. Si è inoltre riscontrato che, quanto più è elevato il grado di specializzazione di un testo scientifico, tanto maggiore risulta la densità dei termini specialistici in esso presenti, proprio a motivo di bandire ogni possibile equivoco o incomprensione nella comunicazione.

Scienziati e tecnologi hanno come obiettivo primario l'avanzamento delle conoscenze e – nei limiti posti dagli interessi particolari e dalla concorrenza imprenditoriale – tendono a mettere in comune il frutto del proprio lavoro. Hanno bisogno di comunicare in modo efficace, rapido e preciso con i loro colleghi, prescindendo dalle limitazioni imposte dallo spazio e dal tempo. Proprio per questo sono portati a privilegiare una lingua di comunicazione internazionale che, a costo di comprimere le diversità culturali dei singoli, sia in grado di favorire le relazioni e gli scambi. Per essere conosciuti in ambito internazionale, scienziati e tecnologi devono far circolare le loro ricerche e i loro studi su riviste pubblicate in inglese, lingua che utilizzano abitualmente anche per presentare le loro relazioni ai congressi internazionali. Occorre, però, tenere ben presente che il fatto che la maggior parte della produzione scientifica si esprima in inglese può risultare fortemente penalizzante per lo sviluppo armonico delle singole lingue nazionali. Se uno scienziato non curasse l'adeguata circolazione dei termini specialistici del suo settore d'interesse anche nella propria lingua madre, ne determinerebbe un progressivo impoverimento che potrebbe arrivare fino alla completa atrofia lessicale dei settori maggiormente specializzati, nei quali opera un più ristretto numero di addetti. Ma anche i grandi settori specialistici più direttamente coinvolti nel fenomeno del-

la globalizzazione – si pensi all'informatica e alla telematica, alle nuove tecnologie in genere, all'economia e alla nuova economia – sono quelli che presentano nell'italiano contemporaneo il più alto tasso di prestiti integrali, di parole cioè originariamente estranee al sistema linguistico.

Metodo di lavoro della terminologia

Il metodo di lavoro proprio della terminologia è detto *onomasiologico* perché si propone di denominare le singole unità di conoscenza individuate in un settore specialistico. Accade di rado, però, che tale attività assuma anche un'effettiva funzione *onomaturgica*, ovvero che il terminologo sia chiamato a imporre un nome alle unità rilevate. Più abitualmente, il suo lavoro consiste in un'opera paziente di inventariazione e sistematizzazione delle conoscenze accumulate in un settore specialistico per classificarle mediante i termini che, denominandole, le rappresentano. Nella maggior parte dei casi, si tratta di espressioni linguistiche, semplici o complesse, ma possono essere anche segni, simboli, formule o elementi grafici. I termini sono convenzionalmente scelti per rappresentare – nella forma più precisa, concisa e adeguata – la definizione di ciascun nucleo della conoscenza elaborata in un determinato settore[23] e, al tempo stesso, essi costitui-

[23] *«Y es así porque en todas las representaciones los símbolos utilizados – igual que los términos, que son en definitiva representaciones – no son más que substituciones de la definición subyacente del concepto de que se trate; y las definiciones del concepto están establecidas con palabras»*.
BERTHA M. GUTIÉRREZ RODILLA, *La ciencia empieza en la palabra. Análisis e historia del lenguaje científico*, Barcellona, Ediciones Peninsula, 1998, p. 30.
L'argomento è ripreso poco dopo: *«Como ya hemos dicho, los términos son marcas sustitutivas de sus correspondientes definiciones, porque es la definición – completa y precisa – la única representación lingüística ade-*

scono il veicolo della comunicazione specialistica, sia tra gli specialisti dello stesso settore, sia nel trasferimento delle conoscenze acquisite, a livello divulgativo o di scambio interlinguistico (Cabré 1993)[24]. L'attività terminologica consiste, dunque, nell'ordinamento e nella classificazione del patrimonio cognitivo e comunicativo di un dominio. Pertanto, la rappresentazione concettuale e terminologica di un settore specialistico realizzata da un terminologo può costituire una via d'accesso privilegiata attraverso la quale anche il documentalista e il traduttore possono avvicinarsi al settore specialistico di cui debbono occuparsi per il proprio lavoro, per farsene un'idea complessiva o per approfondire aspetti di più specifico interesse.

Il prodotto dell'attività terminologica consiste nella rappresentazione linguistico-documentaria delle unità concettuali individuate mediante l'analisi di un corpus di fonti scritte, le più rappresentative e autorevoli del dominio specialistico esaminato: trattati e manuali, documentazione tecnica, riviste di settore e, se esistono, anche fonti legislative e normative (leggi, trattati, protocolli, convenzioni, norme tecniche nazionali e internazionali). Secondo la natura specifica di ciascun dominio, possono essere presi in considerazione anche glossari o repertori terminologici precedentemente realizzati e, talvolta, anche periodici divulgativi. L'intento è quello di prelevare materiale da un corpus che offra il panorama più fedele possibile delle conoscenze sviluppate nel settore, attraverso i termini effettivamente utilizzati dagli specialisti.

La rilevazione di forme diverse di lessicalizzazione dei termi-

cuada de un concepto. Reemplazan las definiciones – normalmente largas – por una expresión más económica pero de igual precisión; esto significa que no son intrínsecamente precisos, sino que sólo lo son si su definición lo es en sí misma».

Ivi, p. 32.

24 CABRÉ, M.T., *op. cit.*, 1993, p. 71.

ni che designano le medesime unità concettuali (come nei casi di varianti formali, sinonimi, varianti di scuola, varianti diatopiche o geografiche, impiego di termini abbreviati e acronimi) permette di fornire un quadro che consente di evidenziare il ventaglio di sfumature che si delineano all'interno di quello che si tenderebbe a considerare un linguaggio unitario, monolitico e prescrittivo. In esso emergono, in misura sempre più significativa, le variazioni che si producono all'interno di scuole di pensiero diverse, delle differenti aree regionali nelle quali si parlano lingue a diffusione intercontinentale, di consuetudini editoriali che impongono o consentono agli autori dei contributi anche l'uso di sigle estemporanee, spesso del tutto arbitrarie e prive di una reale efficacia che non sia quella dettata dal mero tentativo di risparmiare qualche riga di testo.

La registrazione delle varianti non comporta di per sé l'abbandono della vocazione originaria della terminologia, cioè la tendenza all'unificazione e alla normalizzazione dei termini, che è esigenza avvertita principalmente nella comunicazione internazionale, e che costituisce uno dei tratti che maggiormente connotano l'attività terminologica. A questa istanza, ancora oggi fortemente condivisa, si sono affiancati anche altri punti di vista, che tendono a considerare con maggiore attenzione le terminologie nel loro rapporto con le singole culture, e quindi con le lingue che le esprimono. Si va affermando, insomma, un orientamento maggiormente improntato al pragmatismo e alla funzionalità.

Occorre precisare che, anche a causa della sempre più frequente divulgazione attraverso la rete Internet, nella forma di banche dati, le risorse terminologiche disponibili per la consultazione non sempre sono presentate seguendo l'ordinamento in precedenza più consueto, ovvero quello detto *concettuale*, *tematico*, *metodico* o *sistematico*. A differenza dell'ordinamento alfabetico, privilegiato dai repertori lessicografici per la facilità di consultazione, l'ordinamento concettuale è senz'altro più coerente con l'impianto e le finalità delle raccolte terminologiche,

spesso destinate alla consultazione di utenti che conoscono approfonditamente la materia. Ne risulta, infatti, una sorta di quadro sinottico del settore, che permette di valutare l'approfondimento dell'indagine compiuta dal terminologo e facilita anche le operazioni di modifica e aggiornamento del repertorio stesso[25].

Questo tipo di ordinamento, ispirandosi a un criterio vagamente enciclopedico, consente anche a chi non è esperto di farsi un'idea d'insieme e di esplorare i nessi della rete concettuale del settore specialistico preso in esame.

Repertori terminografici

Come appena ricordato, la produzione terminografica risponde, nella gran parte dei casi, alle esigenze di un pubblico esperto, che ricorre ad essa per controlli e verifiche, ma anche per cercare gli equivalenti di un termine in altre lingue. I repertori terminografici sono, infatti, prevalentemente multilingui e si propongono di rispondere alle esigenze di una rapida comunicazione interlinguistica. Recentemente hanno avuto diffusione anche piccoli prontuari terminologici destinati alla divulgazione di terminologie specialistiche, perlopiù relative alle nuove tecnologie di grande impatto sociale. Si tratta spesso di repertori monolingui o corredati del solo equivalente nella lingua d'origine delle tecnologie stesse, ma anche di piccole raccolte terminologiche multilingui: l'essenzialità del contenuto è dettata dalle esigenze di una diffusione capillare e al tempo stesso economica, che tiene conto della rapida obsolescenza cui vanno soggette le tecnologie e delle conseguenti ripercussioni nel cambiamento delle terminologie. L'orientamento alla pianificazione linguistica e

[25] Konferenz der Übersetzungsdienste westeuropäischer Staaten (KÜWES), *Empfehlungen für die Terminologiearbeit*, Berna, Schweizerische Bundeskanzlei, 1990, cap. 6 *Klassification*.

l'attenzione ai criteri della socioterminologia hanno introdotto elementi innovativi rispetto alla tradizionale presentazione terminografica: la segnalazione delle varianti e la presenza dei contesti d'uso sono quelli di maggior rilievo.

La terminografia raccoglie e presenta i termini dei linguaggi specialistici e tende all'uniformazione dei concetti e all'unificazione delle denominazioni, per consentire una comunicazione più diretta e efficace. La norma che rappresenta ha quindi la funzione di modello prescrittivo, benché l'orientamento di tipo variazionista vada raccogliendo un favore sempre più diffuso.

Il criterio privilegiato è quello sincronico, che testimonia l'impegno a contribuire a una comunicazione sempre aggiornata; per questa ragione, le norme internazionali sono sottoposte a revisione periodica[26].

Il corpus terminografico è selezionato con criteri rigorosi di rappresentatività e generalmente si limita alla documentazione scritta, privilegiando le fonti più autorevoli.

I repertori tendono a escludere la polisemia: un termine si connota per la relazione di univocità stabilita con il suo designato e solo in casi piuttosto rari può avere sinonimi. È possibile però riscontrare casi di omonimia, cioè di termini identici che designano oggetti distinti, ma questo può accadere in settori specialistici diversi tra loro.

La frequenza d'uso di un termine non ha alcuna rilevanza ai fini dell'inclusione in un repertorio terminografico: è stato osservato che il grado di specializzazione di un termine è inversamente proporzionale alla sua frequenza d'uso (Müller, 1985)[27].

---

[26] *«Der Vorrang der Begriffe hat zwangsläufig dazu geführt, daß die terminologische Sprachbetrachtung synchronisch ist. Für die Terminologie ist das Wichtigste an einer Sprache das Begriffssystem, das ihr zugrunde liegt»*.
WÜSTER, E., *op. cit*., 1979, p. 2.

[27] BODO MÜLLER, *Le français d'aujourd'hui*, Elsass A. (traduzione di), ed. rivista e ampliata dall'autore, Parigi, Éditions Klincksieck, 1985, p. 188.

Le entrate coincidono con le unità di conoscenza elaborate in un settore specialistico e, quindi, sono rappresentate da termini, che possono essere semplici o composti, e talvolta accompagnati dall'indicazione della categoria grammaticale.

La definizione descrive le caratteristiche e le funzioni dell'oggetto o del concetto denominato dal termine.

La contestualizzazione dei termini non è obbligatoria. Quando è presente, appare improntata all'esigenza della verifica documentaria, testimoniando l'uso effettivo dei termini in fonti di diversa natura (ufficiale, divulgativa) e attestando le variazioni geografiche e culturali.

L'ordinamento terminografico è prevalentemente concettuale, perché destinato alla consultazione di utenti che conoscono approfonditamente la materia (Wüster, 1979)[28]. Questo tipo di ordinamento può rivelarsi utile nell'estendere controlli e verifiche a elementi contigui a quello da cui ha avuto origine la ricerca e fornisce, inoltre, una sorta di quadro sinottico di un dominio del quale può giovarsi anche chi, non esperto, volesse farsene un'idea d'insieme. Tuttavia, le finalità divulgative e l'uso delle tecnologie informatiche sono alla base dell'impiego dell'ordinamento alfabetico, che sembra incontrare una diffusione in costante crescita.

Documentazione della terminologia

L'attività terminologica si ispira a un criterio di natura sincronica, che presuppone una costante attualizzazione del patrimonio terminologico-concettuale raccolto per ciascun settore. Si pensi, per esempio, alle norme tecniche, nazionali e internazionali, istituzionalmente sottoposte a revisione e aggiornamento periodico, anche nella parte relativa alla terminologia e al vocabolario, allo

[28] WÜSTER, E., *op. cit.*, 1979, par. 9.42, pp. 126-127.

scopo di assicurare precisione e efficacia alla comunicazione in ciascun settore. Gli enti nazionali e internazionali preposti all'unificazione, alla normalizzazione e alla standardizzazione sono ormai dotati di siti web presso i quali è possibile consultare il catalogo delle norme prodotte per ogni settore, con le relative date di aggiornamento: si tratta di una fase indispensabile e preliminare a ogni lavoro di documentazione, di traduzione o di redazione tecnica.

Allo stesso modo, è opportuno conoscere l'attività svolta da molte associazioni scientifiche, professionali e di categoria, sul piano nazionale e internazionale. Spesso esse dispongono di centri di documentazione per il settore di cui si interessano. Talvolta, mettono a disposizione, anche attraverso la rete Internet, repertori bibliografici, indicazioni relative ai dizionari specialistici, alla regolamentazione legislativa e normativa del settore. E, in molti casi, i siti di queste istituzioni sono dotati di motori di ricerca, che permettono di compiere interrogazioni puntuali mediante l'impiego di termini specifici, oppure presentano glossari o liste selezionate di parole chiave, che possono rivelarsi di grande utilità per il documentalista e il traduttore.

Negli ultimi anni sono state costituite molte associazioni nazionali e internazionali di terminologia, con l'obiettivo di censire, incrementare e coordinare la produzione di risorse terminologiche nei vari settori specialistici. Esse forniscono un servizio informativo e di documentazione sui repertori terminologici esistenti, sia a stampa, sia disponibili per la consultazione in linea. Raccolgono anche informazioni bibliografiche sulla produzione teorica e applicativa nel settore della terminologia. Una segnalazione particolare merita l'iniziativa intrapresa dall'Associazione europea per la terminologia (Aet-Eaft)[29] per la costituzione di un Server europeo di informazione terminologica (Etis),[30] che si

29 <http://www.eaft-aet.net/it/indice/>.
30 <http://www.computing.surrey.ac.uk/ai/etis/>.

propone come centro unitario di raccolta e smistamento di informazioni in ambito europeo. Questa iniziativa intende soddisfare una delle esigenze oggi più fortemente avvertite, quella di ridurre l'alto tasso di frammentazione delle risorse terminologiche. Infatti, nonostante l'opera preziosa e efficace svolta dalle associazioni nazionali e dai centri di consulenza terminologica, non esiste ancora oggi un nucleo unitario di riferimento e di orientamento per chi voglia entrare in contatto con il mondo complesso e articolato della terminologia.

Nella stessa direzione sono stati rivolti anche gli sforzi di coloro che hanno contribuito alla formazione di vere e proprie *reti di lavoro* tra quanti operano nei più disparati settori dell'attività terminologica: *Nordterm*, un circuito che collega i Paesi del Nord Europa[31]; il *Rint*, Rete internazionale di neologia e terminologia tra i Paesi dell'area francofona; *Riterm*, Rete iberoamericana di terminologia tra i Paesi di espressione linguistica spagnola e portoghese[32]; e *Realiter*, la Rete panlatina di terminologia, che si pone come luogo di raccordo e di scambio tra i Paesi di lingua neolatina[33].

Ma la risorsa più preziosa, che spesso si rivela anche la più ardua da raggiungere e da consultare, è costituita dalle numerose banche di dati terminologici. Nel 1996, il Rint ha realizzato un *Inventaire des banques de terminologie*, disponibile su Internet e attualmente in corso di aggiornamento, che raccoglie schede informative su un gran numero di banche terminologiche, grandi e piccole, private e pubbliche. Ciascuna scheda mostra le finalità, la consistenza, il tipo e la quantità di dati raccolti, e offre notizie sulle formalità e le condizioni di accesso alla banca dati.

La letteratura tecnica che descrive le basi di dati terminologici seleziona all'interno di esse un gruppo ristretto, che viene abi-

31 <http://www.nordterm.net/>.
32 <http://www.riterm.net/>.
33 <http://www.realiter.net/>.

tualmente designato con l'appellativo di *grandi banche dati*. Si tratta di banche dati multilingui, le più importanti e consistenti, le prime a essere implementate, tra la fine degli anni Sessanta e l'inizio dei Settanta: *Eurodicautom*, oggi sostituita da *IATE*[34], *Termium*[35], *le Grand dictionnaire terminologique*[36], *Lexis* e *Team* (Gagnon, 1994)[37]. Queste prime banche dati erano state ideate con lo scopo di superare la rapida obsolescenza dei dizionari e repertori a stampa in rapporto alla continua evoluzione delle terminologie, e per approntare uno strumento che facilitasse il lavoro di documentalisti, traduttori e interpreti. L'impianto di queste banche dati richiedeva però l'investimento di risorse considerevoli, dal punto di vista finanziario, informatico e del lavoro umano, senza tenere conto della complessità e lentezza delle operazioni di aggiornamento, o della necessità di una rete capillare per la diffusione e la consultazione dei dati: soltanto da alcuni anni, grazie a Internet, quest'ultimo aspetto ha trovato una positiva soluzione.

Il clima di sfiducia e scetticismo che si era prodotto attorno a queste grandi banche dati, nella seconda metà degli anni Settanta e nei primi anni Ottanta, ne aveva determinato un lento e progressivo abbandono, culminato con la diffusione dei personal computer, che hanno contribuito allo sviluppo di banche dati di dimensioni più ridotte, e finalizzate alla raccolta di terminologie di settori particolari. Presto, però, si è dovuto constatare che la nuova soluzione costringeva a reduplicare il lavoro necessario per costruire piccole banche dati settoriali ad uso personale da parte dei vari professionisti impegnati nel medesimo settore di attività. In questo modo finiva per essere incentivata anche la

34 <http://iate.europa.eu/iatediff/switchLang.do?success=mainPage&lang=it>.
35 <http://www.btb.termiumplus.gc.ca/site/termium.php?lang=fra&cont=00>.
36 <http://gdt.oqlf.gouv.qc.ca/>.
37 RENÉ GAGNON, *Les grandes banques de terminologie*, in «Meta: journal des traducteurs», vol. 39, n. 3, 1994, pp. 498-499.

mancanza di omogeneità delle terminologie, troppo spesso affidate alla sensibilità e all'intuizione dei singoli, e prive di ogni forma di controllo e di validazione. Di conseguenza, nell'ultimo decennio si è registrato un forte impulso a varie forme di collaborazione: associazioni, centri di consulenza e di servizio. La contemporanea diffusione di nuove tecnologie per la distribuzione e comunicazione dei dati, il cd-rom e le reti informatiche, ha riportato alla ribalta l'utilità delle grandi banche dati terminologiche centralizzate.

Bibliografia

Adamo G., (a cura di), *Ricerca e terminologia tecnico-scientifica: Atti della giornata di studio (Roma, 27 novembre 1992)*, in «Lexicon philosophicum. Quaderni di terminologia filosofica e storia delle idee», vol. 7, Firenze, Leo S. Olschki, 1994

ADAMO, G., *Développement harmonisé et distribution des ressources terminologiques: le Réseau panlatin de terminologie - Realiter*, in «Terminologies nouvelles», n. 14, dicembre 1995, pp. 77-81

ADAMO, G., *Tra lessicologia e terminologia*, in «Lexicon philosophicum. Quaderni di terminologia filosofica e storia delle idee», Lamarra A., Palaia R. (a cura di), vol. 10, Firenze, Leo S. Olschki, 1999, pp. 1-17

ALPÍZAR CASTILLO, R., *¿Cómo hacer un diccionario científico-técnico?*, Buenos Aires, Editorial Memphis, 1997

ARNTZ, R., PICHT, H., *Einführung in die Terminologiearbeit*, ed. 3, Hildesheim, Georg Olms, 1995

AUGER, P., *Implantabilité et acceptabilité terminologiques: les aspects linguistiques d'une socioterminologie de la langue du travail*, in Implantation des termes officiels. Actes du séminaire, Rouen, 6-8 dicembre 1993, in «Terminologies nouvelles», n. 12, 1994, pp. 47-57

Beccaria G.L. (a cura di), *Linguaggi settoriali e lingua comune, I linguaggi settoriali in Italia*, ed. 4, Torino, Bompiani, 1983

CABRÉ, M.T., *La terminología. Teoría, metodología, aplicaciones*, Barcellona, Editorial Antártida/Empúries, 1993

CABRÉ, M.T., *On diversity and terminology*, in «Terminology», vol. 2, n. 1, 1995, pp. 1-16

CABRÉ, M.T., *Elementos para una teoría de la terminología: hacia un paradigma alternativo*, in «El lenguaraz», vol. 1, n. 1, Buenos Aires, 1998, pp. 59-78

CABRÉ, M.T., *La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*, Barcellona, IULA-Università Pompeu Fabra, 1999

CARNAP, R., *Einheit der Wissenschaft durch Einheit der Sprache*, in Travaux du IX Congrès International de Philosophie. Congrès Descartes. L'Unité de la Science: la Méthode et les méthodes, Bayer R.,(Ed.), Parigi, Hermann, 1937, pp. 51-57

DAHLBERG, I., *Grundlagen universaler Wissensordnung. Probleme und Möglichkeiten eines universalen Klassificationssystem des Wissens*, Monaco, Verlag Dokumentation Saur KG, 1974

DARDANO, M., *I linguaggi scientifici*, in Storia della lingua italiana, Serianni L., Trifone P. (a cura di), vol. 2, Torino, Einaudi, 1994, pp. 497-551

DARDANO, M., *I linguaggi scientifici nell'italiano di oggi*, in La terminologia tecnica e scientifica. Attualità e prospettive, Ministero dell'Università e della Ricerca Scientifica e Tecnologica, Supplemento al Notiziario «Università e Ricerca», Roma, Istituto Poligrafico e Zecca dello Stato, 1996, pp. 11-20

De Mauro T. (a cura di), *Studi sul trattamento linguistico dell'informazione scientifica*, Roma, Bulzoni Editore, 1994

DE MAURO, T., *Minisemantica. Dei linguaggi non verbali e delle lingue*, ed. 3, Roma-Bari, Laterza, 1995

Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers. Nouvelle impression en facsimilé de la première édition de 1751-1780, vol. 5, Stuttgart-Bad Cannstatt, Friedrich Frommann Verlag (Günther Holzboog), 1966

FELBER, H., *Terminology Manual*, Parigi, Unesco-Infoterm, 1984

Foundations of the Unity of Science. Toward an International Encyclopedia of Unified Science, Neurath O., Carnap R., Morris C., (Eds.), ed. 3, 2 voll., Chicago-London, The University of Chicago Press, 1971

GAGNON, R., *Les grandes banques de terminologie*, in «Meta: journal des traducteurs», vol. 39, n. 3, 1994, pp. 498-520

GUTIÉRREZ RODILLA, B.M., *La ciencia empieza en la palabra. Análisis e historia del lenguaje científico*, Barcellona, Ediciones Peninsula, 1998

HOHNHOLD, I., *Übersetzungorientierte Terminologiearbeit. Eine Grundlegung für Praktiker*, Stuttgart, InTra, 1990

Konferenz der Übersetzungsdienste westeuropäischer Staaten (KÜWES), *Empfehlungen für die Terminologiearbeit*, Bern, Schweizerische Bundeskanzlei, 1990

LOTTE, D.S., *Principes d'établissement d'une terminologie scientifique et technique*, in Textes choisis de terminologie, Rondeau G., Felber H. (a cura di), Québec, Girsterm, 1981, pp. 3-53

Magris M., Musacchio M.T., Rega L., Scarpa, F. (a cura di), *Manuale di terminologia. Aspetti teorici, metodologici e applicativi*, Milano, Hoepli, 2002

MARCHI, M.A., *Dizionario tecnico-etimologico-filologico*, Milano, Giacomo Pirola, 2 voll., 1828-1829

MÜLLER, B., *Le français d'aujourd'hui*, Elsass A. (traduzione di), ed. rivista e ampliata dall'autore, Parigi, Éditions Klincksieck, 1985

NENCIONI, G., *Linguistica e terminologia tecnico-scientifica*, in Ricerca e terminologia tecnico-scientifica: Atti della giornata di studio (Roma, 27 novembre 1992), Adamo G., (a cura di), in «Lexicon philosophicum. Quaderni di terminologia filosofica e storia delle idee», vol. 7, Firenze, Leo S. Olschki, 1994, pp. 5-12

NEURATH, O., *Prognosen und Terminologie in Physik, Biologie, Soziologie*, in Travaux du IX Congrès International de Philosophie. Congrès Descartes. L'Unité de la Science: la Méthode et les méthodes, Bayer R., (Ed.), Parigi, Hermann, 1937, pp. 77-85

PICHT, H., *En record d'E. Wüster. La multidisciplinarietat de la terminologia*, in Terminologia. Selecció de textos d'E. Wüster, Cabré M.T (a cura di), Barcellona, Università di Barcellona, Servei de la llengua catalana, 1996, pp. 253-287

PÖCKL, W., *Französisch: Fachsprachen. Langues de spécialité*, in Lexikon der Romanistischen Linguistik (LRL), G. Holtus - M. Metzeltin - Ch. Schmitt (Eds.), Tübingen, Max Niemeyer Verlag, 1990, pp. 267-282

PUCCI, C.R., *Gli standard internazionali ISO nel settore documentario*, in «AIDA Informazioni», vol. 13. n. 1, gennaio-marzo 1995, pp. 29-34; anche in: *Raccolta delle Pubblicazioni FUB 1995* (0B04095), Roma, Fondazione Ugo Bordoni, 1995, pp. 25-31

REY, A., *La terminologie. Noms et notions*, ed. 2, Parigi, Presses Universitaires de France, 1992

REY, A., *Essays on terminology*, Sager J.C. (traduzione di) (Ed.), Amsterdam-Philadelphia, John Benjamins, 1995

RONDEAU, G., *Introduction à la terminologie*, Montréal, Centre éducatif et culturel, 1981

Rousseau L.J. (a cura di), *Actes de la table ronde sur les banques de terminologie tenue à Québec les 18 et 19 janvier 1996*, in «Terminologies nouvelles», n. 15, 1996, pp. 3-163.

SAGER, J.C., *A practical course in terminology processing*, Amsterdam-Philadelphia, John Benjamins, 1990

Travaux du IX Congrès International de Philosophie. Congrès Descartes. L'Unité de la Science: la Méthode et les méthodes, Bayer R.,(Ed.), Parigi, Hermann, 1937

VAN CAMPENHOUDT, M., *Abrégé de terminologie multilingue*, 1997 <http://www.termisti.refer.org/theoweb1.htm>
WÜSTER, E., *Die vier Dimensionen der Terminologiearbeit*, in «Mitteilungsblatt für Dolmetscher und Übersetzer», vol. 15, n. 2, 1969, pp. 1-6, n. 5, pp. 1-5
WÜSTER, E., *Die allgemeine Terminologielehre, ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften*, in «Linguistics. An International Review», n. 119, 1974, pp. 61-106
WÜSTER, E., *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*, Vienna, Springer, 1979

Sitografia

<http://www.eaft-aet.net/it/indice/>
<http://www.computing.surrey.ac.uk/ai/etis/>
<http://www.nordterm.net/>
<http://www.riterm.net/>
<http://www.realiter.net/>
<http://iate.europa.eu/iatediff/switchLang.do?success=mainPage&lang=it>
<http://www.btb.termiumplus.gc.ca/site/termium.php?lang=fra&cont=00>
<http://gdt.oqlf.gouv.qc.ca/>

Estrazione Terminologica Automatica e Indicizzazione: Scenari Applicativi, Problemi e Possibili Soluzioni

SIMONETTA MONTEMAGNI*

1. Introduzione

Il ricorso a metodi e tecniche di estrazione automatica di terminologia[1] settoriale da corpora di dominio, ovvero da insiemi di documenti relativi a uno specifico settore della conoscenza, rappresenta una sempre più diffusa pratica di supporto al processo di indicizzazione di collezioni documentali, inteso come l'operazione volta all'individuazione delle voci indice che ne costituiscono il contenuto concettuale. L'obiettivo di questo contributo è una rivisitazione critica di esperienze condotte all'interno di diversi scenari applicativi in cui i risultati del processo di estrazione automatica di terminologia sono utilizzati per la costruzione di vocabolari controllati o di thesauri[2] sulla base dei quali è condotto il processo di indicizzazione.

* Istituto di Linguistica Computazionale del Consiglio Nazionale delle Ricerche.

1 Per estrazione terminologica si intende l'identificazione e il recupero da una collezione documentale di termini che sono ritenuti significativi rispetto al dominio al quale i documenti si riferiscono. Tale operazione può essere eseguita manualmente, semi-automaticamente o automaticamente: in quanto segue, ci concentreremo sulle modalità automatica e semi-automatica di estrazione terminologica.

2 Nel caso del thesaurus, il vocabolario controllato si arricchisce di relazioni semantiche tra i termini identificati. In particolare si distingue tra rela-

Nella sua forma più semplice, un vocabolario controllato è costituito dal lessico che rappresenta un sapere specialistico, per esempio un elenco dei termini specifici di una disciplina (arte, medicina, economia, ecc.). Un vocabolario controllato di questo tipo può essere costruito manualmente da esperti di dominio, oppure essere acquisito in modo semi-automatico facendo uso di metodi e tecniche di trattamento automatico del linguaggio. Nel secondo caso, si va dalla soluzione più elementare di scartare dal vocabolario sottostante a una collezione di testi, assunta come rappresentativa delle conoscenze relative a uno specifico settore del sapere, le cosiddette *stop-words* (ovvero, parole semanticamente *vuote* come articoli, preposizioni, pronomi ecc.) fino al ricorso a tecniche avanzate di filtraggio dei termini rilevanti, ad esempio di estrazione terminologica.

Oltre agli inevitabili costi di costruzione di un vocabolario controllato, l'indicizzazione condotta in relazione a risorse costruite manualmente presenta numerosi problemi. Ad esempio, non è garantito l'allineamento tra il vocabolario controllato e la terminologia usata nei testi per convogliare i contenuti, con potenziali e non indifferenti ripercussioni a livello dell'indicizzazione se condotta in modo automatico. Un'altra difficoltà è connessa con la valutazione della terminologia complessa, ovvero costituita da sequenze di più parole: su che base un esperto decide se una sequenza di parole rappresenta un termine complesso oppure se il contenuto sottostante vada ricondotto alle singole parole che lo compongono? Nell'intuizione dell'esperto, per quanto radicata nella sua conoscenza approfondita del dominio, vi è inevitabilmente un margine di soggettività.

Consideriamo ora il caso dell'indicizzazione condotta in rela-

zioni di equivalenza, per la gestione dei sinonimi, delle varianti, ecc., gerarchiche, per correlare concetti legati da un rapporto genere-specie o parte-tutto, e associative, per definire i restanti tipi di relazioni che possono sussistere tra due o più concetti.

zione a vocabolari controllati costruiti in modo automatico con filtraggio di *stop-words*. Come (Chung, Nation, 2004) osservano, *«it seems that even after eliminating the stop words, the most frequent words from a specialized corpus are not all true terms but include many general words used across a wide range of subjects»*[3]. In questo caso, non tutte le unità di indicizzazione sono rilevanti rispetto al dominio. Un ulteriore problema connesso con questo tipo di approccio riguarda la composizione del vocabolario controllato che contiene soltanto termini singoli o *unità terminologiche monorematiche*, composte da un'unica parola. Ciò è in contrasto con quanto sappiamo della terminologia specialistica, che è prevalentemente costituita da termini complessi, o *unità terminologiche polirematiche* costituite da sequenze di più parole: si vedano in proposito (Jackendoff, 1997)[4], (Nakagawa, Mori, 2003)[5] e (Chung, Nation, 2004)[6].

L'ultima opzione è costituita dall'indicizzazione condotta in relazione a vocabolari controllati costruiti in modo semi-automatico mediante tecniche di estrazione di terminologia di dominio. In linea di principio, oggi questa rappresenta la soluzione ottimale che supera i limiti rilevati in relazione agli altri approcci: l'allineamento con i testi è garantito così come l'inclusione all'interno del vocabolario controllato di terminologia polirematica selezionata su base statistica; inoltre, il vocabolario così acquisito dovrebbe contenere solo i termini rilevanti per il dominio di indagine. Questa evoluzione è ben delineata da (Manning et

---

3 TERESA MIHWA CHUNG, PAUL NATION, *Identifying technical vocabulary*, in «System», vol. 32, 2004, p.259.

4 Cfr. RAY JACKENDOFF, *Twistin' the night away*, in «Language», vol. 73, 1997, pp. 534-559.

5 Cfr. HIROSHI NAKAGAWA, TATSUNORI MORI, *Automatic Term Recognition based on Statistics of Compound Nouns and their Components*, in «Terminology», vol. 9, n. 2, 2003, pp. 201-209.

6 Cfr. CHUNG, T.M., NATION, P., *op. cit.*

alii, 2008) nel loro libro sul recupero di informazioni da testi, che affermano:

> *The general trend in Information Retrieval systems over time has been from standard use of quite large stop lists (200–300 terms) to very small stop lists (7–12 terms) to no stop list whatsoever. Web search engines generally do not use stop lists. Some of the design of modern IR systems has focused precisely on how we can exploit the statistics of language so as to be able to cope with common words in better ways*[7].

Il ricorso a tecniche di estrazione automatica di terminologia di dominio da collezioni documentali si presenta dunque come la direzione da seguire nella selezione dei termini di indicizzazione da includere in un vocabolario controllato: ciò può riguardare la costruzione ex novo di un vocabolario controllato, oppure la sua estensione a partire da una base compilata in modo manuale, o l'aggiornamento di una versione precedente.

Se l'utilizzo di tecniche di estrazione terminologica appare assodato, rimane da valutare se le tecniche correnti siano sempre adeguate ed efficaci per il trattamento di linguaggi settoriali di diversi domini del sapere. Per quanto gli obiettivi dell'indicizzazione e dell'estrazione terminologica coincidano parzialmente, vi sono importanti differenze che possono rendere il risultato dell'estrazione terminologica non del tutto adeguato ai fini del processo di indicizzazione. Se l'obiettivo dell'indicizzazione è quello di trovare termini in grado di discriminare un documento da un altro, quello dell'estrazione di terminologia è l'identificazione di termini settoriali che designano concetti di un dominio specifico: ne consegue che un termine di indicizzazione può non

[7] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008, p. 27.

costituire un termine di dominio così come, nel caso di alcune collezioni documentali, i termini settoriali possono non rappresentare unità di indicizzazione utili.

Partendo dall'analisi dei risultati ottenuti con sistemi di estrazione terminologica in diversi scenari applicativi in relazione a diversi domini del sapere, il presente contributo indaga questi interrogativi e illustra alcune soluzioni avanzate per il superamento – o almeno il ridimensionamento – delle criticità rilevate.

L'articolo è organizzato come segue. Dopo un inquadramento della nozione di linguaggio settoriale con particolare riguardo al rapporto con la lingua comune e sue variazioni d'uso (sezione 2) segue una breve rassegna delle tecniche correnti per l'estrazione automatica di terminologia di dominio da collezioni documentali (sezione 3). Una volta inquadrata la tipologia di problemi che si trovano a fronteggiare i sistemi di estrazione terminologica in relazione a collezioni documentali rappresentative di diverse varietà di linguaggi settoriali (sezione 4), la seconda parte di questo contributo discute possibili soluzioni messe a punto per il superamento dei problemi enucleati e i risultati raggiunti all'interno di diversi scenari applicativi (sezione 5).

2. Il linguaggio della comunicazione specialistica: rapporto con la lingua comune e varietà d'uso

Nella ricerca linguistica italiana si osserva un alto grado di variazione nella designazione della nozione di *language for special purposes*: le denominazioni spaziano da *lingue speciali*, *lingue per scopi speciali*, *lingue di specializzazione*, *linguaggi settoriali*, *micro lingue*, *sottocodici*, *linguaggi specialistici*, e sono spesso usate con accezioni non del tutto equivalenti. Ad esempio, (Sobrero, 1993)[8] distingue tra *lingue specialistiche*, caratterizza-

[8] Cfr. ALBERTO SOBRERO, *Lingue speciali*, in Introduzione all'italiano con-

te da un alto grado di specializzazione come il linguaggio della medicina, della fisica, dell'informatica o del diritto, e *lingue settoriali*, che mostrano un minor grado di specializzazione come quella della pubblicità, della politica o della burocrazia: la differenza tra i due tipi si colloca principalmente a livello del lessico (regole di formazione delle parole, quantità di terminologia specialistica all'interno del vocabolario del testo). (Dardano, 1987)[9], invece, opta per un'unica classe dei cosiddetti *linguaggi settoriali* all'interno della quale distingue tra linguaggi *forti*, altamente organizzati e stabili dal punto di vista lessicale (come il linguaggio della matematica), e linguaggi *deboli*, caratterizzati da organizzazioni lessicali meno strutturate, come il linguaggio giuridico. Se da un lato questo estremo grado di variabilità nella denominazione di tale nozione rappresenta senza dubbio un indizio della mancanza, nella terminologia linguistica italiana odierna, di una definizione univoca della nozione di lingua utilizzata nella comunicazione specialistica, dall'altro mette in evidenza che la tipologia dei linguaggi usati all'interno dei domini del sapere è varia. In quanto segue, ci riferiremo a questa nozione con il termine di *linguaggi settoriali*.

Ai fini del presente contributo, due sono gli aspetti rilevanti del dibattito sui linguaggi settoriali: il rapporto tra questi e la lingua comune e l'esistenza di varietà d'uso all'interno di uno stesso linguaggio settoriale. Si tratta di due questioni parzialmente correlate, in quanto il rapporto con la lingua comune diventa un fattore che contribuisce significativamente all'identificazione di diverse varietà d'uso all'interno dello stesso linguaggio settoriale. Questi due aspetti sono ampiamente dibattuti nella letteratura

temporaneo. La variazione e gli usi, Sobrero A. (a cura di), Roma-Bari, Laterza, 1993, pp. 237-277.

9 Cfr. MAURIZIO DARDANO, *Linguaggi settoriali e processi di riformulazione*, in Parallela 3. Linguistica contrastiva / Linguaggi settoriali / Sintassi generativa, Dressler W. et alii (a cura di), Tubinga, Narr, 1987, pp. 134-145.

linguistica ma, come vedremo in seguito, non sono stati affrontati in modo sistematico nella definizione dei correnti metodi e tecniche di estrazione automatica di terminologia settoriale da corpora di dominio.

Il rapporto tra la lingua comune e i diversi linguaggi settoriali è stato ampiamente dibattuto nella letteratura italiana e internazionale[10]. Si tratta di una distinzione che sfugge a una caratterizzazione univoca. (Varantola, 1986)[11] osserva al riguardo quanto segue: *«Definitions of LSP [Language for Special Purposes] or SL [Specialized Language] versus GL [General Language] abound; none is universally applicable, for obvious reasons. Basically we are dealing with two intuitively correct assumptions that are good as working concepts but which resist a clear-cut definition and delimitation»*. Piuttosto che mirare a definire un confine che separi nettamente la lingua comune dal linguaggio settoriale, si è andata affermando l'idea che i due tipi di linguaggio rappresentano i due poli di un continuum che si estende dalla lingua comune ai linguaggi settoriali e caratterizzato da una gamma di livelli intermedi[12]. I due estremi di questo continuum presentano divergenze significative dal punto di vista linguistico, pragmatico e funzionale (Cabré, 1999)[13]: per quanto riguarda il

[10] Per una rassegna del dibattito sul problema si rinvia a: sul versante nazionale italiano, Cfr. STEFANIA CAVAGNOLI, *La comunicazione specialistica*, Roma, Carocci, 2007; a livello internazionale, Cfr. MARIA TERESA CABRÉ, *Terminology: Theory, Methods, and Applications*, Amsterdam, John Benjamins, 1999.

[11] KRISTA VARANTOLA, *Special Language and General Language: Linguistic and Didactic Aspects*, in «Unesco ALSED-LSP Newsletter», vol. 9, n. 2, dicembre 1986, p.10.

[12] Cfr. in proposito, tra gli altri: GUY RONDEAU, JUAN SAGER, *Introduction à la terminologie*, ed. 2, Chicoutimi, Gatan Morin, 1984; VARANTOLA, K., *op. cit.*; GIANFRANCO PORCELLI, *Principi di Glottodidattica*, Brescia, La Scuola, 1994.

[13] Cfr. CABRÉ, M.T., *op. cit.*

piano linguistico, i tratti divergenti spaziano tra i diversi livelli di descrizione linguistica, da quello lessicale, per il quale si registrano le differenze più significative, a quelli morfologico e sintattico (ad esempio, vi sono costruzioni sintattiche o formazioni morfologiche che sono parte del linguaggio comune ma non si registrano, se non in modo del tutto sporadico, all'interno dei linguaggi settoriali).

Nella transizione dalla lingua comune al linguaggio settoriale si osservano varietà linguistiche intermedie. Questa prospettiva sul rapporto tra lingua comune e linguaggi settoriali introduce il secondo problema che intendiamo affrontare in questa sede, ovvero che all'interno della comunicazione specialistica si osservano dimensioni di variazione che danno luogo a diversi tipi di linguaggi settoriali, anche all'interno dello stesso dominio del sapere. La Figura 1 riporta il grafico proposto da (Rondeau, 1983)[14] e riprodotto in (Cabré, 1999)[15] che visualizza il complesso rapporto che lega lingua comune (LC) e linguaggi settoriali (LS) di cui si distinguono diverse varietà a seconda del grado di specializzazione.

Nella figura ciascun settore corrispondente a un linguaggio settoriale presenta variazioni di tipo verticale, corrispondenti a diversi livelli di comunicazione, che vanno da un livello altamente specialistico e specializzato, in cui gli attori della comunicazione sono esclusivamente esperti di dominio, a livelli comunicativi più vicini all'utente comune come quello della divulgazione o della comunicazione didattica, in cui gli attori sono esperti e non esperti (nel caso della divulgazione) o futuri esperti (nella comunicazione di tipo didattico). Questo tipo di variazione verticale all'interno della comunicazione specialistica è vi-

[14] GUY RONDEAU, *Introduction à la terminologie*, Québec, Gaëtan Morin éditeur, 1983.

[15] CABRÉ, M.T., *op. cit.*, p. 69.

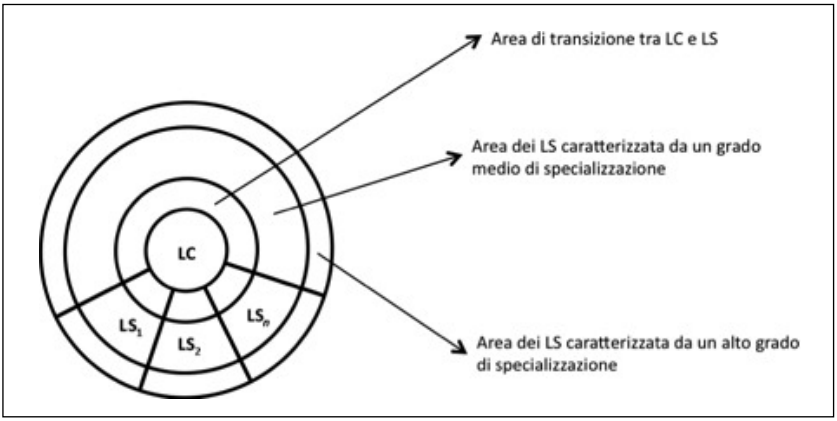


Figura 1. Rapporto tra lingua comune e linguaggi settoriali.

sta da (Lavinio, 2004)[16] come una variazione di tipo diastratico in quanto legata al livello socio-culturale degli interlocutori. Questa variazione di natura verticale si intreccia e sovrappone parzialmente con una variazione di tipo diafasico, ovvero legata alle modifiche che avvengono in una situazione comunicativa, sulla base del contesto, degli interlocutori, degli scopi della comunicazione e connessa ai diversi registri della lingua: a seconda della situazione all'interno della quale si colloca la comunicazione specialistica e delle sue finalità, si possono riconoscere sottovarietà di uno stesso linguaggio settoriale.

Sulla base di quanto detto finora, appare difficile marcare confini netti nella descrizione dei linguaggi settoriali, sia quando si vada a rapportarli alla lingua comune, sia quando se ne vadano a identificare sottovarietà definite sulla base della situazione comunicativa all'interno della quale sono usati. Ciò nonostante, i linguaggi settoriali rappresentano un'unica varietà che si diffe-

16 Cfr. CRISTINA LAVINIO, *Comunicazione e linguaggi disciplinari. Per un'educazione linguistica trasversale*, Roma, Carocci, 2004.

renzia nettamente dalla lingua comune: come afferma (Cabré, 1999), *«from a theoretical and methodological standpoint, it is important to establish the concept of special language in the singular»*[17]. Tuttavia, continuando con Cabré, la nozione astratta di linguaggio settoriale può essere suddivisa in sottovarietà a seconda della situazione comunicativa all'interno della quale sono usate, con importanti ripercussioni a livello dell'uso terminologico. A questo punto, la nozione unitaria di linguaggio settoriale si frammenta: a seconda che si tratti di comunicazione legata all'elaborazione del sapere o all'apprendimento o alla sua applicazione o alla sua divulgazione, il tipo di terminologia usata per convogliare gli stessi contenuti varierà in modo significativo, presentando diverse *miscele* di terminologia specialistica e lessico comune.

Finora, si è parlato di linguaggi settoriali e relative sottovarietà in termini molto astratti. Tuttavia, in questa sede il nostro interesse è legato al fatto che tali linguaggi si manifestano nei testi specialistici, che contengono – oltre agli elementi specialistici – quelli della lingua comune. Quella di linguaggio settoriale è un'astrazione costruita a partire dai testi. E' con i testi che ci confronteremo nel prosieguo di questo articolo, in quanto costituiscono il punto di partenza del processo di estrazione terminologica.

3. Estrazione automatica di terminologia specialistica da corpora di dominio

Sempre più centrali per lo sviluppo di applicazioni reali di gestione della conoscenza (o Knowledge Management), i sistemi di estrazione automatica di terminologia sono finalizzati all'identificazione e all'estrazione di unità terminologiche monorematiche (come *accordo*, *produttore* o *presidente*) e polirematiche (co-

[17] CABRÉ, M.T., *op. cit.*, p.76.

me *procedimento amministrativo*, *Ministro dell'ambiente*, *incenerimento dei rifiuti pericolosi*, *assistenza reciproca*, *contratto di multiproprietà*) da corpora di dominio. Questo compito, le cui denominazioni spaziano nella letteratura sull'argomento da *terminology extraction* a *automatic term recognition* fino a *terminology mining*, rappresenta il primo e ormai consolidato passo nel processo incrementale di estrazione di conoscenza ontologica (denominato *Ontology Learning*[18]) da collezioni documentali (Buitelaar et alii, 2005)[19]. Si parte dall'assunto di base che i termini costituiscono la rappresentazione linguistica dei concetti specifici di un dominio e per questo motivo il compito di estrazione terminologica rappresenta il primo e fondamentale passo verso l'accesso al contenuto di collezioni documentali.

Il processo di estrazione terminologica si articola in due passi fondamentali:

1) identificazione delle potenziali unità terminologiche, siano esse monorematiche oppure polirematiche;
2) filtraggio della lista dei termini candidati al fine di discriminare la terminologia di dominio da non-termini (o parole comuni).

Queste due fasi del processo estrattivo possono essere basate su diversi tipi di evidenza, ovvero linguistica, statistica oppure una combinazione dei due: quest'ultimo rappresenta il caso più frequente.

---

18 Per *Ontology Learning* si intende il processo di supporto automatico o semi-automatico nello sviluppo di ontologie di dominio, attraverso l'acquisizione di conoscenza a partire dai testi. Un'ontologia è una rappresentazione formale e condivisa di un dato dominio di conoscenza.

19 Cfr. PAUL BUITELAAR, PHILIPP CIMIANO, BERNARDO MAGNINI, *Ontology Learning from Text: An Overview*, in Ontology Learning from Text: Methods, Evaluation and Applications, Buitelaar P. et alii (a cura di), in «Frontiers in Artificial Intelligence and Applications Series», vol. 123, IOS Press, 2005, pp. 3-12.

L'identificazione dei potenziali termini avviene in relazione al testo linguisticamente annotato, ovvero arricchito con informazione relativa alla struttura linguistica sottostante: tipicamente, a tal fine viene utilizzato il testo annotato morfo-sintatticamente oppure segmentato sintatticamente in costituenti sintattici elementari non ricorsivi detti *chunk* (Abney, 1991), (Federici et alii, 1996)[20]. Al livello di annotazione morfo-sintattica, a ogni parola (*token*[21]) del testo viene associata informazione relativa alla categoria grammaticale che la parola ha nel contesto specifico (ad es. sostantivo, verbo, aggettivo); nell'annotazione sintattica non ricorsiva, il testo viene segmentato in *chunk*, ovvero sequenze di parole del testo che vanno da un'unità grammaticale (tipicamente, una preposizione, un ausiliare, un [pre]determinatore o un ausiliare) fino alla prima unità lessicale semanticamente *piena* selezionata dall'unità grammaticale (esempi di *chunk* di tipo nominale sono: *un difficile problema* oppure *la mia prima casa*).

Il testo arricchito con informazione linguistica viene analizzato da una mini-grammatica deputata al riconoscimento delle strutture linguistiche che formano potenziali termini. Nel caso di unità terminologiche monorematiche, si farà riferimento alle categorie grammaticali: tipicamente, verranno identificati come candidati tutti i *token* etichettati come nomi, anche se in linea di principio la terminologia di un dominio specifico include anche aggettivi o verbi che designano proprietà o eventi tipici del dominio. In questa sede, ci limiteremo a considerare terminologia

---

20 Cfr. STEVEN ABNEY, *Parsing by chunks*, in Principle-based Parsing: Computation and Psycholinguistics, Berwick R.C. et alii (a cura di), Dordrecht, Kluwer, 1991, pp. 257-278. Per questo tipo di annotazione in relazione alla lingua italiana, cfr. STEFANO FEDERICI, SIMONETTA MONTEMAGNI, VITO PIRRELLI, *Shallow Parsing and Text Chunking: a View on Underspecification in Syntax*, in Proceedings of Workshop On Robust Parsing and Eight Summer School on Language, Logic and Information, Praga, Repubblica Ceca, 12-16 agosto 1996, pp. 35-44.

21 Ogni parola unità distinta del testo.

di tipo nominale. L'identificazione delle unità terminologiche polirematiche può avvenire in relazione al testo morfo-sintatticamente annotato o segmentato in costituenti sintattici non ricorsivi. In entrambi i casi, la mini-grammatica sarà finalizzata al riconoscimento di sequenze di categorie grammaticali o di *chunk* che corrispondono a potenziali unità polirematiche: si va da semplici casi di modificazione aggettivale (es. *inquinamento atmosferico*), a casi di modificazione con un complemento preposizionale (es. *diritto di recesso*), fino a strutture più complesse che combinano diversi tipi di modificatori (es. *encefalopatia spongiforme bovina*, *somatotropina bovina di ricombinazione*, *agenzia europea di valutazione dei medicinali* oppure *trattamento degli oli usati*).

La seconda fase del processo estrattivo è volta a verificare se e in quale misura un termine candidato rappresenti un termine valido per il dominio considerato. A questo scopo, nella letteratura sull'estrazione terminologica vengono utilizzate una serie di misure statistiche finalizzate a discriminare la terminologia di dominio da parole comuni, caratterizzate da una designazione generica. In particolare, l'estrazione di unità monorematiche è tipicamente realizzata sulla base della distribuzione di frequenza all'interno del corpus, oppure su misure di rilevanza statistica tipiche dell'Information Retrieval quali la *TF/IDF* (*Term Frequency/Inverse Document Frequency*) (Salton, Buckley, 1998), (Baeza-Yates, Ribeiro-Neto, 1999)[22]. Per le unità polirematiche, oltre alle misure già menzionate, vi sono metodi e tecniche che si basano sull'assunto di base che se due o più parole formano un termine è molto probabile che nell'uso linguistico relativo a quel dominio esse tendano a ricorrere insieme in maniera statistica-

[22] Cfr. GERARD SALTON, CHRIS BUCKLEY, *Term-Weighting Approaches in Automatic Text Retrieval*, in «Information Processing and Management», vol. 24, n. 5, 1988, pp. 513-523; Cfr. RICARDO BAEZA-YATES, BERTHIER RIBEIRO-NETO, *Modern Information Retrieval*, New York, ACM Press,1999.

mente significativa. La significatività del legame sussistente tra le parole che formano il termine viene calcolata attraverso il ricorso a misure di associazione che considerano la frequenza di co-occorrenza delle parole che compongono l'unità terminologica polirematica in relazione alle occorrenze totali delle singole parole che la formano: per menzionarne alcune, la misura della *Mutual Information* (Church, Hanks, 1990)[23] o della *Log-likelihood* (Dunning, 1993)[24] per arrivare al più recente e più sofisticato metodo denominato *C/NC-value* (Frantzi et alii, 2000)[25] che rappresenta uno standard *de facto* nel settore dell'estrazione terminologica (Vu et alii, 2008)[26]. Seguendo (Kageura, Umino, 1996)[27], la varietà di misure utilizzate per l'estrazione di terminologia settoriale da corpora di dominio può essere ricondotta a due classi fondamentali, a seconda che valutino:

a) l'unità di un termine (*unithood*), ovvero la forza di associazione che lega le parole che formano un'unità terminologica polirematica. Ovviamente, questo tipo di misura si applica solo nel caso di unità terminologiche formate da più parole;
b) la pertinenza rispetto al dominio (*termhood*), espressa co-

[23] Cfr. KENNETH WARD CHURCH, PATRICK HANKS, *Word association norms, mutual information, and lexicography*, in «Computational Linguistics», vol. 16, n. 1, 1990, pp. 22-29.

[24] Cfr. TED DUNNING, *Accurate Methods for the Statistics of Surprise and Coincidence*, in «Computational Linguistics», vol. 19, n. 1, 1993, pp. 61-74.

[25] Cfr. KATERINA FRANTZI, SOPHIA ANANIADOU, HIDEKI MIMA, *Automatic recognition of multi-word terms*, in «International Journal of Digital Libraries», vol. 3, n. 2, 2000, pp.117-132.

[26] Cfr. THUY VU, AITI AW, MIN ZHANG, *Term Extraction Through Unithood and Termhood Unification*, in Third International Joint Conference on Natural Language Processing. Proceedings of the Conference, Hyderabad, India, 07-12 gennaio 2008, pp. 631-636.

[27] Cfr. KYO KAGEURA, BIN UMINO, *Methods of automatic term recognition: a review*, in «Terminology», vol. 3, n. 2, 1996, pp. 259-289.

me misura del grado di rilevanza di una parola all'interno del dominio considerato, ovvero di quanto un termine candidato costituisca un'unità rappresentativa del contenuto della base documentale. Diversamente dal caso precedente, questa classe di misure si applica a unità terminologiche sia monorematiche sia polirematiche.

Mentre le misure come la frequenza grezza o *TF/IDF* così come il *C/NC-value* sono riconducibili alla seconda classe finalizzata alla quantificazione della rilevanza rispetto al dominio di un termine candidato, le misure di riconducibili alla prima classe, che includono la *Log-likelihood* e la *Mutual Information*, catturano piuttosto la forza e la stabilità dell'associazione che lega le parole che formano un termine polirematico. Data la complementarità di questi due tipi di misure, recentemente si registrano vari tentativi di combinarli ai fini dell'acquisizione di termini di dominio (Vu et alii, 2008)[28].

Nonostante le differenze rilevate, le misure viste finora si basano tutte sulla distribuzione dei termini candidati all'interno di uno stesso dominio, studiato attraverso una collezione documentale che ne convoglia i contenuti tipici. Un'altra classe di approcci all'estrazione terminologica si basa ancora su evidenza di tipo distribuzionale, ma rilevata attraverso un'analisi contrastiva inter-dominio: in questo caso, l'estrazione di unità terminologiche monorematiche e polirematiche è condotta a partire dal confronto della distribuzione dei termini candidati nel corpus di acquisizione rispetto a un corpus di riferimento (detto anche *corpus di contrasto*). In questo modo, la lista finale di unità terminologiche estratte conterrà quelle unità che sono maggiormente rilevanti nel corpus di acquisizione rispetto al corpus di riferimento. A questo scopo è stata sviluppata una serie di metodi in grado di computare la misura della diversa rilevanza di unità ter-

[28] Cfr. VU, T. et alii, *op. cit.*

minologiche all'interno dei due corpora oggetto dell'analisi contrastiva. La possibilità di discriminare termini e non-termini è così empiricamente realizzata sulla base del confronto della loro distribuzione in un corpus di dominio (il corpus di acquisizione) rispetto a un altro corpus: il corpus di riferimento è generalmente rappresentativo della lingua comune, ma a seconda del tipo di analisi contrastiva che si vuole condurre potrebbe anche essere relativo ad un altro dominio specialistico (cfr. sezione 5.2). Questo tipo di approccio può essere usato direttamente per l'identificazione dei termini all'interno di collezioni documentali di dominio (Basili et alii, 2001)[29], così come può essere utilizzato in combinazione con le misure precedenti (Bonin et alii, 2010a)[30].

Qualsiasi sia la tecnica adottata, il risultato del processo di estrazione automatica di terminologia da corpora di dominio dovrà essere validato e filtrato da parte di esperti che saranno supportati nelle decisioni finali non solo dalla loro competenza del dominio analizzato, ma anche da evidenza statistica che riflette la significatività dei termini acquisiti, sia essa costituita dalla rilevanza rispetto al dominio (intra-dominio o inter-dominio), oppure dalla forza di associazione che lega le parole all'interno di termini poliremiatici. Per questo motivo, il processo di costruzione di vocabolari controllati basato su questo approccio viene definito complessivamente come semi-automatico.

---

29 Cfr. ROBERTO BASILI, ALESSANDRO MOSCHITTI, MARIA TERESA PAZIENZA, FABIO MASSIMO ZANZOTTO, *A contrastive approach to term extraction*, in Atti della «4th Conference on Terminology and Artificial Intelligence (TIA-2001)», Nancy, 3-4 maggio 2001; Cfr. CHUNG, T.M., NATION, P., *op. cit.*; Cfr. ANSELMO PENAS, FELISA VERDEJO, JULIO GONZALO, *Corpus-Based Terminology Extraction Applied to Information Access*, in Proceedings of the Corpus Linguistics 2001 Conference, Università di Lancaster, 29 marzo – 2 aprile 2001, Rayson P., Wilson A., McEnery T., Hardie A., Khoja S. (ed.), pp. 458-465.

30 Cfr. FRANCESCA BONIN, FELICE DELL'ORLETTA, SIMONETTA MONTEMAGNI, GIULIA VENTURI (a), *A Contrastive Approach to Multi-word Extraction*

4. Questioni aperte nel processo di estrazione terminologica da corpora

Al termine della breve rassegna sui metodi e le tecniche correntemente usati per l'estrazione di terminologia da corpora di dominio riportata nella precedente sezione, appare legittimo chiedersi se la loro affidabilità ed efficacia possano essere influenzate dal tipo di linguaggio settoriale usato nel corpus impiegato per l'acquisizione. Concludendo la loro illustrazione del metodo *C/NC-value* testato su un corpus di dominio biomedico, (Frantzi et alii, 2000)[31] affermano che «*although we have shown that the method performs well for this text type of corpora, we are cautious in making this claim for other types of special language corpora, before conducting appropriate experiments*»: ciò lascia chiaramente intravedere che l'efficacia del metodo possa variare in relazione al tipo di corpus di acquisizione.

I sistemi di estrazione terminologica sono nati in relazione a collezioni di testi caratterizzati da un lessico altamente specialistico e rivolti a un pubblico di esperti, come ad esempio la letteratura biomedica. Se i risultati raggiunti su questa tipologia di testi sono ormai più che soddisfacenti, rimane una questione aperta: il loro *rendimento scientifico* in relazione a corpora rappresentativi di domini non altamente specialistici e/o composti da testi rivolti ad un ampio pubblico.

Gli approcci correnti al problema non sembrano aver affrontato in modo sistematico i diversi ordini di difficoltà connessi con la varia tipologia di linguaggi settoriali descritta nella sezione 2. In relazione a ciò, (Cabré, 1999)[32] ricorda come le maggiori

from Domain-specific Corpora, in Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 17-23 maggio 2010, pp. 3222–3229.

31 FRANTZI, K. et alii, *op. cit.*, p. 130.

32 Cfr. CABRÉ, M.T., *op. cit.*

difficoltà siano dovute proprio al confine non sempre così netto tra linguaggi settoriali e lingua comune, nonché al costante scambio biunivoco che li lega. La questione riguarda in particolare la difficoltà di estrarre terminologia rilevante di dominio, ovvero di distinguere tra termini del dominio (lessico settoriale) e non-termini (lessico comune), tenendo in considerazione la varie dimensioni di variazione all'interno di questa classe di linguaggi. In particolare, i problemi sono connessi con la difficoltà di estrarre la terminologia a partire da corpora rappresentativi di linguaggi settoriali caratterizzati da diversi livelli di specializzazione, oppure da collezioni di testi appartenenti a diversi tipi di registri.

La discriminazione tra termini settoriali e parole comuni non è tuttavia l'unico aspetto che non trova una risposta adeguata nei sistemi correnti di estrazione terminologica automatica. Un ulteriore problema riguarda il trattamento di testi all'interno dei quali si osserva una commistione di tipologie di termini, non sempre nettamente distinguibili tra di loro. Questo è il caso, ad esempio, dei testi giuridici: una peculiarità della lingua del diritto, che contribuisce alla sua eterogeneità dal punto di vista terminologico, è dovuta agli stretti e biunivoci rapporti con la lingua comune da un lato e con i linguaggi settoriali dall'altro (Cortelazzo, 1997), (Venturi, 2011)[33]. Ne consegue che nei corpora rappresentativi della lingua del diritto si intrecciano il lessico del dominio giuridico, quello proprio della materia legislata così come del linguaggio comune. Ai fini del presente studio, possiamo de-

[33] Cfr. MICHELE CORTELAZZO, *Lingua e diritto in Italia. Il punto di vista dei linguisti*, in La lingua del diritto. Difficoltà traduttive. Applicazioni didattiche, Atti del primo Convegno Internazionale, Milano, 5-6 ottobre 1995, Schena L. (a cura di), Roma, CISU (Centro d'Informazione e Stampa Universitaria), 1997, pp. 35-50; Cfr. GIULIA VENTURI, *Lingua e diritto: una prospettiva linguistico-computazionale*, Tesi di Dottorato, Università degli Studi di Torino, Scuola di Dottorato in Studi Umanistici, 2011.

finirli corpora *multi-dominio*: si tratta di una tipologia di testi che trova nei corpora giuridici un esempio prototipico, ma che non è circoscritta ad essi. Altri esempi di tale tipologia di corpora includono: il corpus dei testi che descrivono le linee di attività del CNR (Guarasci, 2006)[34] in cui si intreccia terminologia relativa alla gestione delle linee di attività con la terminologia del settore disciplinare a cui tali attività si riferiscono; oppure corpora della Pubblica Amministrazione, come quello trattato in (Taverniti, 2008)[35], in cui la terminologia relativa all'oggetto della comunicazione (i beni e i servizi informatici oggetto di acquisto) si combina con la terminologia propria del linguaggio burocratico.

Ai fini dell'indicizzazione di tali corpora, è molto importante poter discriminare tra i diversi tipi di contenuti. (Francesconi et alii, 2010)[36] hanno recentemente proposto un approccio alla rappresentazione formalizzata della conoscenza giuridica basato sulla distinzione tra conoscenza tecnico-giuridica e conoscenza del mondo regolato: il modello suggerito prevede infatti due distinti livelli di organizzazione, ovvero il *Domain Independent Legal Knowledge level* (DILK) che include concetti giuridici, e il *Domain Knowledge level* (DK) all'interno del quale sono resi espliciti i principali concetti rappresentativi di un determinato dominio di conoscenza regolato dalle norme. Questa doppia ar-

34 Cfr. ROBERTO GUARASCI, *Estrazione terminologica e gestione della conoscenza*, in «iged.it», n. 3, 2006, pp. 46-51.

35 Cfr. MARIA TAVERNITI, *Tra terminologia e documentazione: estrazione automatica di voci indice da corpora documentali della Pubblica Amministrazione*, in «Aida Informazioni», a. 26, n. 1-2, gennaio-giugno 2008, pp. 239-250.

36 Cfr. ENRICO FRANCESCONI, SIMONETTA MONTEMAGNI, WIM PETERS, DANIELA TISCORNIA, *Integrating a Bottom-Up and Top-Down Methodology for Building Semantic Resources for the Multilingual Legal Domain*, in Semantic Processing of Legal Texts, Francesconi E. et alii (a cura di), in «LNCS/LNAI», Springer-Verlag, vol. 6036, 2010, pp. 95-127.

ticolazione, riproposta al livello della terminologia acquisita, crea i presupposti per una indicizzazione multi-dimensionale del testo, dove sicuramente la prospettiva del dominio specialistico legislato rappresenta forse la chiave di accesso privilegiata. A conoscenza di chi scrive, le tecniche e i metodi di estrazione terminologica automatica correnti non si sono mai confrontati con casi di acquisizione di terminologia rilevante da corpora *multi-dominio* con il duplice fine di acquisire la terminologia rilevante e di classificarla secondo il dominio di appartenenza.

Nel momento in cui i sistemi di estrazione terminologica da corpora di dominio stanno diventando uno strumento sempre più usato in compiti di Knowledge Management quali l'indicizzazione automatica di testi, la costruzione di ontologie di dominio ecc., le sfide delineate sopra diventano aspetti che non possono essere ignorati nel processo di estrazione terminologica automatica e per i quali vanno trovate soluzioni operative appropriate. In quanto segue, partendo dall'esperienza condotta in diversi scenari applicativi con una piattaforma per l'estrazione da testi di conoscenza terminologico-ontologica (descritta nella sezione 5) verranno illustrati i problemi rilevati e le soluzioni proposte.

5. T2K: una piattaforma per l'estrazione di conoscenza ontologica da collezioni documentali

Text-to-Knowledge, in breve T2K, è una piattaforma software progettata e sviluppata congiuntamente dall'Istituto di Linguistica Computazionale *Antonio Zampolli* del CNR di Pisa e dal Dipartimento di Linguistica dell'Università di Pisa, che si propone di offrire una batteria integrata di strumenti avanzati di analisi linguistica del testo, analisi statistica e apprendimento automatico del linguaggio, destinati a offrire una rappresentazione accurata del contenuto di una base documentale non strutturata, per scopi di indicizzazione avanzata e navigazione intelligente (Del-

l'Orletta et alii, 2008)[37]. T2K trasforma le conoscenze implicitamente codificate all'interno di un corpus di testi in conoscenza esplicitamente strutturata: il risultato finale di questo processo interpretativo spazia dall'acquisizione di conoscenze lessicali e terminologiche complesse alla loro organizzazione in strutture proto-concettuali.

Per arrivare a identificare i concetti rilevanti e più caratterizzanti i documenti di un certo dominio di interesse, T2K impiega lo stato dell'arte della ricerca in linguistica computazionale. I termini acquisiti da T2K possono essere unità lessicali monorematiche come *monitoraggio* o *audit* oppure unità lessicali polirematiche come *Quadro Comunitario di Sostegno*, *obiettivi specifici*, *progetto integrato*, *autorità di gestione*, *autorità di pagamento*, ecc. Per quanto riguarda le unità monorematiche, il processo estrattivo opera sul testo annotato a livello morfo-sintattico[38] e lemmatizzato[39] e avviene sulla base della loro frequenza all'interno del corpus di acquisizione. Diverso è il caso delle unità terminologiche polirematiche, la cui estrazione si articola in due fasi: la prima finalizzata all'identificazione dei potenziali

37 Cfr. FELICE DELL'ORLETTA, ALESSANDRO LENCI, SIMONE MARCHI, SIMONETTA MONTEMAGNI, VITO PIRRELLI, GIULIA VENTURI, *Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio*, in «Aida Informazioni», a. 26, n. 1-2, gennaio-giugno 2008, pp. 185-206.

38 Lo scopo dell'annotazione morfo-sintattica è l'assegnazione a ogni parola (o *token*) del testo dell'informazione relativa alla *categoria grammaticale* (o *parte del discorso*) che la parola ha nel contesto specifico (ad es. nome, verbo, aggettivo). Questa informazione viene talora integrata da ulteriori specificazioni morfologiche (ad es. riguardanti categorie flessionali come persona, genere, numero, ecc.).

39 Il processo di lemmatizzazione consiste nel ricondurre ogni parola del testo al relativo esponente lessicale o *lemma* (tipicamente l'infinito per i verbi, oppure il singolare maschile per gli aggettivi, corrispondente approssimativamente all'esponente lessicale delle voci di un dizionario).

termini sulla base di una mini-grammatica operante sul testo segmentato sintatticamente in *chunk*, la seconda in cui la forza di associazione tra le parole che compongono il termine candidato viene stimata applicando la misura associativa detta *Log-likelihood*, che si è dimostrata produrre risultati sensibilmente migliori rispetto ad altre misure statistiche in quanto più robusta nel caso di dati linguistici con bassa frequenza di occorrenza. La compilazione di un repertorio di terminologia di dominio sulla base delle concrete attestazioni nei testi costituisce il risultato della prima fase operativa di T2K sulla base del quale è possibile condurre un'indicizzazione terminologica dei documenti. Si noti che in T2K è possibile nonché auspicabile validare il risultato del processo automatico di estrazione terminologica, in modo che il glossario di termini automaticamente acquisito possa diventare una risorsa di riferimento (ovvero rappresentativa dei termini di un dominio) sulla base della quale condurre l'indicizzazione dei testi.

I termini che formano il glossario terminologico automaticamente acquisito e validato dall'esperto di dominio sono a loro volta raggruppati secondo diverse relazioni di similarità semantica, che vanno dalle relazioni gerarchiche di iperonimia/iponimia (denominate anche BT *Broader Term* e NT *Narrower Term* nella terminologia dei thesauri per far riferimento rispettivamente al concetto più generico e a quello più specifico) a classi di termini semanticamente correlati (o RT, *Related Term* secondo la terminologia dei thesauri), ovvero termini genericamente correlati al termine di partenza da rapporti di implicazione e/o associazione semantica[40]. L'organizzazione e la strutturazione dei termini secondo le relazioni appena delineate rappresenta il risultato della seconda fase operativa di T2K, al termine della qua-

[40] In tal caso si parla anche di *quasi-sinonimi*: si tratta di una relazione di sinonimia relativa, nel senso che i termini sono considerati *sinonimi* essenzialmente ai fini dell'indicizzazione.

le è possibile condurre un'indicizzazione concettuale dei testi. Anche in questo caso, il risultato del processo automatico di estrazione di strutture proto-concettuali dovrà essere validato dall'esperto di dominio che costruirà l'ontologia di riferimento sulla base della quale condurre l'indicizzazione concettuale dei testi.

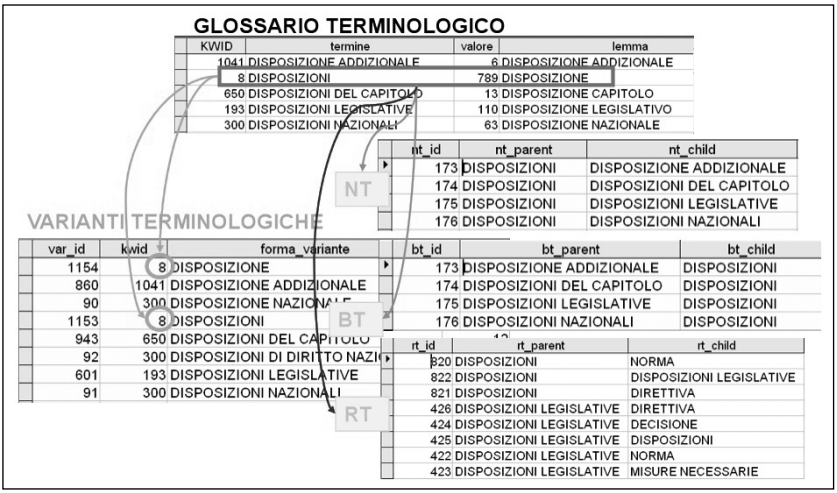


Figura 2. Frammento della base di conoscenza di T2K costruita dinamicamente a partire dai testi.

La Figura 2 riporta un frammento della base di conoscenza costruita dinamicamente a partire dai testi con la piattaforma software T2K. Le unità terminologiche monorematiche e polirematiche acquisite – integrate da informazione riguardante le relative varianti testuali (morfologiche ma anche strutturali) – sono organizzate all'interno di strutture proto-concettuali. Si vedano in particolare le tabelle contrassegnate come BT e NT, che contengono relazioni gerarchiche rispettivamente di iperonimia e iponimia e RT, che include relazioni tra termini genericamente correlati corrispondenti a rapporti di implicazione e/o associazione semantica.

La piattaforma T2K nella versione descritta sopra, d'ora in avanti denominata T2K_v.1, è stata utilizzata in diversi scenari applicativi: per la gestione documentale nella Pubblica Amministrazione (PA) (progetti *Traguardi* e *Pubblicamente.it* del *Formez*); per l'indicizzazione di contenuti didattici multimediali nell'E-learning (progetto PEKITA *Personalized Knowledge In The Air* in collaborazione con Università della Calabria e Siemens Italdata); per l'acquisizione di conoscenza ontologica da cataloghi di prodotti nell'ambito del progetto europeo VIKEF (*Virtual Information and Knowledge Environment Framework*, IP 507173); per lo sviluppo di risorse ontologiche a supporto del *drafting* legislativo nel progetto europeo DALOS (*Drafting Legislation with Ontology-based Support*, eParticipation project n. 2006/01/024). Inoltre, T2K è stato oggetto di sperimentazione con basi documentali di varia natura: per menzionarne alcune, la documentazione scientifica del CNR (Guarasci, 2006)[41]; corpora di testi giuridico-legislativi (Venturi, 2006)[42] e di letteratura linguistico-computazionale (Montemagni, 2007)[43]; per la costruzione di un vocabolario di indicizzazione per la gestione normalizzata e l'estrazione di conoscenza di documenti relativi ai pareri obbligatori sugli acquisti di beni e servizi informatici proposti dalle pubbliche amministrazioni centrali (Taverniti, 2008)[44] o sui temi dell'efficienza energetica e dell'applicazione di fonti rinnovabili agli usi finali civili (Oliveri et alii, 2010)[45].

41 Cfr. GUARASCI, R., *op. cit.*

42 Cfr. GIULIA VENTURI, *L'ambiente, le norme, il computer. Studio linguistico-computazionale per la creazione di ontologie giuridiche in materia ambientale*, Tesi di Laurea Specialistica, Università di Pisa, dicembre 2006.

43 Cfr. SIMONETTA MONTEMAGNI, *Acquisizione automatica di termini da testi: primi esperimenti di estrazione e strutturazione di terminologia metalinguistica*», in Lessicologia e metalinguaggio: Atti del Convegno, Macerata, 19-21 dicembre 2005, Poli D. (a cura di), Roma, Il Calamo, 2007.

44 Cfr. TAVERNITI, M., *op. cit.*

45 Cfr. ELISABETTA OLIVERI, CONCETTA BARONIELLO, ANTONIETTA FOLINO,

Nel corso degli ultimi due anni, la piattaforma T2K_v.1 è stata utilizzata in scenari applicativi che hanno portato alla luce nuove sfide per affrontare le quali sono state progettate e sperimentate soluzioni software innovative. Le principali novità della nuova piattaforma software T2K, d'ora innanzi denominata T2K_v.2, riguardano i componenti utilizzati per l'annotazione linguistica del testo e per l'estrazione di terminologia di dominio.

In T2K_v.2, l'annotazione linguistica è condotta mediante componenti di analisi del testo basati su metodi statistico-quantitativi in linea con il paradigma dominante nel settore della linguistica computazionale che è rappresentato da sistemi basati su algoritmi di apprendimento automatico supervisionato. Secondo questo approccio, il compito di *annotazione linguistica* viene modellato come un compito di classificazione probabilistica: a ogni passo di computazione il sistema sceglie l'annotazione più probabile data la parola in input, i suoi tratti descrittivi, il contesto e le annotazioni linguistiche già identificate. A partire da un corpus di addestramento, annotato con informazione morfo-sintattica e sintattica, viene costruito un modello probabilistico per l'annotazione linguistica del testo. In particolare, per quanto riguarda l'annotazione morfo-sintattica il componente utilizzato (Dell'Orletta, 2009)[46] risulta tra gli strumenti più precisi e affidabili secondo la campagna di valutazione di strumenti per l'analisi automatica dell'italiano EVALITA-2009[47].

ROSSELLA SCAIOLI, *Terminologia, lessici specialistici e strutture tassonomiche nel dominio dell'efficienza energetica e dell'applicazione di fonti rinnovabili agli usi finali civili*, Contributo alla «VI Giornata Scientifica della Rete Panlatina di Terminologia», Università dell'Algarve, Faro, Portogallo, 14 maggio 2010.

46 Cfr. FELICE DELL'ORLETTA, *Ensemble system for Part-of-Speech tagging*, in Atti della «11th Conference of Evaluation of NLP and Speech Tools for Italian (EVALITA) 2009», Reggio Emilia, 12 dicembre 2009.

47 Cfr. EVALITA, *Poster and Workshop Proceedings of the 11th Conference*

Il componente di estrazione terminologica all'interno di T2K_v.2 implementa una nuova strategia che opera sul testo morfo-sintatticamente annotato e procede in due fasi, la prima volta all'identificazione all'interno del corpus di acquisizione di unità terminologiche rilevanti per il dominio, la seconda basata sul confronto della distribuzione inter-dominio dei termini estratti nella fase precedente per una validazione della loro pertinenza.

Per quanto riguarda la prima fase, la maggiore novità riguarda l'acquisizione delle unità terminologiche polirematiche, la cui estrazione è basata sul metodo denominato *C/NC-value* (Frantzi et alii, 2000)[48] che appartiene alla classe delle misure di rilevanza rispetto al dominio (*termhood*). Questa misura tiene conto simultaneamente di quattro aspetti caratterizzanti il termine candidato, ovvero a) la sua frequenza di occorrenza all'interno del corpus di acquisizione, b) la sua frequenza di occorrenza come sottostringa di altri termini candidati, c) il numero di diversi termini candidati che lo contengono come sottostringa, e d) il numero di parole di cui si compone il termine candidato. Questa misura, denominata *C-value*, risulta particolarmente utile per il trattamento di terminologia complessa che include al suo interno altri termini. La lista di termini candidati definita sulla base del *C-value* viene ulteriormente rivista prendendo in considerazione informazione relativa ai contesti di occorrenza (*NC-value*).

I risultati ottenuti al termine della fase 1 per entrambe le tipologie di termini estratti (ovvero unità monorematiche e polirematiche) vengono filtrati sulla base di una funzione, chiamata *funzione di contrasto*, che valuta dal punto di vista quantitativo quanto un termine della lista estratta al passo precedente sia specifico di un certo dominio. Per calcolare la specificità del termine, sulla base della quale viene definito un nuovo ordinamento

of the Italian Association for Artificial Intelligence, Reggio Emilia, 12 Dicembre 2009. <http://www.evalita.it/2009/proceedings>.

48 Cfr. FRANTZI, K. et alii, *op. cit.*

dei termini in base alla pertinenza rispetto al dominio, viene considerata la distribuzione del termine sia nel corpus di acquisizione sia in un corpus differente, detto *corpus di contrasto*. La funzione di contrasto utilizzata, chiamata *Contrastive Selection of multi–word terms (CSmw)*, si è rivelata particolarmente adatta per l'analisi di variazioni distribuzionali di eventi a bassa frequenza (come appunto l'occorrenza di un termine polirematico). Se per una descrizione dettagliata del metodo si rinvia a (Bonin et alii, 2010a)[49], in questa sede vale la pena ricapitolare quali siano i principali elementi di novità dell'approccio proposto.

Contrariamente a (Penas et alii, 2001)[50], (Chung, Nation, 2004)[51] e (Basili et alii, 2001)[52], la fase di analisi contrastiva viene condotta in relazione alle unità terminologiche polirematiche acquisite nel corso della precedente fase: ciò è possibile grazie alla nuova funzione di contrasto che può essere applicata anche a eventi caratterizzati da basse frequenze. Questo previene potenziali problemi quali l'inclusione, nel risultato finale, di unità polirematiche non rilevanti ma lessicamente *governate* da una testa che è stata identificata come unità monorematica specifica per il dominio, oppure l'esclusione di unità polirematiche rilevanti che non sono state acquisite perché la loro testa lessicale non è stata selezionata come specifica per il dominio (Bonin et alii, 2012)[53]. Illustriamo quanto detto finora con un esempio, tratto da un esperimento di estrazione terminologica condotto su

49 Cfr. Bonin, F. et alii, *op. cit.*, 2010(a).
50 Cfr. Penas, A. et alii, *op. cit.*
51 Cfr. Chung, T.M., Nation, P., *op. cit.*
52 Cfr. Basili, R. et alii, *op. cit.*
53 Cfr. Francesca Bonin, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi, *Lessico settoriale e lessico comune nell'estrazione di terminologia specialistica da corpora di dominio*, in Lessico e lessicologia. Atti del XLIV Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI 2010), Viterbo, 27-29 settembre 2010, Ferreri S. (a cura di), 2012, pp. 207–220.

un corpus di articoli scientifici sul cambiamento climatico. Seguendo un approccio contrastivo di tipo tradizionale, l'acquisizione dell'unità terminologica polirematica *effetto serra* è subordinata alla identificazione dell'unità monorematica *effetto* come rilevante per il dominio: nel caso in cui l'unità monorematica *effetto* non sia stata selezionata come rilevante per il corpus di acquisizione, neanche l'unità polirematica, di cui essa è la testa, sarà estratta, sebbene essa sia significativa per il dominio. Ma se l'unità monorematica *effetto* è stata selezionata come rilevante, allora anche polirematiche come *effetto domino*, se ricorrenti nel testo, potranno essere qualificate come termini di dominio. Nell'approccio proposto, ciò non si verifica in quanto la funzione di contrasto opera direttamente sulla lista delle unità terminologiche polirematiche acquisite al passo precedente.

La nuova strategia di estrazione terminologica è stata verificata in diversi contesti applicativi e con risultati incoraggianti in relazione a corpora di leggi e sentenze (Venturi, 2011)[54], corpora web di cultura italiana (relativi ai domini di letteratura, arte e linguistica), corpora di referti relativi ad esami radiologici (Pirrelli et alii, 2010)[55]. In quanto segue, si riportano i risultati di esperimenti condotti in questi scenari applicativi che illustrano in dettaglio come la soluzione proposta sia in grado di fornire risultati più precisi e affidabili in relazione alle situazioni problematiche descritte nella sezione 4.

5.1 *Lessico settoriale vs lessico comune*[56]

Partiamo dal problema della discriminazione tra lessico settoriale e lessico comune. Riportiamo di seguito i risultati di un espe-

54 Cfr. VENTURI, G., *op. cit.*

55 Cfr. VITO PIRRELLI, ALESSANDRO LENCI, SIMONETTA MONTEMAGNI, FELICE DELL'ORLETTA, EMILIANO GIOVANNETTI, SIMONE MARCHI, *Connect To Life (modulo semantico): Rapporto Finale*, Rapporto Tecnico, 2010.

56 Parti di quanto segue sono riprese da: BONIN, F. et alii, *op. cit., 2012*, sezione 5.1; FRANCESCA BONIN, FELICE DELL'ORLETTA, GIULIA VENTURI, SI-

rimento condotto nell'ambito del progetto *Panorama FIRB: arte, lingua e letteratura italiana* (n. RBNE07C4R9, finanziato dal Ministero dell'Istruzione, dell'Università e della Ricerca) con T2K_v2 su un corpus di testi del settore della storia dell'arte. La sfida, nel caso specifico, è rappresentata dall'acquisizione di lessico settoriale da un corpus di testi caratterizzati da un livello non particolarmente alto di specializzazione. In particolare, l'estrazione di unità terminologiche monorematiche e polirematiche è stata condotta a partire da un corpus di testi di storia dell'arte estratti da siti di cultura italiana sul web (per un totale di 326.066 parole) costruito da esperti di dominio. Se tale corpus (denominato da ora in avanti ARTE) si presenta omogeneo rispetto al dominio, esso appare alquanto eterogeneo per quanto riguarda la tipologia di registri linguistici in esso testimoniati in ragione della natura variegata del web: in ARTE sono contenuti testi specialistici, così come testi divulgativi rivolti a un pubblico più vasto.

Per la fase di analisi *contrastiva* è stato selezionato un corpus di riferimento rispetto al quale confrontare la distribuzione delle unità terminologiche estratte dal corpus di acquisizione ARTE. Dato l'obiettivo di filtrare dal risultato finale il lessico comune, in questo esperimento come *corpus di contrasto* è stato usato il corpus PAROLE, un corpus di italiano contemporaneo di circa 3 milioni di parole (Marinelli et alii, 2003)[57].

MONETTA MONTEMAGNI (b), *Contrastive filtering of domain specific multiword terms from different types of corpora*, in Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications, Beijing, Cina, Coling 2010 Organizing Committee, agosto 2010, pp. 76-79, sezione 3.

57 Cfr. RITA MARINELLI, LISA BIAGINI, REMO BINDI, SARA GOGGI, MONICA MONACHINI, PAOLA ORSOLINI, EUGENIO PICCHI, SERGIO ROSSI, NICOLETTA CALZOLARI, ANTONIO ZAMPOLLI, *The Italian PAROLE corpus: an overview*, in «Linguistica Computazionale», Special Issue in «Computational Linguistics in Pisa», Zampolli A. et alii (a cura di), voll. 16-17, t. 1, Pisa-Roma, IEPI, 2003, pp. 401–421.

La Tabella 1 esemplifica il risultato della prima fase di estrazione, riportando le prime 10 unità terminologiche monorematiche e polirematiche delle rispettive liste ordinate per valori decrescenti di frequenza (unità monorematiche) e *C/NC-value* (unità polirematiche). Come si può notare, le liste risultanti da questa fase includono sia termini come *artista,* appartenenti evidentemente al lessico specialistico del settore artistico, sia voci come *anno* appartenenti piuttosto al lessico comune (marcate in corsivo nella tabella).

Ordinamento sulla base del filtro statistico (frequenza vs *C/NC-value*)			
	Unità monorematiche		**Unità polirematiche**
1	Arte	1	*Punto di vista*
2	Opera	2	Opera d'arte
3	Artista	3	Storia dell'arte
4	*Anno*	4	Arte contemporanea
5	Mostra	5	Figura umana
6	Parte	6	Bene culturale
7	Pittura	7	Storico dell'arte
8	*Secolo*	8	Movimento artistico
9	Forma	9	Produzione artistica
10	*Tempo*	10	*Anno scorso*

Tabella 1. Frammento delle liste di unità monorematiche e polirematiche estratte dopo la prima fase di estrazione terminologica. In corsivo i *non-termini.*

La seconda fase estrattiva, basata sulla distribuzione inter-dominio dei termini estratti, consiste nel riordinamento della lista (a tal fine vengono selezionati i primi 600 termini) sulla base della significatività dei termini di ARTE rispetto al corpus di contrasto[58]. È nel corso di questa fase di confronto della distribuzio-

[58] La soglia è stata stabilita su base sperimentale.

ne dei termini nei due corpora che il lessico settoriale viene distinto da quello comune. Grazie all'analisi contrastiva, le unità terminologiche precedentemente individuate come rilevanti per il corpus di acquisizione, ma non necessariamente per il dominio di acquisizione, vengono riordinate sulla base di un valore di contrasto. Da questa lista, vengono selezionati i termini risultanti alle prime 300 posizioni[59].

La Tabella 2 illustra il risultato della fase di analisi contrastiva che, come si può notare, ha consentito di filtrare nelle posizioni più alte della lista termini particolarmente specifici non solo per il corpus di acquisizione in sé, ma anche per il dominio trattato. Ad esempio, l'unità polirematica *anno scorso,* di pertinenza del lessico comune ma che occupava la decima posizione nella lista dei termini in Tabella 1, viene filtrata dopo la fase di confronto con il corpus di riferimento, scendendo oltre la trecentesima posizione.

Ordinamento sulla base delle funzione di contrasto (confronto PAROLE)			
	Unità monorematiche		**Unità polirematiche**
1	Artista	1	Opera d'arte
2	Pittura	2	Figura umana
3	Pittore	3	Movimento artistico
4	Scultura	4	Produzione artistica
5	Arte	5	Arte contemporanea
6	Mostra	6	Pittore italiano
7	Dipinto	7	Percorso espositivo
8	Affresco	8	Elemento architettonico
9	Architettura	9	Storia dell'arte
10	Museo	10	Storico dell'arte

Tabella 2. Frammento della lista finale di unità monorematiche e polirematiche estratte.

[59] La soglia è stata stabilita su base sperimentale.

La valutazione dei risultati raggiunti è stata condotta confrontando il glossario ottenuto con un Thesaurus di dominio[60], seguita da una fase di validazione da parte di esperti. Da questa duplice valutazione è emerso un aumento significativo dei termini di dominio estratti, che sono passati da 61,33% al termine della fase 1 al 79,40% a conclusione dell'analisi contrastiva, con un incremento relativo[61] registrato di +29,34%.

La discriminazione tra lessico settoriale e lessico comune può rappresentare un problema, sebbene di portata più limitata, anche nel caso di letteratura specialistica. In quanto segue, riportiamo i risultati di due esperimenti condotti con due corpora su tematiche di tipo ambientale, ovvero: a) letteratura scientifica (costituito da articoli scientifici sul cambiamento climatico per un totale di 397.297 *token*) e b) le voci di Wikipedia riconducibili al settore *Ecologia e Ambiente* per un totale di 174.391 *token*. Come nel precedente esperimento, ai fini dell'analisi contrastiva è stato usato il corpus PAROLE.

L'estrazione terminologica è stata condotta in due fasi, focalizzandosi sulla terminologia complessa. In particolare, i primi 2.000 termini risultanti dalla prima fase di analisi basata sul *C/NC-value* sono stati ordinati sulla base della funzione di contrasto *CSmw*. Da entrambe le liste ordinate di termini risultanti dalla prima e seconda fase di analisi sono stati estratti i primi 300 termini che sono stati sottoposti a valutazione. La valutazione è stata condotta semi-automaticamente: prima, i termini estratti sono stati confrontati con il *Thesaurus EARTh*[62] contenente 12.398

60 Il Thesaurus è stato fornito dal Dipartimento di Storia delle Arti dell'Università di Pisa.

61 Per tenere sotto controllo l'effetto della fase di analisi contrastiva, si è fatto ricorso all'incremento relativo (IR) ottenuto dividendo l'incremento assoluto osservato nel risultato della seconda fase per la numerosità dei termini estratti al termine della prima fase.

62 *Environmental Applications Reference Thesaurus*. <http://uta.iia.cnr.it/earth.htm#EARTh%202002>.

termini ambientali; i termini che non hanno trovato una corrispondenza all'interno della risorsa di riferimento selezionata sono stati manualmente validati da esperti di dominio. L'incremento relativo osservato nella selezione di terminologia rilevante rispetto al dominio al termine della seconda fase di analisi contrastiva è di +11,30% nel caso di Wikipedia e di +12,82% nel caso del corpus di articoli scientifici. Confrontando questo dato con l'incremento relativo osservato nel precedente esperimento in relazione a testi di storia dell'arte (+29,34%), possiamo concludere che questa strategia di analisi è particolarmente promettente con corpora caratterizzati da un linguaggio non altamente specialistico come quello della storia dell'arte ma ottiene miglioramenti significativi, anche se inferiori rispetto al caso precedente, anche nel caso di corpora di letteratura specialistica.

5.2 *Estrazione di terminologia multi-dominio*[63]

Come già accennato nella sezione 4, non si verifica sempre il caso che i corpora specialistici siano caratterizzati da un lessico espressione di un unico dominio di conoscenza e nettamente separato da quello comune. Un esempio di questo tipo è costituito dal dominio giuridico contraddistinto da una commistione di tipologie di termini, non sempre nettamente distinguibili tra di loro. (Agnoloni et alii, 2009)[64] così come (Lenci et alii, 2009)[65] ri-

[63] Parti di quanto segue sono riprese da: BONIN, F. et alii, *op. cit.*, 2010b, sezione 5.2; BONIN, F. et alii, *op. cit.*, 2012, sezione 3.

[64] Cfr. TOMMASO AGNOLONI, LORENZO BACCI, ENRICO FRANCESCONI, WIM PETERS, SIMONETTA MONTEMAGNI, GIULIA VENTURI, *A two-level knowledge approach to support multilingual legislative drafting*, in Law, Ontologies and the Semantic Web - Channelling the Legal Information Flood, Breuker J. et alii (a cura di), in «Frontiers in Artificial Intelligence and Applications», Springer, vol. 188, 2009, pp. 177-198.

[65] Cfr. ALESSANDRO LENCI, SIMONETTA MONTEMAGNI, VITO PIRRELLI, GIULIA VENTURI, *Ontology learning from Italian legal texts*, in Law, Ontologies and the Semantic Web - Channelling the Legal Information Flood,

portano che i metodi tradizionali di estrazione terminologica su questo tipo di testi arrivano ad acquisire repertori terminologici nei quali le diverse tipologie di termini sono mescolate. Questi ultimi notano anche che il repertorio terminologico acquisito contiene un maggior numero di termini giuridici rispetto a quelli relativi alla materia legislata. Ciò viene ricondotto alla bassa frequenza (e alto rango[66]) di quest'ultimo tipo di termini nel corpus di testi giuridici di partenza, in accordo con la legge di Zipf, secondo la quale la frequenza di una parola è inversamente proporzionale al suo rango.

Abbiamo provato ad affrontare questa situazione applicando il metodo contrastivo descritto in precedenza. Riportiamo di seguito i risultati di un esperimento condotto con T2K_v2 su una collezione di direttive europee in materia ambientale per un totale di 394.088 parole (d'ora in avanti AMB), reperito dalla versione disponibile on-line del Bollettino Giuridico Ambientale[67]. In questo caso, la metodologia contrastiva di estrazione terminologica ha svolto un duplice ruolo, finalizzato non solo a discriminare il lessico rilevante in AMB dal lessico comune, ma anche a distinguere il lessico del diritto da quello del dominio ambientale. A questo scopo sono stati usati due corpora di riferimento: il corpus PAROLE e un corpus di direttive europee in materia di protezione del consumatore (per un totale di 72.210 parole, d'ora in avanti CONS). In questo caso, l'analisi si è concentrata sull'estrazione di unità terminologiche poliematiche.

Analogamente agli esperimenti precedenti, è stata estratta una lista di 600 unità terminologiche poliematiche ordinate per va-

Breuker J. et alii (a cura di), in «Frontiers in Artificial Intelligence and Applications», vol. 188, Springer, 2009, pp. 75-94.

66 Nella lista delle parole di un testo ordinata per valori decrescenti di frequenza, il *rango* si riferisce alla posizione che una data parola occupa all'interno della lista.

67 <http://extranet.regione.piemonte.it/ambiente/bga/index.htm>.

lori decrescenti sulla base dei valori di *C/NC-value*; in questo caso si osserva la compresenza di unità appartenenti sia al lessico comune (es. *anno successivo*) sia al lessico del diritto (es. *norma nazionale*), sia a quello ambientale (es. *effetto serra*).

La fase di analisi contrastiva in questo caso è stata suddivisa in due passi, ciascuno condotto rispetto a un diverso corpus di riferimento: prima, con il corpus PAROLE per discriminare i termini rilevanti per AMB (sia giuridici sia ambientali) dai *non-termini*; in seconda battuta, con CONS per distinguere tra i termini del lessico del diritto e quelli del lessico ambientale. Il primo passo dell'analisi contrastiva ha riguardato le prime 600 unità terminologiche; da questa lista di unità riordinate sulla base della loro rilevanza per AMB, sono state selezionate le prime 300 su cui si è incentrata la seconda fase di analisi contrastiva basata sul confronto con CONS, volta a distinguere le unità proprie del lessico del diritto da quelle del dominio ambientale.

La Tabella 3 riporta nelle due colonne iniziali le prime 10 unità terminologiche della lista estratta al termine della fase 1, nelle ultime due colonne le prime e ultime cinque posizioni della lista risultante dalla doppia analisi contrastiva (fase 2). Come si può vedere, mentre a conclusione della fase 1 i termini appartenenti al lessico del diritto (in corsivo) si affiancano a termini ambientali (in grassetto) nelle prime posizioni della lista, nella lista finale i termini dei due lessici settoriali sono riordinati in modo da essere distinti (ovvero, la testa della lista contiene i termini ambientali mentre nella coda si concentrano quelli del diritto).

La valutazione è avvenuta in modo semi-automatico, analogamente ai precedenti esperimenti. Come risorse di riferimento per la valutazione dei risultati conseguiti, sono stati selezionati il *Dizionario Giuridico* (Edizioni Simone)[68] e il *Thesaurus EARTh* sopra citato. Anche in questo caso, i risultati raggiunti dimostra-

[68] <http://www.simone.it/cgi-local/Dizionari/newdiz.cgi?index,5,A>.

Ordinamento sulla base del filtro statistico (*C/NC-value*)	Unità polirematiche	Ordinamento sulla base della funzione di contrasto (confronto con CONS)	Unità polirematiche
1	*parlamento europeo*	1	**valore limite**
2	*autorità competente*	2	**sostanza pericolosa**
3	**valore limite**	3	**salute umana**
4	*valore limite di emissione*	4	**effetto serra**
5	*stato membro*	5	**sviluppo sostenibile**
6	**limite di emissione**	296	*diritto nazionale*
7	**sostanza pericolosa**	297	*testo della disposizione*
8	*destinatario della presente direttiva*	298	*disposizione essenziale del diritto interno*
9	*misura necessaria*	299	*disposizione nazionale*
10	**sviluppo sostenibile**	300	*funzionamento del mercato interno*

Tabella 3. Frammenti delle liste ordinate di unità polirematiche estratte al termine delle fasi 1 e 2.

no l'efficacia di questo approccio all'estrazione terminologica: mentre, infatti, dopo l'estrazione sulla base del *C/NC-value* il 65,34% dei primi 300 termini della lista ordinata era costituito da unità polirematiche del lessico ambientale (38,67%) e del lessico del diritto (26,67%), al termine della doppia analisi contrastiva le unità terminologiche ambientali sono aumentate fino al 43,33% e quelle del lessico del diritto fino al 29,33%. Per quanto riguarda il lessico ambientale l'incremento relativo è del 23,81%.

6. Conclusioni

I sistemi di estrazione automatica di terminologia da corpora di dominio sono oggi considerati maturi per poter essere integrati in applicazioni reali, per l'indicizzazione automatica di basi documentali e l'accesso su base semantica ai contenuti. I migliori risultati sono ottenuti nei casi di acquisizione di terminologia di dominio da testi caratterizzati da un lessico altamente specialistico e rivolti ad un pubblico di esperti, come ad esempio la letteratura biomedica. Il rendimento scientifico di tali sistemi decresce significativamente quando la collezione documentale usata come corpus di acquisizione non appartenga alla classe dei testi altamente specialistici. Infatti, l'analisi di testi che occupano una posizione intermedia nel continuum tra linguaggi altamente specialistici e lingua comune rappresenta una sfida tuttora aperta per tali sistemi. Un'ulteriore e non secondaria sfida riguarda la necessità di distinguere, all'interno di un corpus rappresentativo di un unico linguaggio settoriale, i termini appartenenti a diversi domini del sapere; ad es. il lessico del diritto da quello proprio della materia legislata nel caso di corpora giuridici. Ad oggi, a conoscenza di chi scrive, nessun sistema automatico ha affrontato questi due ordini di problemi in modo sistematico.

Partendo dall'analisi critica dei risultati ottenuti con un approccio *standard* all'estrazione terminologica in diversi scenari applicativi, il presente contributo raccoglie gli sforzi condotti per cercare di colmare le lacune e i limiti identificati nei sistemi correnti di estrazione terminologica, fornendo una risposta al problema dell'acquisizione di terminologia da corpora non altamente specialistici e da corpora *multi-dominio*. I risultati conseguiti, sebbene ancora preliminari, sono incoraggianti: gli scenari applicativi trattati sono vari, con un incremento relativo nella terminologia rilevante estratta che va dall'11/12% nel caso di testi specialistici (cfr. sezione 5.1), a più del 23% nel caso di corpora

giuridici (cfr. sezione 5.2), per arrivare fino al 29% registrato nel caso del corpus web di storia dell'arte (cfr. sezione 5.1). Altri esperimenti con risultati interessanti sono stati condotti con corpora di sentenze e corpora di referti diagnostici provenienti da reparti di Senologia Radiologica di diversi ospedali; mentre per quanto riguarda le sentenze la valutazione dei risultati è ancora in corso, nel secondo caso si è osservato un incremento altrettanto significativo rispetto a quanto riportato sopra.

Sulla base dei risultati raggiunti, si può affermare che T2K_v2, il prototipo software che implementa la nuova strategia estrattiva illustrata nelle precedenti pagine, sa far fronte in modo più che soddisfacente alle sfide poste da corpora non altamente specialistici o *multi-dominio*, fornendo così una prima risposta ai desiderata espressi in (Oliveri et alii, 2010)[69] che concludono il loro articolo auspicando *«una fase di estrazione terminologica tematica che recuperi solo i termini rappresentativi del contenuto concettuale dei documenti e al tempo stesso del dominio di riferimento»*.

Potenziali ed interessanti estensioni del metodo contrastivo per l'estrazione di terminologia di dominio includono il trattamento di variazioni di registro all'interno dello stesso linguaggio settoriale così come la ricostruzione dell'evoluzione diacronica di un lessico settoriale.

Per quanto riguarda le variazioni di registro, (Oliveri et alii, 2010)[70] nel loro studio sulla terminologia specialistica nel dominio dell'efficienza energetica e dell'applicazione di fonti rinnovabili agli usi finali civili si sono trovati a trattare con sottocorpora di testi caratterizzati da diversi livelli di specializzazione e da diverse finalità comunicative (ovvero articoli e rendicontazioni scientifiche, riviste di settore, leggi e norme, opuscoli e linee

69 Cfr. OLIVERI, E. et alii, *op. cit*.
70 Cfr. OLIVERI, E. et alii, *op. cit*.

guida). Il processo estrattivo è stato condotto separatamente per ciascun sottocorpus al fine di poter estrarre termini specialistici e termini appartenenti al linguaggio comune, tra i quali, nel thesaurus, sono state stabilite relazioni di equivalenza. Un'analisi contrastiva di tali collezioni di documenti, condotta con corpora di contrasto adeguatamente selezionati, dovrebbe poter rendere possibile l'estrazione della terminologia settoriale tipica di ogni registro, fornendo così all'esperto ulteriore evidenza utile per l'arricchimento del vocabolario controllato o del thesaurus.

Un altro aspetto importante riguarda l'evoluzione diacronica della terminologia settoriale, che va di pari passo con l'evolversi delle conoscenze all'interno di un dominio. L'estrazione terminologica non va ad operare su collezioni documentali chiuse che cambiano raramente se non mai, come ad esempio la produzione letteraria di un autore del passato. Un sistema di estrazione terminologica si trova tipicamente a trattare con collezioni documentali aperte e dinamiche, continuamente aggiornate con nuovi documenti, comprendenti nuovi contenuti e dunque nuova terminologia. La domanda è se sia possibile utilizzare il metodo estrattivo qui proposto anche per identificare termini in entrata e/o in uscita. Primi esperimenti in questa direzione, condotti sulla lingua comune, hanno fornito risultati interessanti (Montemagni, 2010)[71]. Il monitoraggio terminologico-lessicale alla ricerca di parole *in entrata* condotto su un corpus giornalistico tratto da *La Repubblica* degli anni 2002-2005 usando come corpus di contrasto un corpus della stessa testata di un periodo antecedente (1992-1995) ha identificato nelle unità monorematiche come *tsunami*, *devolution*, *web*, *info*, *sms*, *bipartisan*, *dvd*, ecc. o nelle

[71] Cfr. SIMONETTA MONTEMAGNI, *Tecnologie linguistico-computazionali per il monitoraggio della lingua italiana*, presentazione tenuta nell'ambito della Giornata di Studio «Lo stato della lingua. Il CNR e l'italiano nel terzo millennio», Roma, Consiglio Nazionale delle Ricerche - Dipartimento Identità Culturale, 8 marzo 2010.

unità polirematiche *milione di euro*, *influenza aviaria*, *cellule staminali*, *reality show* e *digitale terrestre* le parole caratterizzanti il corpus 2002-2005 rispetto a quello 1992-1995. Il monitoraggio terminologico-lessicale finalizzato alla ricerca delle parole *in via di recessione*, condotto analizzando il corpus del 1992-1995 in relazione a quello del 2002-2005 (usato per l'analisi contrastiva), ha fatto emergere voci quali *minimum tax*, *miliardo di marchi*, *svalutazione della lira*, *patto in deroga* oppure *quadripartito*, *pidiessini*, *rublo* come espressioni (polirematiche o monorematiche) che stanno scomparendo dall'uso linguistico. Seguendo questo approccio, sarebbe quindi possibile qualificare le voci di un vocabolario controllato o di thesaurus non solo in rapporto a un registro linguistico, ma anche sull'asse diacronico.

Ringraziamenti

Gli strumenti e le tecnologie illustrati in questo articolo sono il frutto del lavoro di collaborazione tra due gruppi di ricerca, rispettivamente dell'Istituto di Linguistica Computazionale *Antonio Zampolli* del Consiglio Nazionale delle Ricerche di Pisa, e del Dipartimento di Linguistica *Tristano Bolelli* dell'Università di Pisa, nell'ambito prima del laboratorio interistituzionale DY-LAN Lab, e oggi dell'ItaliaNLP Lab. In particolare, Felice Dell'Orletta e Giulia Venturi hanno contribuito in misura sostanziale al disegno e allo sviluppo di T2K_v2 per l'estrazione di terminologia settoriale da corpora di dominio di cui sono stati qui delineati i principi fondamentali.

Bibliografia

ABNEY, S. *Parsing by chunks*, in Principle-based Parsing: Computation and Psycholinguistics, Berwick R.C. et alii (a cura di), Dordrecht, Kluwer, 1991, pp. 257-278

AGNOLONI, T., BACCI, L., FRANCESCONI, E., PETERS, W., MONTEMAGNI, S., VENTURI, G. *A two-level knowledge approach to support multilingual legislative drafting*, in Law, Ontologies and the Semantic Web - Channelling the Legal Information Flood, Breuker J. et alii (a cura di), in «Frontiers in Artificial Intelligence and Applications», Springer, vol. 188, 2009, pp. 177-198

BAEZA-YATES, R., RIBEIRO-NETO, B. *Modern Information Retrieval*, New York, ACM Press, 1999

BASILI, R., MOSCHITTI, A., PAZIENZA M.T., ZANZOTTO F.M., *A contrastive approach to term extraction*, in Atti della «4th Conference on Terminology and Artificial Intelligence (TIA-2001)», Nancy, 3-4 maggio 2001

BONIN, F., DELL'ORLETTA, F., MONTEMAGNI, S., VENTURI, G. (a) *A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora*, in Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 17-23 maggio 2010, pp. 3222–3229

BONIN, F., DELL'ORLETTA, F., VENTURI, G., MONTEMAGNI, S. (b) *Contrastive filtering of domain specific multi-word terms from different types of corpora*, in Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications, Beijing, Cina, agosto 2010, Coling 2010 Organizing Committee, pp. 76-79

BONIN, F., DELL'ORLETTA, F., MONTEMAGNI, S., VENTURI, G., *Lessico settoriale e lessico comune nell'estrazione di terminologia specialistica da corpora di dominio*, in Lessico e lessicologia. Atti del XLIV Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI 2010), Viterbo, 27-29 settembre 2010, Ferreri S. (a cura di), 2012, pp. 207–220

BUITELAAR, P., CIMIANO, P., MAGNINI, B., *Ontology Learning from Text: An Overview*, in Ontology Learning from Text: Methods, Evaluation and Applications, Buitelaar P. et alii (a cura di), in «Frontiers in Artificial Intelligence and Applications Series», vol. 123, IOS Press, 2005, pp. 3-12

CABRÉ, M.T., *Terminology: Theory, Methods, and Applications*, Amsterdam, John Benjamins, 1999

CAVAGNOLI, S., *La comunicazione specialistica*, Roma, Carocci, 2007

CHUNG, T.M., NATION P., *Identifying technical vocabulary*, in «System», vol. 32, 2004, pp. 251-263

CHURCH, K.W., HANKS, P., *Word association norms, mutual information, and lexicography*, in «Computational Linguistics», vol. 16, n.1, 1990, pp. 22-29

CORTELAZZO, M., *Lingua e diritto in Italia. Il punto di vista dei linguisti*, in La lingua del diritto. Difficoltà traduttive. Applicazioni didattiche. Atti del primo Convegno Internazionale, Milano, 5-6 ottobre 1995, Schena L. (a cura di) Roma, CISU (Centro d'Informazione e Stampa Universitaria), 1997, pp. 35-50

DARDANO, M., *Linguaggi settoriali e processi di riformulazione*, in Parallela 3. Linguistica contrastiva / Linguaggi settoriali / Sintassi generativa, Dressler W. et alii (a cura di), Tubinga, Narr, 1987, pp. 134-145

DELL'ORLETTA, F., *Ensemble system for Part-of-Speech tagging*, in Atti della «11th Conference of Evaluation of NLP and Speech Tools for Italian

(EVALITA) 2009», Reggio Emilia, 12 dicembre 2009

DELL'ORLETTA, F., LENCI, A., MARCHI, S., MONTEMAGNI, S., PIRRELLI, V., VENTURI, G., *Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio*, in «Aida Informazioni», a. 26, n. 1-2, gennaio-giugno 2008, pp. 185-206

DUNNING, T., *Accurate Methods for the Statistics of Surprise and Coincidence*, in «Computational Linguistics», vol. 19, n. 1, 1993, pp. 61-74

EVALITA, *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, 12 dicembre 2009
<http://www.evalita.it/2009/proceedings>

FEDERICI, S., MONTEMAGNI, S., PIRRELLI, V., *Shallow Parsing and Text Chunking: a View on Underspecification in Syntax*, in Proceedings of Workshop On Robust Parsing and Eight Summer School on Language, Logic and Information, Praga, Repubblica Ceca, 12-16 agosto 1996, pp. 35-44

FRANCESCONI, E., MONTEMAGNI, S., PETERS, W., TISCORNIA, D., *Integrating a Bottom-Up and Top-Down Methodology for Building Semantic Resources for the Multilingual Legal Domain*, in Semantic Processing of Legal Texts, Francesconi E. et alii (a cura di), in «LNCS/LNAI», Springer-Verlag, vol. 6036, 2010, pp. 95-127

FRANTZI, K., ANANIADOU, S., MIMA, H., *Automatic recognition of multi-word terms*, in «International Journal of Digital Libraries», vol. 3, n. 2, 2000, pp.117-132

GUARASCI, R., *Estrazione terminologica e gestione della conoscenza*, in «iged.it», n. 3, 2006, pp. 46-51

JACKENDOFF, R., *Twistin' the night away*, in «Language», vol. 73, 1997, pp. 534-559

KAGEURA, K., UMINO, B., *Methods of automatic term recognition: a review*, in «Terminology», vol. 3, n. 2, 1996, pp. 259-289

LAVINIO, C., *Comunicazione e linguaggi disciplinari. Per un'educazione linguistica trasversale*, Roma, Carocci, 2004

LENCI, A., MONTEMAGNI, S., PIRRELLI, V., VENTURI, G., *Ontology learning from Italian legal texts*, in Law, Ontologies and the Semantic Web - Channelling the Legal Information Flood, Breuker J. et alii (a cura di), in «Frontiers in Artificial Intelligence and Applications», vol. 188, Springer, 2009, pp. 75-94

MANNING, C.D., RAGHAVAN, P., SCHÜTZE, H., *Introduction to Information Retrieval*, Cambridge University Press, 2008

MARINELLI, R., BIAGINI, L., BINDI, R., GOGGI, S., MONACHINI, M., ORSOLINI, P., PICCHI, E., ROSSI, S., CALZOLARI, N., ZAMPOLLI, A., *The Italian PARO-*

LE corpus: an overview, in «Linguistica Computazionale», Special Issue in «Computational Linguistics in Pisa», Zampolli A. et alii (a cura di), voll. 16-17, Tomo I, Pisa-Roma, IEPI, 2003, pp. 401–421

MONTEMAGNI, S., *Acquisizione automatica di termini da testi: primi esperimenti di estrazione e strutturazione di terminologia metalinguistica*», in Lessicologia e metalinguaggio: Atti del Convegno, Macerata, 19-21 dicembre 2005, Poli D. (a cura di), Roma, Il Calamo, 2007

MONTEMAGNI, S., *Tecnologie linguistico-computazionali per il monitoraggio della lingua italiana*, presentazione tenuta nell'ambito della Giornata di Studio «Lo stato della lingua. Il CNR e l'italiano nel terzo millennio», Roma, Consiglio Nazionale delle Ricerche - Dipartimento Identità Culturale, 8 marzo 2010

NAKAGAWA, H., MORI, T., *Automatic Term Recognition based on Statistics of Compound Nouns and their Components*, in «Terminology», vol. 9, n. 2, 2003, pp. 201–209

OLIVERI, E., BARONIELLO, C., FOLINO, A., SCAIOLI, R., *Terminologia, lessici specialistici e strutture tassonomiche nel dominio dell'efficienza energetica e dell'applicazione di fonti rinnovabili agli usi finali civili*, Contributo alla «VI Giornata Scientifica della Rete Panlatina di Terminologia», Università dell'Algarve, Faro, Portogallo, 14 maggio 2010

PENAS, A., VERDEJO, F., GONZALO, J., *Corpus-Based Terminology Extraction Applied to Information Access*, in Proceedings of the Corpus Linguistics 2001 Conference, Università di Lancaster, 29 marzo – 2 aprile 2011, Rayson P., Wilson A., McEnery T., Hardie A., Khoja S. (ed.), pp. 458-465

PIRRELLI, V., LENCI, A., MONTEMAGNI, S., DELL'ORLETTA, F., GIOVANNETTI, E., MARCHI, S., *Connect To Life (modulo semantico): Rapporto Finale*, Rapporto Tecnico, CNR-ILC-Dylan LAB, TR-008, 2010

PORCELLI, G., *Principi di Glottodidattica*, Brescia, La Scuola, 1994

RONDEAU, G., *Introduction à la terminologie*, Québec, Gaëtan Morin éditeur, 1983

RONDEAU, G., SAGER, J., *Introduction à la terminologie*, ed. 2, Chicoutimi, Gatan Morin, 1984

SALTON, G., BUCKLEY, C., *Term-Weighting Approaches in Automatic Text Retrieval*, in «Information Processing and Management», vol. 24, n. 5, 1988, pp. 513-523

SOBRERO, A., *Lingue speciali*, in Introduzione all'italiano contemporaneo. La variazione e gli usi, Sobrero A. (a cura di), Roma-Bari, Laterza, 1993, pp. 237-277

TAVERNITI, M., *Tra terminologia e documentazione: estrazione automatica di voci indice da corpora documentali della Pubblica Amministrazione*, in «Aida Informazioni», a. 26, n. 1-2, gennaio-giugno 2008, pp. 239-250

VARANTOLA, K., *Special Language and General Language: Linguistic and Didactic Aspects*, in «Unesco ALSED-LSP Newsletter», vol. 9, n. 2, dicembre 1986, pp. 10-20

VENTURI, G., *L'ambiente, le norme, il computer. Studio linguistico–computazionale per la creazione di ontologie giuridiche in materia ambientale*, Tesi di Laurea Specialistica, Università di Pisa, dicembre 2006

VENTURI, G., *Lingua e diritto: una prospettiva linguistico-computazionale*, Tesi di Dottorato, Università degli Studi di Torino, Scuola di Dottorato in Studi Umanistici, 2011

VU, T., AW, A., ZHANG, M., *Term Extraction Through Unithood and Termhood Unification*, in Third International Joint Conference on Natural Language Processing. Proceedings of the Conference, Hyderabad, India, 07-12 gennaio 2008, pp. 631-636

Sitografia

<http://uta.iia.cnr.it/earth.htm#EARTh%202002>
<http://extranet.regione.piemonte.it/ambiente/bga/index.htm>
<http://www.simone.it/cgi-local/Dizionari/newdiz.cgi?index,5,A>

Data Mining e Text Mining*

GIORGIO GAMBOSI** – MAURIZIO LANCIA***

Introduzione

Lo sviluppo tecnologico ha consentito di raccogliere ed immagazzinare enormi quantità di dati, con una ricchezza di informazioni potenzialmente accessibili che eccede quanto era nelle intenzioni di chi li aveva raccolti, ovvero aveva progettato la banca dati. Tali informazioni possono essere particolarmente preziose per i processi decisionali, consentendo di estrarre informazioni di sintesi (conoscenza nascosta) da un insieme di dati apparentemente non correlati e di difficile interpretazione.

L'ambito scientifico rappresenta uno dei contesti in cui tale tipo di situazione può presentarsi, a causa della elevatissima quantità di dati che possono essere raccolti mediante la moderna strumentazione. Un esempio eclatante, in tal senso, viene dal mondo dell'astrofisica. In quel contesto esistono archivi di enormi dimensioni, vere e proprie biblioteche digitali, contenenti gli insiemi di dati raccolti in occasione di campagne attuate tramite l'u-

---

* Questo contributo rappresenta una revisione critica e un aggiornamento del lavoro *Data Mining e Text Mining* realizzato da Maurizio Lancia e Andrea Lapiccirella e apparso in Guarasci R. (a cura di), *Dal documento all'informazione*, Milano, ITER, 2008, pp. 277-314.

** Università di Roma Tor Vergata, Dipartimento di Ingegneria dell'Impresa.

*** Ufficio Sistemi Informativi e documentali del Consiglio Nazionale delle Ricerche.

tilizzo di satelliti. Le *strisciate sperimentali* di questo tipo di dati sono molto ridondanti (raccolgono moltissimi parametri non indipendenti fra loro) ed il loro contenuto informativo eccede di molto quanto strettamente necessario al raggiungimento dell'obiettivo primario per il quale la campagna di raccolta era stata disegnata. Questo fa sì che oggi le banche in cui sono raccolti tali dati siano oggetto di studi ed analisi relativi a tematiche completamente diverse da quelle che si volevano studiare all'atto del disegno della campagne di raccolta. In questa maniera, lavorando unicamente su dati già raccolti, si effettuano vere e proprie campagne virtuali (si parla di osservatori virtuali) che spaziano e fondono assieme tutto ciò che già c'è, arrivando a conclusioni e scoperte di conoscenze completamente nuove e assolutamente non considerate all'atto del disegno dell'esperimento originale. È inutile dire che in questa maniera si ottengono considerevoli risparmi ed il rapporto benefici su costo di tali nuove indagini è assolutamente favorevole.

Tuttavia, se non si dispone di strumenti che consentano di individuare ed accedere alle informazioni nascoste, queste enormi quantità di dati rischiano di essere inefficaci per prendere decisioni, arrivando così a costituire più una perdita che un *asset*. Per far fronte a questo problema si è sviluppata negli ultimi anni un'area disciplinare che, prendendo spunto da altre discipline, si è posta l'obiettivo di fornire gli strumenti e le metodologie di estrazione della conoscenza all'interno di banche dati di grandi dimensioni.

Il processo strutturato di estrazione di conoscenza dai dati, ovvero il processo di selezione, esplorazione e modellazione di basi dati molto consistenti al fine di scoprire correlazioni, similitudini (*patterns*), sequenze e tendenze (*trends*) è conosciuto come *knowledge discovery in databases* (KDD).

Il Data Mining è una particolare fase di questo processo e consiste nell'applicazione di analisi di dati e di algoritmi di scoperta che portano all'individuazione di *pattern* tra i dati. Per far

questo usa principalmente metodi statistici ed impiega una o più tecniche di apprendimento computerizzato con lo scopo di produrre un modello che formalizzi le non ovvie conoscenze individuabili nei dati.

C'è molta confusione circa l'esatto significato dei termini *Data Mining* e *KDD*, che vengono spesso considerati sinonimi. Diversi autori tuttavia, nell'ultimo periodo, hanno preferito usare il termine KDD per descrivere l'intero processo strutturato di estrazione della conoscenza e *Data Mining* come la fase di applicazione all'interno del processo di uno specifico algoritmo per l'individuazione dei *pattern*.

Il processo di scoperta della conoscenza è strettamente legato allo sviluppi di sistemi di *data warehousing* (finalizzati a mantenere collezioni di dati integrati, che si evolvono nel tempo, accessibili secondo viste diverse e utilizzati per processi di *decision support*) e può essere in qualche modo considerato una sua estensione, una sua evoluzione. Lo sviluppo di un sistema di Data Mining dovrebbe avvenire partendo da un *Data Warehouse* (DWH) già implementato, o comunque da ambienti certificati dove i dati siano stati regolarizzati, in modo che l'analisi possa essere fatta su dati accurati, integri e omogenei, il più possibile *ripuliti* da incongruenze che potrebbero influenzare la precisione dei risultati.

Il Text Mining è una specializzazione del Data Mining applicata al contesto di collezioni di testi non strutturati (agenzie stampa, pagine web, *e-mail*, libri e articoli in versione digitale, e più in generale a qualsiasi *corpus* di documenti). Questo tipo di applicazione è particolarmente utile per individuare gruppi tematici, classificare i documenti, scoprire associazioni nascoste (legami tra argomenti o tra autori, *trend* temporali, ecc.), addestrare motori di ricerca, estrarre concetti per la creazione di ontologie (*ontology learning*).

Il processo di estrazione della conoscenza (KDD)

Il processo di estrazione della conoscenza è un processo interattivo ed iterativo e, indipendentemente dal tipo di applicazione specifica, può essere schematizzato attraverso le seguenti fasi:

1. Definizione dell'obiettivo;
2. Selezione dei dati;
3. *Pre-processing* (Pulizia dei dati);
4. Trasformazione e Riduzione;
5. Data Mining;
6. Interpretazione/valutazione dei risultati;
7. Rappresentazione dei risultati.

La natura iterativa del processo viene evidenziata dallo schema seguente che mette in evidenza come la fase di valutazione può portare da una semplice ridefinizione dei parametri di analisi utilizzati ad una ridefinizione dell'intero processo (a partire dai dati estratti).

Figura 1. Il processo KDD[1].

[1] <http://www.cineca.it>.

All'interno del processo di estrazione della conoscenza alcune funzioni, come l'individuazione dei *pattern*, possono essere svolte automaticamente, altre richiedono il coinvolgimento di varie professionalità (esperti del dominio applicativo, specialisti in analisi dati, informatici). In generale, comunque, l'intero processo presuppone il monitoraggio da parte di esperti, anche ai fini dell'individuazione delle tecniche di estrazione di conoscenza più efficaci nel contesto considerato, oltre che dell'eventuale *tuning* (ottimizzazione) di tali tecniche.

Definizione degli obiettivi

Nel processo di estrazione della conoscenza, per ottenere i migliori risultati è importante comprendere al meglio il dominio applicativo di riferimento ed individuare chiaramente gli obiettivi che si intendono raggiungere attraverso tale attività.

La perfetta definizione di tali obiettivi è forse la fase più delicata del processo e fondamentale per la riuscita dell'intero progetto. Condizione essenziale è parlare con i cosiddetti esperti del settore (esperti di dominio) per capire con esattezza cosa va fatto e quali sono le esigenze da soddisfare. Si può ad esempio voler incrementare il numero di possibili clienti che risponderanno positivamente ad una campagna promozionale. Ciò si traduce in due differenti obiettivi: *incrementare il tasso di risposta* e *incrementare il valore di una risposta*. Per essi sono necessari due modelli completamente diversi.

Selezione dei dati

Il secondo passo del processo consiste nella selezione dei dati su cui sarà attivato il processo di scoperta. In questa fase risulta particolarmente importante individuare le fonti da cui reperire i dati necessari per il raggiungimento degli obiettivi precedentemente definiti e valutare l'eventuale ambiente informatico che ne consenta la gestione. L'integrazione dei dati provenienti da fonti di dati diverse – che possono essere interne o esterne, in funzio-

ne della necessità di utilizzare nuovi elementi non presenti nel sistema informativo – deve portare alla creazione di una matrice di dati (*data set*) disegnata coerentemente con il modello dati definito sulla base degli obiettivi individuati.

Un ambiente che favorisce l'integrazione dei dati è rappresentato da un sistema di Data Warehouse. Esso fornisce un unico punto di accesso a dati di tipo operativo, che possono avere un immenso valore strategico nel processo decisionale se opportunamente organizzati e certificati. Se non si dispone di un Data Warehouse, si dovrà ricorrere all'utilizzo di alcuni *tool* che offrono la possibilità di pre-elaborare i dati e prepararli per la fase successiva. Da sottolineare comunque che non è possibile utilizzare un Data Warehouse direttamente per il Data Mining, poiché è necessario un gran lavoro di preparazione dei dati che devono essere utilizzati per le analisi.

Pre-processing (Pulizia dei dati)

Una volta selezionati i dati dalle varie fonti interne ed esterne ed organizzati in un unico contenitore (la matrice di dati) si rende necessario valutarne la qualità e procedere ad una attenta preparazione dei dati stessi.

Poiché i metodi di Data Mining si pongono l'obiettivo di cogliere caratteristiche presenti nei dati e non note a priori, è molto importante individuare eventuali contaminazioni e cercare di porvi rimedio, per non correre il rischio di trarre conclusioni che siano basate su caratteristiche che derivano da vizi nella raccolta o nella registrazione.

È particolarmente applicabile in questo ambito il detto *Garbage In, Garbage Out*, che sta a significare che se si ha della spazzatura in ingresso al sistema non si può che ottenere della spazzatura in uscita.

È quindi necessario effettuare uno studio preliminare che consenta di identificare le caratteristiche dei dati ed eliminare le eventuali contaminazioni presenti al fine di eseguire la fase di

Data Mining in modo effettivo ed efficiente. Solitamente si individuano tre categorie di problemi:

- *valori mancanti*: la presenza di dati mancanti può essere affrontata in modi diversi. Nella maggior parte dei casi essi indicano informazioni perse o non inserite all'origine. Alcune tecniche consentono di trattare queste situazioni scartando i *record* contenenti valori mancanti oppure mediante sostituzione di tali valori con valori calcolati (es. la media dei valori delle altre osservazioni corrispondenti) o rilevati per altre osservazioni molto simili. Queste soluzioni, tuttavia, potrebbero non essere coerenti con gli obiettivi definiti;
- *dati anomali*: rappresentano errori casuali nei valori delle variabili. Quando si analizzano grandi *data set* tali errori possono assumere svariate forme e tipologie. Possono essere presenti dati replicati (*record* doppi), valori errati (non ammissibili) e/o valori che si discostano moltissimo dagli altri valori, collocandosi al di fuori dell'intervallo atteso, e che, se considerati nelle analisi, possono modificare significativamente i risultati. Purtroppo non si hanno delle tecniche precise per la loro gestione;
- *dati incerti*: la precisione dei dati è di particolare rilevanza nei sistemi di scoperta della conoscenza. Se i dati si trovano in una situazione molto grave di incertezza, la migliore soluzione è riciclare sull'individuazione delle fonti per ottenerne di migliori.

Come già detto la *pulizia* dei dati, in molti casi, può essere fatta automaticamente. Non è tuttavia realistico pensare di essere capaci di rimuovere tutte le contaminazioni in un'unica fase: alcune anomalie nei dati potranno essere scoperte solo durante il processo di Data Mining stesso. Ciò dimostra la natura iterativa del processo.

Trasformazione e Riduzione

Un trattamento dei dati particolare è richiesto per quegli strumenti di Data Mining che possono accettare solamente specifici valori in *input*. Ad esempio, tecniche come le reti neurali o come quelle utilizzate per estrarre regole di associazione – che verranno descritte in seguito – riescono a lavorare meglio se i dati hanno un valore numerico compreso tra *0* e *1*. In questo caso, devono essere effettuate operazioni di normalizzazione dei dati, facendo sì che i valori numerici cadano all'interno di un *range* specificato, e di conversione dei dati, trasformando i dati categorici in dati numerici.

Quando il numero di variabili o caratteristiche che descrivono i dati da analizzare è particolarmente elevato può essere necessario applicare alcune tecniche (dette di *feature selection*) che consentono di selezionare solo quelle caratteristiche che sono maggiormente utili e rilevanti nel processo di Data Mining, scartando quelle tendenzialmente più ridondanti e/o insignificanti. Il processo di Data Mining, infatti, tende a risentire in negativo, sia in termini di efficienza che di qualità dei risultati, della presenza di un eccessivo numero di *feature*: ciò è particolarmente vero per tutti i metodi di tipo statistico applicati nell'ambito di tale processo, che risentono negativamente della elevata sparsità dei dati derivante dal fare riferimento a spazi ad elevata dimensione (in cui ciascuna dimensione corrisponde ad una caratteristica).

Inoltre, la riduzione del numero delle variabili, oltre a portare ad un miglioramento delle prestazioni, può facilitare la visualizzazione dei dati e favorire la selezione del modello da utilizzare. L'attività di identificazione delle variabili dipendenti (attributi di output), indipendenti (attributi di input) e correlate fra loro viene anche indicata col termine di *data pruning*.

Data Mining

È la fase del processo KDD in cui vengono applicati iterativamente particolari metodi per estrarre conoscenza ovvero cer-

care interessanti *pattern*, regole, o sequenze ripetute all'interno di grandi quantità di dati.

La conoscenza estratta da una sessione di Data Mining fa tipicamente riferimento ad un modello, che comprende le informazioni e le scoperte durante il processo di mining. In pratica, è una topologia delle relazioni che mappa quali condizioni di *input* hanno influenza su specifiche condizioni di *output* e come alcune relazioni possono influenzare significativamente altre relazioni. Il modello costituisce la rappresentazione formale, tipicamente matematica, degli aspetti fondamentali di un fenomeno e ne riproduce le caratteristiche essenziali: nel nostro caso spiega l'effetto degli *input* sugli *output*. Come tale, può essere raffinato e messo a punto per una maggiore precisione attraverso un processo iterativo di comprensione dei dati che sono alla sua base.

Scelta dei metodi

Il primo passo di questa fase consiste nell'individuazione di quei metodi che meglio possono portare agli obiettivi che si vogliono raggiungere e che meglio si adattano al tipo di dati da analizzare all'interno di una strategia di estrazione. Esistono numerose tecniche (metodi) di Data Mining. Tuttavia, tutti i metodi di Data Mining applicano l'apprendimento basato sull'induzione. Questo è un processo che genera definizioni di concetti generali sull'osservazione di esempi specifici.

Le strategie di Data Mining possono essere classificate come strategie supervisionate e strategie non supervisionate. Il Data Mining supervisionato è un approccio *top down* applicabile quando si sa cosa si sta cercando e assume spesso la forma di modelli previsionali. In pratica vengono costruiti modelli tramite l'utilizzo di attributi di *input* per predire i valori di attributi di *output*. In tale ambito ricadono le tecniche di regressione (previsione del valore di uno o più attributi numerici sulla base dei va-

lori degli attributi di input) e di classificazione (previsione della classe di appartenenza di un elemento, all'interno di una partizione, in funzione dei valori degli attributi). Le tecniche supervisionate fanno invariabilmente uso di insiemi di *esempi* (*training set*): gruppi di elementi per i quali i valori degli attributi di output (numerici o booleani/discreti, come nel caso della classificazione) sono noti. L'uso di algoritmi di *addestramento* (*training*) permette di indurre da tali esempi leggi generali che governano la relazione tra attributi di input e di output.

Le strategie di apprendimento supervisionato possono essere ulteriormente classificate in base al fatto che i modelli siano progettati per descrivere una condizione corrente o per predire una condizione futura.

Il Data Mining non supervisionato è un approccio bottom up, vale a dire in cui si lascia che i dati stessi indichino un risultato. Non esiste un attributo di *output*: tutti gli attributi utilizzati per costruire i modelli sono variabili indipendenti. Nell'approccio non supervisionato, si fa tipicamente l'ipotesi che esista un modello (spesso generativo) dell'insieme dei dati che descrive con la sua struttura l'informazione da estrarre con il processo di mining. Un esempio significativo è offerto dai modelli generativi: in questo caso si assume che l'intero insieme dei dati sia stato generato in modo stocastico sulla base di un modello statistico (una o più distribuzioni di probabilità) di cui viene definita la struttura, ma non i relativi parametri, che sono associati all'informazione da estrarre. Il processo di mining viene a coincidere allora con l'inferenza statistica dei parametri del modello dai dati[2].

Naturalmente, essendo la scelta del tipo di modello, o in generale della tecnica da utilizzare, sostanzialmente arbitraria, sarà

[2] L'inferenza statistica è un procedimento attraverso il quale è possibile indurre le caratteristiche di una popolazione, intesa in senso statistico come un insieme di elementi oggetto di indagine, a partire da una sua parte, detta campione.

necessario, una volta estratta l'informazione sulla base di tale scelta, valutare in qualche modo la qualità dei risultati ottenuti, eventualmente confrontandola con quella raggiunta mediante l'applicazione di tecniche e modelli diversi.

Una strategia di Data Mining si applica a un insieme di dati utilizzando uno o più metodi. Una tecnica specifica di Data Mining è definita da un algoritmo e da un modello, inteso come struttura di conoscenze associata (es. alberi, insieme di regole, distribuzioni di probabilità). I metodi sono applicabili a qualsiasi ambito di indagine, tuttavia la scelta di quale metodo utilizzare nella fase di analisi dipende essenzialmente dal tipo di problema oggetto di studio e dal tipo di dati disponibili per l'analisi.

Sulla base di tali considerazioni, per orientarsi meglio nella scelta dei metodi, alcuni dei quali saranno descritti in seguito, si preferisce suddividere gli stessi in quattro grandi classi, come evidenziato in (Giudici, 2004)[3]:

- *Metodi esplorativi*, che presentano forti analogie con le tecniche di tipo OLAP (*On Line Analytical Processing*) e con i metodi propri dell'analisi dei dati (*query tools*). Essi si basano su metodologie interattive e, solitamente, visuali, che hanno lo scopo di trarre le prime conclusioni ipotetiche dalla massa di dati disponibili, oltre che fornire indicazioni su eventuali trasformazioni degli stessi;
- *Metodi descrittivi* (detti anche non supervisionati), basati su modelli simmetrici, privi di ipotesi di casualità, che hanno lo scopo di descrivere l'insieme dei dati in un modo parsimonioso. Attraverso tali metodi è possibile sintetizzare delle osservazioni che vengono pertanto classificate in gruppi non noti a priori (analisi di *cluster*[4], mappe di

[3] PAOLO GIUDICI, *Data mining e statistica*, in «Statistica & Società», vol. 3, n. 1, 2004, pp. 5-6.

[4] Per cluster si intendono raggruppamenti di oggetti omogenei in termini di similarità delle relative caratteristiche.

Kohonen) o delle variabili, che vengono tra loro relazionate, secondo legami non noti a priori (modelli log-lineari, modelli grafici);
- *Metodi previsivi* (denominati anche supervisionati): gruppo di metodologie con l'obiettivo di spiegare una o più variabili in funzione di tutte le altre, ricercando, sulla base dei dati, delle regole di segmentazione e di valorizzazione (*scoring*) delle osservazioni;
- *Metodi locali*, per i quali l'obiettivo dell'analisi non è, come nei casi precedenti, la descrizione delle caratteristiche del database nel suo complesso (analisi globale), ma l'individuazione di caratteristiche peculiari, relative a sottoinsiemi di interesse del database (analisi locale).

Interpretazione, valutazione dei risultati

Lo scopo dell'interpretazione e della valutazione è determinare la validità di un modello e la relativa applicabilità a problemi esterni all'ambito del test. Se si ottengono risultati accettabili si trasforma la conoscenza acquisita in termini comprensibili dagli utenti. La fase di interpretazione, attraverso un processo iterativo, può suggerire di ritornare ai passi precedenti per ulteriori attività di raffinamento. In questa fase si può ricorrere a tecniche di visualizzazione per analizzare i modelli estratti.

Una volta che il modello è stato creato e raffinato ad un accettabile grado di accuratezza, esso può essere usato in due modi:

1. in modo *descrittivo*, permettendo agli utenti finali di studiare le relazioni scoperte tra i dati per migliorare la comprensione dei fattori chiave che influenzano il *business*;
2. in modo *predittivo*, determinando la più probabile condizione di output associata agli *input* forniti.

Rappresentazione dei risultati

L'ultimo obiettivo del processo di estrazione della conoscenza consiste nell'utilizzo di ciò che è stato appreso attraverso la

sua applicazione ad altri sistemi per le azioni del caso.

L'attività di *reporting* dei risultati è particolarmente importante. Essa può assumere diverse forme. In genere è possibile usare un qualsiasi *report writer* o *tool* grafico per rendere accessibili i risultati del processo.

L'attività di *reporting* si sviluppa attraverso due differenti funzioni:

- analisi dei risultati degli algoritmi di *pattern recognition* (riconoscimento di forme);
- applicazione dei risultati dell'algoritmo di *pattern recognition* a dati nuovi.

Non si vuole solamente esaminare ciò che si è appreso, ma si vorrebbero applicare le informazioni classificate e segmentate che si sono ottenute. In molti casi il *reporting* può essere effettuato usando i tradizionali *query tools* per database, tuttavia, si stanno affermando nuove tecniche.

Data Mining: Modellistica e Tecnologie

La fase di *scoperta* ed estrazione della conoscenza da un coacervo amplissimo e multiforme di informazioni variamente raccolte e variamente immagazzinate è certamente il momento più affascinante ed importante dell'intero processo di KDD. In questa fase emerge come fondamentale il Data Mining che, a sua volta, oggi è sinonimo di un insieme multiforme di algoritmi, anche indicati con il termine di metodologie dal momento che si fa riferimento al metodo generale cui l'algoritmo si richiama, ognuno dei quali garantisce l'ottenimento di specifici obiettivi significativi anche in ambiti più generali quali:

- *il machine learning;*
- *la pattern recognition;*
- *il data cleansing;*
- ecc.

Come sopra accennato il Data Mining è un mondo composito nel quale diverse famiglie di algoritmi di analisi lavorano in modo differente sull'intera massa di dati e colgono di volta in volta obiettivi specifici e parzialmente differenti l'uno dall'altro. Le singole estrazioni di dati ottenute tramite l'applicazione di una data metodologia possono essere integrate con estrazioni derivate dall'applicazione di una metodologia differente e così modificate possono far da base all'applicazione di una terza metodologia. In questa maniera si viene a creare quella che può essere definita come una vera e propria catena di scoperta della conoscenza.

All'interno delle metodologie riteniamo interessante analizzare in modo particolare:

- *i query tool;*
- la visualizzazione;
- gli strumenti OLAP;
- I K primi vicini *(the k-nearest neighbors)*;
- le regole di associazione;
- gli alberi di decisione;
- le reti neurali;
- gli algoritmi genetici.

I query tool

Questi sono gli strumenti più semplici e dovrebbero essere usati come primi anelli di quella catena di scoperta della conoscenza cui prima si accennava. Nei fatti, come quando esplorando per la prima volta una landa sconosciuta si cercano i riferimenti fondamentali, così nella ricerca della conoscenza nascosta in un insieme multiforme di dati si inizia, laddove la struttura del DWH lo permetta, con delle *query* SQL dirette con cui si estraggono grosse masse di dati che rispondono a livello macro agli obiettivi finali che ci si proponeva. In questo modo si arriva a scoprire l'80% circa dell'informazione che si cerca. Il restante 20%, che peraltro potrebbe essere di natura vitale per il raggiun-

gimento dell'obiettivo finale, rimane nascosto e deve essere raggiunto tramite tecniche ed algoritmi più avanzati.

Nonostante queste limitazioni, come primo strumento di ricerca, i *query tool* permettono di esplorare il *terreno* in maniera sicura, evidenziando dettagli primari e strutture generali di riferimento, elementi questi che serviranno di base per l'applicazione di strumenti di ricerca via via più sofisticati.

La Visualizzazione

Un altro strumento, peraltro più moderno dei *query tool*, di prima esplorazione dell'insieme di dati per trovare *pattern*, ovvero regolarità nella struttura dei dati stessa, è la visualizzazione.

Per capire la natura dei processi di visualizzazione poniamoci di fronte ad un classico problema di *marketing*: capire chi è propenso ad acquisire un dato bene. Per far ciò possiamo farci aiutare creando un semplice grafico cartesiano di *data scatter*, funzione di grafica ormai standard anche in Excel. I passi che portano dal nostro insieme di dati ad una loro rappresentazione in un *data scatter diagram* significativo per il nostro proposito finale sono i seguenti:

- Si individua nel database quell'insieme di persone per cui sono noti due differenti attributi di natura numerica, per esempio le persone di cui si conoscono al contempo l'età e l'entità delle loro entrate personali annue;
- Per le persone individuate in questo insieme si individua, sempreché detta informazione sia presente nel database, il valore della caratteristica che si vuole investigare, i.e. la propensione o meno ad acquistare un dato bene;
- Si rappresentano le singole persone come punti in un piano cartesiano in cui l'ascissa rappresenta l'età e l'ordinata rappresenta l'entità delle entrate personali annue. Le singole persone trasformate in punti si posizionano nello spazio cartesiano a due dimensioni in funzione della loro specifica età ed entità delle entrate;

- Si può dare ai punti/persona un colore differente a seconda che la singola persona rappresentata sia propensa o non sia propensa ad acquisire il bene su cui si sta svolgendo l'indagine di mercato.

Se il quadro grafico finale vedrà punti di colore differente distribuirsi sul piano in maniera assolutamente casuale e sovrapposta, ciò vorrà dire che età ed entità annue delle entrate personali non sono due variabili correlate all'acquisizione del bene voluto.

Laddove, invece, emergano insiemi definiti di punti di ugual colore confinati a diverse zone dello spazio cartesiano rappresentato, significa che età e entità delle entrate sono correlabili alla propensione delle persone ad acquisire il prodotto in questione e laddove il nostro database contenga un campione sufficientemente ampio di persone possiamo ragionevolmente pensare di applicare la relazione individuata (la propensione o meno ad acquisire un dato prodotto in funzione di età e reddito) a persone non contenute nel campione, ma che possiedano le stesse caratteristiche.

Il processo può essere esteso a coppie di caratteristiche numeriche delle persone differenti da età ed entità delle entrate (ad esempio distanza dell'abitazione dal centro commerciale e numero di figli di età inferiore ai diciotto anni presenti in famiglia), sempre che detti attributi abbiano la natura di un indice numerico continuo, per poi fare considerazioni analoghe a quelle sopra descritte.

Si può passare alla grafica tridimensionale esplorando triplette di caratteristiche e si può utilizzare una tavolozza di colori o sfumature per rappresentare i singoli punti in funzione della caratteristica che si sta investigando (ad esempio rappresentare in una scala a quattro colori la propensione ad acquisire un dato bene, sempre laddove questa informazione numerica sia presente nel nostro database), ecc.

Da quanto sopra descritto si capisce come la visualizzazione rappresenti uno strumento di primo intervento di esplorazione del nostro spazio dei dati e come sostituisca l'analogo ma più an-

tico strumento dei *query tool*.

L'accoppiamento tecnologico di grafica avanzata e *query* sta negli ultimi anni ricevendo un grande impulso ed è solo recentemente che sono arrivati sul mercato strumenti quali *l'object oriented three-dimensional toolkit*, ovvero la grafica tridimensionale interattiva con la quale si rappresentano in tre dimensioni i risultati di *query* che individuano oggetti con tre attributi numerici che servono per l'appunto a definire le tre coordinate spaziali dell'oggetto da rappresentare.

Una volta passati dal dato alla sua posizione in uno spazio (bi- o tridimensionale) è possibile assumere che punti vicini l'uno all'altro siano simili ed a volte è possibile individuare dei *pattern* interessanti, ovvero insiemi significativi di punti, anche con un semplice esame visivo.

A volte, invece, l'individuazione di *nuvole* di punti che emergono dal tutto in modo statisticamente significativo deve avvenire mediante algoritmi specifici. Ricordiamoci a questo punto di cosa succede se vogliamo esplorare in via grafica oggetti in funzione di un numero di caratteristiche superiori a tre. È lo stesso problema che si ha se si vuole rappresentare graficamente una funzione di quattro variabili:

$$f(x,y,z,t).$$

Per farlo in maniera completa ed assumendo che si possa rappresentare in via grafica solo una funzione di al massimo tre coordinate (x,y,z), si devono fissare valori successivi di t. Per esempio porre $t=1$ e rappresentare interamente la $f(x,y,z,1)$. Quindi porre $t=2$ e rappresentare interamente la $f(x,y,z,2)$.

Si può ripetere la stessa procedura, esplorando gradualmente l'intero intervallo nell'ambito del quale la variabile t è compresa e così si arriva a rappresentare interamente la funzione a quattro variabili.

La stessa struttura di azione si può adottare nel nostro spazio dei dati, solo che gli attributi (i.e. il numero di variabili indipen-

denti della nostra funzione) in base ai quali dobbiamo esplorare lo spazio a volte sono molti di più di quattro e ciò pone il problema di scelta del *taglio* rispetto al quale rappresentiamo la funzione: i.e. la scelta di quali attributi svolgono il ruolo di *x,y,z* e di quali invece svolgono il ruolo della *t* fissata, di volta in volta, ad un dato valore.

L'OLAP

Abbiamo visto come la visualizzazione ci abbia naturalmente portato a considerare il concetto di dimensionalità del dato; nei fatti una tabella costituita da *n* colonne che contengono ognuna un attributo indipendente dagli altri e da una ulteriore, *n+1-esima*, colonna in cui è contenuto un dato che possiamo considerare una variabile dipendente dagli *n* precedenti attributi può essere vista alla stregua di una funzione in uno spazio a *n* dimensioni.

Se ci si pensa, i *manager* in genere fanno sempre domande per la cui risposta è necessario analizzare un problema a *n* dimensioni. In genere essi non si limitano a chiedere solamente quanto è stato venduto nel complesso il mese scorso (la domanda a dimensione zero), ma anche che tipo di rivista è stata venduta ogni mese, per regione, e che età aveva l'acquirente (domanda a quattro dimensioni). Quindi, la vendita viene analizzata in funzione di: prodotto (tipo di rivista), tempo (ogni mese), zona geografica (per regione), età dell'acquirente.

Questo tipo di informazione è per propria natura multidimensionale e le relazioni in questi contesti sono difficili ad individuarsi laddove l'unico strumento di rappresentazione ed indagine siano tabelle a due dimensioni. Peraltro, quando dobbiamo trovare le relazioni fra più dimensioni nemmeno i database relazionali standard ci aiutano (i.e i tradizionali *query tool*). Nei fatti, l'SQL identifica i *record* tramite chiavi[5] e il concetto di mul-

[5] In un database, la chiave primaria è costituita da quegli attributi che permettono di identificare ogni istanza in maniera univoca.

tidimensionalità si traduce in quello di chiavi multiple per singolo *record*; purtroppo c'è un limite al numero di chiavi che in database relazionali possono essere associate ad una singola tabella mentre non vi è un limite alla fantasia del *manager* nel chiedere dati aggregati per insiemi di dimensioni ogni volta differenti.

Il *manager* ad un dato tempo *t* chiede i dati delle vendite in funzione di area geografica, età dell'acquirente e relative entrate annue, al tempo *t* più un minuto lo stesso *manager* chiede sempre i dati di vendita ma questa volta in funzione di carta di credito ed età, possibilmente interrogando online il database di gestione delle casse dei supermercati della compagnia.

Si può ricorrere ai cosiddetti *tool* OLAP. Si tratta di strumenti che per prima cosa prelevano i dati contenuti nel database e poi li memorizzano in un formato intrinsecamente multidimensionale; ciò fatto possono rispondere a qualsivoglia stravagante domanda venga eventualmente fatta dal *manager* di cui sopra. È ovvio che, proprio perché è avvenuto un processo di copia dal database originale al database immagine in memoria, si possono perdere i cambiamenti che in quello stesso istante avvengono nel database di origine. Si deve, perciò, con una certa frequenza, aggiornare la copia in memoria al fine di avere dati il più possibile freschi. Sarebbe peraltro infattibile lavorare direttamente sul database origine perché i *tool* OLAP lancerebbero *query* così complesse da *far sedere* DBMS (*Data Base Management Systems*) anche potenti ed attrezzati, con grave degrado delle prestazioni gestionali e conseguente allungamento delle file di clienti alle casse del supermercato sopra citato.

I *tool* OLAP rivestono un ruolo importante nel Data Mining anche se sono caratterizzati da un limite intrinseco importante: non imparano, essi non creano nuove conoscenze e non possono ricercarne di nuove. Forniscono dati in risposta a domande complesse ma non offrono o individuano interpretazioni.

I K primi vicini

Con questa notazione un po' criptica si indica una famiglia di algoritmi che ancora una volta trasportano in uno spazio cartesiano i nostri dati e ci fa arrivare a delle conclusioni sulla base della vicinanza dei punti.

Per fare un esempio pratico si pensi a due differenti oggetti che possiedono i valori di tre attributi numerici molto simili fra di loro. Una volta che detti attributi siano tradotti in coordinate e che tramite questa trasformazione essi possano venire rappresentati in uno spazio cartesiano a tre dimensioni, i due punti di cui sopra verranno rappresentati uno molto vicino all'altro.

Rifacciamoci, estendendolo, all'esempio già fatto con due coordinate: prendiamo due persone che sono ugualmente propense ad acquisire un bene e ci accorgiamo che tre attributi che caratterizzano queste due persone (età, entità delle loro entrate annue e quantità del loro indebitamento con terzi) sono molto simili. Ecco che se adesso rappresentiamo le due persone in uno spazio cartesiano a tre dimensioni i punti, che per l'appunto rappresentano le due persone, appariranno l'uno vicino dell'altro.

Abbiamo così definito il concetto di *vicinato* (*neighborhood*): *record* che, una volta rappresentati nello spazio cartesiano adottando la metodologia suddetta, appaiano l'uno vicino all'altro si dice che appartengono allo stesso vicinato.

Poniamoci ora il problema di voler predire il comportamento di un insieme di persone relativamente, per esempio, alla possibile acquisizione di un bene, e per far ciò abbiamo a disposizione un database in cui ciascun potenziale cliente è caratterizzato da una serie di attributi (ad esempio età, entità delle sue entrate annue e quantità del suo indebitamento con terzi ecc.).

L'ipotesi base che guida la nostra predizione è che persone dello stesso tipo è probabile si comportino in maniera simile rispetto ad una eventuale acquisizione del bene. Nella nostra traduzione *attributi => coordinate*, l'appartenere ad uno stesso tipo può essere tradotto nell'appartenere ad una stessa regione del-

lo spazio, *id est* appartenere allo stesso vicinato. Basandoci su questa semplice constatazione è possibile costruire un potente algoritmo di autoapprendimento denominato *i K primi vicini (the k-nearest neighbors)*.

La filosofia alla base dell'algoritmo è la seguente: in genere una persona fa ciò che vede fare al suo vicino. In altre parole, laddove si voglia attribuire una misura ad un dato comportamento di un punto/persona (i.e. la propensione all'acquisizione di un dato bene), questa viene calcolata facendo la media della misura effettiva di quello stesso comportamento nei 10 punti/persona che sono più vicini nel *data space* (i.e. nello spazio cartesiano nel quale rappresentiamo i punti persona) al punto in questione.

La lettera *k* nel nome dell'algoritmo *i k primi vicini* sta ad indicare quanti punti (*k* per l'appunto) si prendono per calcolare la proprietà voluta sulla base della media delle proprietà effettive (i.e. sperimentali, ovvero i valori naturali di quella proprietà contenuti nel database originale). Nell'esempio sopra fatto si lavora con un algoritmo a 10 primi vicini.

Il principale vantaggio della presente metodologia è il seguente: è una tecnica di ricerca che ha come proprio riferimento interno lo stesso insieme di dati su cui si effettua la ricerca. Peraltro gli svantaggi, soprattutto dal punto di vista della complessità numerica, sono non pochi e fra di essi maggiormente si rilevano i seguenti:

- In un insieme di oggetti *n*, per poter decidere quali sono i *k* primi vicini di un dato oggetto bisogna misurare la distanza fra l'oggetto preso in considerazione e gli altri n-1 oggetti che fanno parte dell'insieme. Ciò porta a dover calcolare n-1 distanze per ogni oggetto dell'insieme e di conseguenza circa n^2 distanze se si deve prendere in considerazione l'intero insieme. La complessità quadratica[6] è mol-

[6] La complessità misura il tempo di esecuzione di un algoritmo in base ai

to pesante; basti pensare che per un insieme di un milione di oggetti si devono fare mille miliardi di operazioni di definizione della distanza. In genere la complessità degli algoritmi migliori nel Data Mining è dell'ordine $n*log(n)$[7] dove *n* è il numero di *record* o oggetti nel database. Proprio per questo motivo si tende ad usare questa metodologia su piccoli sottoinsiemi degli oggetti da analizzare (i.e. per indagini di natura cosiddetta locale);

- Il numero di attributi di un oggetto e, quindi, il numero di *dimensioni*, è spesso molto più alto di tre. Lavorare in uno spazio altamente multidimensionale porta un elevato grado di complessità poiché esso, fra l'altro, elude le nostre percezioni abituate a lavorare solo in tre dimensioni. Per esempio in tre dimensioni un milione di punti affollano lo spazio occupabile in maniera tale da avere punti non molto distanti gli uni dagli altri. Mentre quello stesso milione di punti occuperebbe uno spazio a 20 dimensioni con una densità molto bassa e sarebbe difficile parlare di vicinanza di punti e quindi di similarità di comportamento di punti/individui.

Ancora una volta si riconferma il seguente principio: non esiste algoritmo di Data Mining in assoluto migliore di tutti gli altri e per ciò stesso tale da far con sicurezza trovare ciò che si cerca. La realtà è la seguente: si devono di volta in volta usare algoritmi specifici per ottenere dei risultati parziali nell'ambito di quella catena di scoperta della conoscenza sopra citata.

valori di input. Nel caso della complessità quadratica il numero delle istruzioni da eseguire è proporzionale al quadrato delle dimensioni dell'input.

7 In questa classe di complessità la crescita del numero di istruzioni eseguite rispetto alla dimensione dell'input è poco più che lineare, quindi il tempo di esecuzione è inferiore rispetto alla complessità quadratica.

Gli alberi di decisione

Per descrivere questo algoritmo partiamo, come nei casi precedenti, pensando di dover trattare una tabella nella quale è data la propensione di *n* individui ad acquisire un bene, l'automobile ad esempio, ed al contempo l'individuo è caratterizzato da tre ulteriori attributi, (i.e. dati numerici continui), in questo caso l'età, le entrate annue e la capacità di ricevere credito. Il tentativo di predire un certo tipo di comportamento è sempre basato sul concetto che se un individuo appartiene ad un gruppo esso si comporterà in genere come il gruppo cui appartiene.

Per fare la predizione voluta ci si chiede, in prima istanza, se dei tre attributi a disposizione l'età sia il fattore determinante nel caratterizzare il comportamento degli individui dell'insieme rispetto alla loro disponibilità ad acquisire una macchina. Questo fatto di per sé implica che solamente conoscendo l'età degli individui del campione si è in grado di predire se un individuo sia propenso o meno ad acquisire una macchina. Laddove ciò fosse vero, è possibile suddividere il nostro insieme di individui in due distinti sottoinsiemi facendo uso della sola età. Si deve investigare se esista o meno un'età limite che possa dividere i possibili compratori dai probabili non compratori di macchine. Analizzato l'insieme rispetto al primo degli attributi a disposizione ed eventualmente rilevata la presenza di questa età limite, si passa al secondo degli attributi a disposizione e si ripete rispetto ad esso la stessa analisi sopra descritta. Analizzato l'insieme rispetto al secondo attributo a disposizione ed, anche in questo caso, rilevata l'eventuale presenza di entrate annue limite con cui si può dividere l'insieme, si passa al terzo attributo e si fanno le stesse domande.

Il suddetto tipo di analisi può essere fatto a cascata, cioè si può analizzare rispetto al secondo attributo, le entrate annue, il sottoinsieme di persone propense ad acquisire una macchina derivante dall'analisi rispetto al primo attributo, l'età, e così via. Alla fine di questo processo iterativo, attributo per attributo si ar-

riva a creare il cosiddetto albero delle decisioni, come da Figura 2, per l'insieme oggetto di indagine in relazione ad una data proprietà, in questo caso la propensione ad acquistare una macchina.

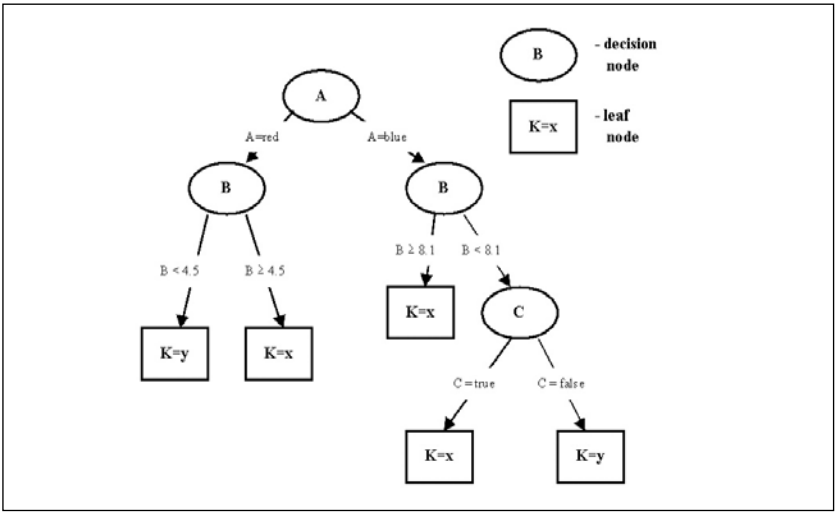


Figura 2. Un semplice albero delle decisioni[8].

Vi sono molti algoritmi e, quindi, molti programmi che eseguono i *task* sopra descritti con grande efficienza anche perché la complessità degli algoritmi stessi è proporzionale a n*log(n) in cui *n* è il numero di oggetti all'interno dell'insieme da analizzare. Stante questa situazione gli algoritmi di strutturazione di un albero di decisione lavorano molto bene in insiemi molto popolati (al contrario di quanto accadeva per gli algoritmi dei *k* primi vicini a causa della loro dipendenza da n^2). Un vantaggio ulteriore dei risultati ottenuti con questa metodologia è che essa è di diretta e naturale comprensione.

8 <http://dms.irb.hr/tutorial/tut_dtrees.php>.

Le regole associative

In genere ai *manager* interessano regole semplici che presumono siano dettate dal mercato, come ad esempio: al 90% delle signore che guidano macchine *spider* rosse e possiedono un cagnolino di piccola taglia piace lo Chanel n° 5. Questo tipo di regole affascina poiché dà un chiaro profilo del cliente sul quale finalizzare una precisa strategia di promozione. Alcune delle metodologie di Data Mining, quelle soprannominate regole associative, premettono di estrarre da grandi basi dati regole come quelle sopra descritte.

Si lavori per esempio su un database nel quale per ogni individuo è memorizzato il genere, il tipo e il colore della macchina, la razza del cane posseduto, e la disponibilità o meno di una serie di beni o di preferenze (l'uso o non uso del profumo, tipo di profumo ecc.). La regola che interessa tanto ai manager letta dal suddetto database si traduce in: il 90% dei *record* in cui il genere è femminile, la macchina è una *spider*, il colore della macchina è rosso, il cane è di taglia piccola, contiene nella colonna profumo la voce Chanel n° 5.

Per prima cosa bisogna dire che le regole associative si trovano in database in cui gli attributi hanno valore binario (sì-no, compra-non compra, alto-basso, ecc.) e qualora gli attributi fossero di natura numerica questi ultimi dovrebbero essere *appiattiti* fino ad assumere una forma binaria; se ad esempio uno degli attributi fosse l'età ed essa avesse una distribuzione nell'insieme di interesse fra i trenta ed i settanta anni dovremmo, prima di applicare un algoritmo associativo, tramutare l'età nel parametro binario *sopra i cinquanta: sì/no*. Una volta applicata questa trasformazione vi sono molti algoritmi che possono ricavare associazioni come quelle sopra descritte e dette associazioni possono comprendere un attributo, due attributi, tre attributi ecc.

È però anche vero che il numero di regole associative cresce esponenzialmente con il numero di differenti attributi che compongono il database oggetto di investigazione. Peraltro non è dif-

ficile trovare regole associative quanto definire se dette regole abbiano o meno un qualche valore. Per esempio nel caso precedente è vero che al 90% delle signore che guidano macchine *spider* rosse e possiedono un cagnolino di piccola taglia piace lo Chanel n°5, ma può risultare anche vero che il numero delle signore che guidano macchine *spider* rosse e possiedono un cagnolino di piccola taglia sia solamente 10 sui 2 miliardi di individui registrati nel database in esame. Quindi la regola associativa estratta non è poi così significativa.

Nei fatti tali algoritmi di ricerca trovano con facilità moltissime associazioni, il problema è quello di identificare indici che misurino il peso delle associazioni trovate aiutandoci a individuare solo quelle significative.

Gli indici in questione sono in genere due:

- gli indici di supporto;
- gli indici di confidenza.

Facciamo un esempio. In una database abbiamo tre attributi binari per individuo registrato: lettore o non lettore di giornali sportivi (*sport, no sport*); lettore o non lettore di giornali generalisti (*gen, no gen*), lettore o non lettore di riviste di moto e motori (*moto, no moto*).

Abbiamo ricavato la seguente regola associativa: chi legge sia giornali sportivi sia giornali generalisti legge anche giornali di moto e motori. In termini formali *sport and gen => moto* ovvero *sport gen moto.*

Quanto è solida questa associazione? Quanto i dati contenuti nel database la supportano? L'indice di supporto ci dà informazioni in questo senso fornendo il rapporto tra il numero di lettori che contemporaneamente leggono *sport gen e moto* e l'intero numero di *record* contenuti nel database.

Purtroppo l'indice di supporto non basta di per sé a darci indicazioni sufficienti. Per esempio, ci possiamo accorgere che l'indice di supporto della nostra associazione a tre (*sport gen e*

moto) è significativo, ma che è anche di molto inferiore all'indice di supporto della sola associazione a due attributi *sport e gen*. In parole povere il fatto che il numero di lettori di tre riviste contemporaneamente sia molto piccolo rispetto a quello binario di partenza, ovvero i lettori delle sole due riviste sportiva e generalista, ci dice che la nostra associazione a tre benché rilevata, è alquanto debole. Entra così in gioco un ulteriore indice, quello di confidenza, che non è altro che il rapporto fra il numero di lettori di *sport gen e moto* e il numero di lettori di *sport e gen*.

Quindi, lavorando per ogni regola e ricavando al contempo per ogni associazione i relativi indici di supporto e di confidenza, si può arrivare a discriminare le associazioni importanti da quelle assolutamente inutili. Ancora una volta la ricerca di regole associative deve partire da un obiettivo prefissato, come ad esempio la conferma di una ipotesi di esistenza di una specifica associazione, altrimenti si è sommersi da una mole di informazioni che confonde più che chiarire. L'applicazione di un algoritmo che trova tante regole non garantisce che tutte quelle trovate siano significative. D'altra parte, se partiamo con un algoritmo più rigido troviamo sì meno associazioni, ma possiamo aver mancato quelle veramente importanti. Si devono quindi testare con criterio diversi algoritmi, valutare i risultati parziali con attenzione, per poi salire di una maglia sulla catena di scoperta della conoscenza.

Le reti neurali

Quando si arriva a voler creare un *tool* di analisi descrittivo, ovvero che implichi processi di tipo *machine learning,* si derivano modelli e strutture provenienti da aree della scienza completamente differenti dalla matematica e dall'informatica tradizionale. Tali sono gli algoritmi basati sulle reti neurali che trovano la loro origine nella ricerca e nella scienza del funzionamento del cervello in uomini ed animali.

Nei fatti, la struttura del cervello vede la presenza di un gran

numero di cellule dette neuroni, nell'uomo circa 10^11, ognuno dei quali è connesso a circa 2.000 altri neuroni tramite le cosiddette sinapsi. Basandosi su questa struttura base che è al contempo semplice, flessibile ed altamente ridondante, il cervello può gestire compiti estremamente complessi.

Tentando di ricreare nelle macchine gli elementi che funzionano come neuroni e sinapsi, dapprima con hardware specifici oggi tramite programmi software, si è aperto il campo di ricerca sulle reti neurali che sono impiegate per eseguire in modo automatico compiti complessi simili a quelli eseguiti dal cervello.

Tipicamente una rete neurale, come rappresentata in Figura 3, consiste in una serie di nodi interconnessi fra loro:

- i nodi di *input* ricevono segnali di ingresso;
- i nodi di *output* passano segnali in uscita;
- i nodi intermedi o nascosti, raggruppati in un numero variabile di strati intermedi, svolgono funzioni elaborative.

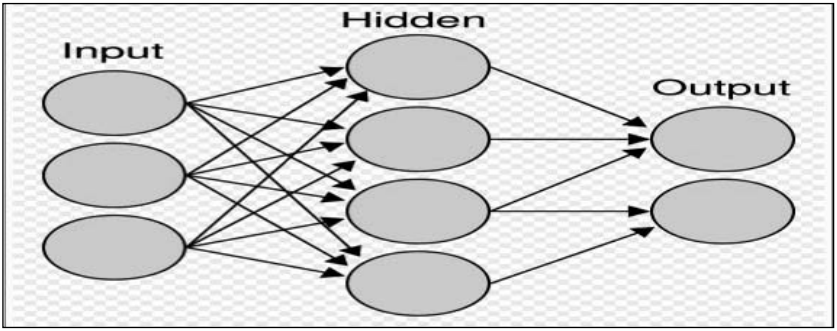


Figura 3. Rete neurale[9].

L'uso delle reti neurali prevede due fasi distinte:

- La fase di cosiddetto *encoding* nella quale si *insegna* alla rete ad eseguire un determinato compito;

9 <http://it.wikipedia.org/wiki/Rete_neurale>.

- La fase di cosiddetto *decoding* nella quale si utilizza la rete per eseguire in via ripetitiva il compito che le è stato insegnato (classificare esempi, fare predizioni, ecc).

Vi sono differenti architetture di reti neurali, ognuna delle quali è caratterizzata da specifiche strutture interconnettive e da differenti strategie di addestramento ed acquisizione dell'esperienza. In questo contesto parleremo di due strutture neurali che hanno un ruolo oggi nella tecnologia informatica:

- Le *Backpropagation Networks* (BPN, letteralmente le reti a retro propagazione), dette anche *Multilayer Perceptrons*;
- Le mappe auto-organizzanti di Kohonen.

Come mostrato nella Figura 4, una BPN consiste di una sem-

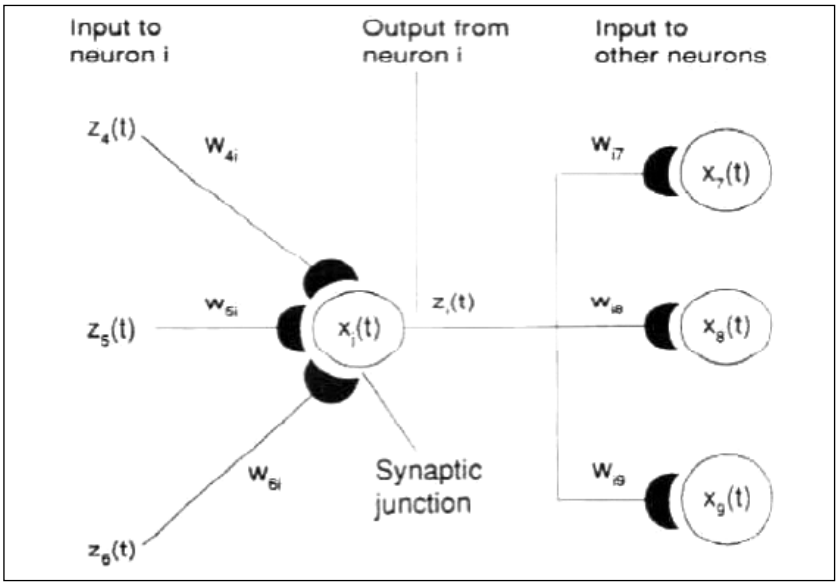


Figura 4. Nodo di un BPN[10].

[10] CHRISTOS STERGIOU, DIMITRIOS SIGANOS, *Neural Networks*. <http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html>.

plice struttura a strati (almeno tre) con i nodi di *input* sul primo strato e i nodi di *output* sull'ultimo. I nodi sugli strati interni sono detti nascosti (*hidden*). I nodi sono collegati da link (sinapsi) mono-direzionali utilizzati per la comunicazione tra di loro: ad ogni sinapsi è associato un peso. La rete opera in modalità *feedforward*, nel senso che l'input ad ogni nodo deriva soltanto da output di nodi posti negli strati precedenti (tipicamente nello strato immediatamente precedente): questo, evidentemente, rende la trasmissione all'interno della rete priva di cicli (feedback).

Allo stato iniziale le BPN hanno sinapsi con pesi w_{ij} scelti a caso. Durante la fase di *encoding* il network è esposto ad un insieme di dati di *training* in input (insieme di dati per cui si conosce la risposta, ovvero l'output) in istanze ripetute, dette, appunto, di *training*. Per ogni istanza di *training*, l'output della BPN è paragonato all'output corretto che ci si aspetta; se vi è una differenza fra *output effettivo* ed *output voluto* i pesi delle sinapsi di tutti i nodi intermedi coinvolti sono di volta in volta mutati sino a che la differenza non scende al disotto di un valore di soglia predeterminato. Il termine *Backpropagation* deriva dal fatto che tale aggiornamento dei pesi delle sinapsi è effettuato a partire dagli ultimi strati, procedendo verso i nodi di input. Una volta che la struttura della BPN si è stabilizzata (i.e. i pesi delle sinapsi non variano più e si hanno i dati di output voluti per i dati di training) le si possono far elaborare nuovi insiemi di dati per ottenere da essi un output che li categorizzi, operando su di essi in maniera analoga a quando si era operato sui dati di *training*.

L'architettura BPN rappresenta spesso un approccio importante ed efficace al problema della classificazione. Comunque, essa presenta alcuni inconvenienti: uno di essi è che la fase di *training* è lunga e necessita di insiemi molto popolati. Inoltre le BPN (così come molti altri classificatori di tipo statistico) forniscono i risultati senza dare alcuna motivazione per razionalizzare il risultato finale fornito (si ricordi ad esempio quanto accadeva con il metodo degli alberi di decisione).

Le Reti Auto Organizzanti di Kohonen (RAOK) possono essere pensate come la controparte artificiale della mappa di neuroni attivati che esistono in zone diverse del cervello. Una RAOK è un insieme di neuroni, le unità, ognuno dei quali è connesso ad un piccolo numero di altre unità chiamate vicini. Per la maggior parte del tempo la RAOK si mantiene in due dimensioni; in essa ogni nodo è caratterizzato da un fattore che è correlato allo spazio oggetto dell'investigazione. Nella fase iniziale la RAOK è caratterizzata da una scelta casuale di vettori, ovvero i fattori di cui sopra. Via via che il *training* avanza i vettori vengono sistemati al fine di garantire la migliore mappatura dello spazio da esplorare.

Le reti neurali funzionano bene in compiti di classificazione ed in questo ambito sono molto usate nel Data Mining, anche se sono delle vere scatole nere da cui scaturiscono i risultati senza che l'utente possa ricostruire in modo immediato la logica della classificazione ottenuta (anche se in realtà l'applicazione di analisi di tipo statistico, come ad esempio lo studio delle correlazioni tra attributi, permette di ottenere informazioni sull'effetto dei valori dei singoli attributi di input sull'output).

Gli Algoritmi Genetici

Gli algoritmi genetici sono una tipologia particolare della classe più generale degli algoritmi di ricerca locale, finalizzati, per l'appunto, alla ricerca dell'elemento (o della configurazione) ottimo all'interno di uno spazio di possibilità molto esteso. Tali algoritmi operano iterativamente a partire da un elemento, considerando un insieme di elementi ad esso simili (secondo un qualche criterio) e verificando se tra di essi esiste un elemento migliore di quello attuale. Gli algoritmi genetici seguono questo approccio, imitando il processo di selezione/evoluzione operante in natura. Essi hanno, per tale motivo, alcuni illustri padri:

- Darwin, che ha individuato nel processo di selezione/evoluzione un meccanismo di ottimizzazione, scelta ed adatta-

mento di una serie di attributi, da lui riferito agli esseri viventi, ma in via di principio applicabile a qualsiasi sistema di cui si voglia trovare un *optimum*;
- Francis Crick e James Watson, che hanno scoperto, attraverso il DNA, il meccanismo molecolare con il quale si attua negli esseri viventi il processo di codifica dell'informazione genetica, oltre che quello di selezione;
- Nils Aall Barricelli, virologo genetista italo norvegese, che ha introdotto e studiato algoritmi che tentano di riprodurre il processo di selezione naturale.

La formula per la costruzione di un algoritmo genetico per la soluzione di un qualsivoglia problema di ottimizzazione è la seguente:
- Strutturare un'adeguata codifica delle istanze del problema, scegliendo un alfabeto (corrispondente alle triplette di base del DNA) limitato ma efficace;
- Inventare un ambiente artificiale nel calcolatore dove le possibili diverse soluzioni al problema posto possano competere fra di loro. Per far ciò bisogna anche strutturare un buon algoritmo di *rating* che possa giudicare il successo o il fallimento di una possibile soluzione, e cioè quella che viene denominata funzione di *fitness* della possibile soluzione;
- Sviluppare modi in cui diverse soluzioni possano accoppiarsi e quindi generarne una nuova. A tal fine, la tecnica più usata è quella detta di *cross-over,* nella quale le stringhe *padre* e *madre* sono tagliate, cambiate e riattaccate l'una all'altra. Oltre al *cross-over*, in questo contesto è stata comunque applicata tutta una serie di operatori di mutazione;
- Avviare la simulazione evolutiva a partire da una popolazione di possibili soluzioni. È importante che questa sia ben bilanciata e varia Ad ogni passo evolutivo si devono scartare le soluzioni cattive eventualmente rimpiazzandole

con la loro *progenie*, ovvero con mutazioni che possiedono una buona *fitness*;
- Terminare il processo quando tutte le soluzioni rimaste in gioco sono buone soluzioni.

Per quanto, in via di principio, estranei al mondo del calcolo, i passi suddetti sono facilmente programmabili e ciò è dimostrato dal successo che questa algoritmica ha avuto a partire dagli anni '60-'70.

I meriti e i demeriti della famiglia degli algoritmi genetici riproducono quelli del processo di selezione che si opera in natura. I due principali svantaggi del metodo sono la larga produzione di soluzioni/individui e la casualità del processo di mutazione. In generale servono grandi potenze di calcolo, e di questo Barricelli si era ben accorto, al fine di raggiungere una soglia di accettabilità delle soluzioni generate. D'altra parte il processo è solido; se una soluzione esiste e si ha a disposizione un buon calcolatore, è ragionevole pensare che l'algoritmo genetico la scopra, anche se in tempi che potrebbero essere non accettabili. Di recente si è visto il nascere di una serie di metodi ibridi, nel cui ambito le reti neurali fornivano un risultato che serviva da input per gli algoritmi genetici. Ad oggi, le applicazioni più importanti degli algoritmi genetici si trovano nello studio degli andamenti dei mercati finanziari e in applicazioni specifiche per il mondo assicurativo.

Text Mining

Il Text Mining è una variazione del Data Mining e consiste nell'applicazione delle sue tecniche a testi non strutturati. La differenza tra il Data Mining ed il Text Mining è che nel Text Mining i dati (sequenze di parole) sono estratti da testi (collezioni di documenti) in linguaggio naturale piuttosto che da database strutturati. I testi esprimono, in realtà, una vasta e ricca gamma

di informazioni, codificate però in una forma che è difficile decifrare automaticamente.

Oggi, con la diffusione delle tecnologie proprie della società dell'informazione e con la crescita smisurata del numero di documenti che possono contenere informazioni interessanti e quindi da analizzare (pagine web, e-mail, libri e articoli in versione digitale, agenzie stampa, ecc.) risulta fondamentale disporre di strumenti automatici per la loro catalogazione ed analisi. Dal momento che i dati testuali non sono strutturati, l'analisi automatica appare comunque più difficile che nel caso del Data Mining operante su valori ben strutturati, eventualmente anche numerici. Operazione cruciale nel Text Mining è quella di estrarre, a partire dal contenuto testuale dei documenti della collezione, un insieme di indicatori numerici, che descrivano in modo il più possibile efficace il contenuto dei documenti, fornendo al tempo stesso un formato trattabile da algoritmi operanti su valori numerici, come quelli di mining.

L'approccio basilare, a tal fine, è basato sulla derivazione di un vettore di occorrenze per ogni documento (una matrice per la collezione), che associa ad ogni termine che occorre nella collezione (e quindi in un lessico da essa derivato) il numero di occorrenze nel documento. Si noti che tale soluzione comporta una pesante perdita di informazione, conseguente al fatto che viene ignorata la sequenza dei termini nel documento, così come, quindi, la vicinanza nelle occorrenze tra termini: per tale motivo, essa prende il nome di *bag of words* (sacca dei termini, ad indicare proprio il fatto che i termini sono considerati senza nessun ordine). Al tempo stesso, l'approccio *bag of words* porta ad associare ad ogni documento un vettore di dimensioni pari, almeno, al numero di termini diversi che occorrono nel documento stesso o, più spesso, al numero di termini diversi nella collezione. Questo fa sì che ogni elemento da considerare nel processo di mining venga modellato come un punto in uno spazio a dimensione molto elevata, dell'ordine di 10^5-10^6, con tutti i proble-

mi derivanti dall'operare in spazi di tale tipo. Tenuto conto del fatto che, da un lato molti termini del lessico non compaiono in uno specifico documento (e quindi forniscono valori pari a 0 nella componente associata del vettore) e che non tutti i termini sono ugualmente interessanti per caratterizzare un documento (termini molto frequenti, che appaiono in praticamente tutti i documenti, come ad esempio articoli e congiunzioni, risultano poco significativi per *comprendere* il contenuto di un documento), risulta importante nel Text Mining utilizzare metodi efficienti di riduzione di dimensionalità, che consentano di individuare insiemi sufficientemente limitati di termini che caratterizzino in modo sufficientemente preciso il contenuto di un documento.

Un problema comune che può essere risolto con il Text Mining consiste nella determinazione dell'appartenenza o meno di un dato documento ad uno specifico argomento (*classificazione*, *raggruppamento tematico*). L'individuazione di gruppi tematici consente di dare un'organizzazione all'informazione disponibile e di individuare argomenti minori, che, anche ad una lettura attenta, potrebbero sfuggire. Le relazioni, inoltre, possono mettere in evidenza *associazioni* nascoste considerando legami tra argomenti apparentemente separati ma che hanno una terminologia comune (legami tra argomenti, o tra autori, *trend* temporali, ecc.).

Altri problemi che possono essere affrontati con il Text Mining riguardano la possibilità di *identificare entità* (es. nomi di geni, nomi di aziende, ecc.) contenute nei documenti (*information extraction*) ed *estrarre concetti* per la creazione di ontologie, ovvero di strutture formali che modellano *in astratto* un insieme di proprietà dei dati e di relazioni tra gli stessi: l'associazione di una collezione di dati ad una ontologia può fornire un importante ausilio al loro trattamento, associando ad ognuno di essi una *semantica* (in modo similare all'associazione di un schema concettuale ai dati presenti in un sistema informativo). Tale operazione, di tipo top down, può risultare però molto difficoltosa, in quanto prevede l'associazione ad ogni elemento della collezione

di un insieme di informazioni (metadati): un approccio *manuale* può risultare impraticabile in presenza (come spesso ormai accade) di collezioni molto estese, mentre l'utilizzo di tecniche automatiche appare ancora problematico, soprattutto in presenza di ontologie abbastanza articolate.

In alternativa, è possibile utilizzare un approccio bottom up di estrazione dell'ontologia a partire dalla collezione dei dati (*ontology learning*).

Il Text Mining può essere usato per analizzare documenti testuali, in particolare scritti in linguaggio naturale riguardanti qualsiasi soggetto: l'esplosione del numero di documenti accessibili via Internet (pagine web, messaggi di email, blog post, news, libri e articoli, ecc.) rende la definizione di tecniche efficaci per l'estrazione di conoscenza da tali enormi collezioni di documenti (così come per la ricerca all'interno di esse) un fattore essenziale per il loro utilizzo. Applicazioni immediate, in tale contesto, sono la ricerca delle informazioni sul Web utili ai fini di una specifica esigenza informativa (*web retrieval*), l'analisi di trend, anche temporali, su larga scala (*issue* particolarmente discusse, *sentiment analysis*, estrazione di relazioni di similitudine tra documenti e tra utenti, sistemi di suggerimento e personalizzazione, monitoraggio politico, ecc.).

Ciò è vero anche in contesti più ristretti, come ad esempio in campo biomedico e specificatamente nella ricerca di interazioni (attivazione, inibizione, ecc.) tra le proteine attraverso l'analisi della letteratura di dominio. Il Text Mining, applicato in questo contesto, non solo può estrarre informazioni dai documenti sulle interazioni fra le proteine, ma può anche andare un passo oltre per scoprire *pattern* nelle interazioni estratte. Queste informazioni potrebbero aiutare a rispondere a importanti domande o suggerire nuovi percorsi da esplorare.

Il Text Mining coinvolge l'applicazione di tecniche proprie ad altre aree, come ad esempio l'information retrieval, l'elaborazione del linguaggio naturale, l'analisi statistica dei dati. Tecniche

che possono costituire fasi di un unico processo di Text Mining e possono essere combinate insieme in un singolo *workflow*. Si può affermare che, analogamente a quanto visto per il KDD, il Text Mining percorre le fasi classiche di un processo di estrazione di conoscenza.

Vediamo ora come si struttura tale processo descrivendo con maggior dettaglio ognuna delle fasi in cui si suddivide:

- *Information retrieval (IR)*: identifica all'interno di una collezione un insieme di documenti che soddisfano una esigenza informativa dell'utente, espressa attraverso una *query*, tipicamente fornendo per ciascuno di essi una valutazione della relativa rilevanza rispetto a tale esigenza. I sistemi IR più conosciuti sono i *search engine* come Google™, che identifica i documenti accessibili su World Wide Web più rilevanti rispetto alle esigenze dell'utente, espresse mediante la specifica di parole chiave. I sistemi IR sono anche utilizzati frequentemente a livello personale, per accedere in modo efficiente ai documenti di interesse presenti su un personal computer (*desktop search*) o per organizzare e ricercare l'insieme, tipicamente vasto, di documenti digitali prodotti all'interno di una organizzazione o di una azienda (*enterprise search*): in questo senso, lo sviluppo e l'utilizzo di tali sistemi risulta elemento fondamentale di tutti i processi di digitalizzazione dell'informazione e di applicazione di procedure informatiche per la relativa gestione, ormai ampiamente praticati in organizzazioni ed aziende. Inoltre, i sistemi IR sono spesso usati anche in ambito bibliotecario, dove i documenti non sono i libri stessi ma *record* digitali contenenti informazioni relative ai libri. Ciò tuttavia sta cambiando con la diffusione delle biblioteche digitali (*digital libraries*) dove i documenti da recuperare sono la versione digitale dei libri e delle riviste (*journal*). I sistemi IR consentono di accedere più velocemente all'argomento di interesse e individuare i lega-

mi con altri argomenti riducendo la quantità di documenti da analizzare. Per esempio, se siamo interessati solo a informazioni riguardanti le interazioni tra proteine, possiamo restringere l'analisi a quei documenti che contengono il nome di una proteina, o alcune forme verbali del verbo *interagire* o uno dei suoi sinonimi;

- *Elaborazione del linguaggio naturale (Natural Language Processing - NLP)*: è uno dei problemi più indagati e difficili nel campo dell'intelligenza artificiale. Si occupa di sviluppare tecniche di analisi del linguaggio *umano*, (quindi non formalizzato come i linguaggi *artificiali* sviluppati in ambito informatico), al fine di consentire la comprensione automatica del testo, sia ricostruendone la struttura sintattica, che estraendo attraverso di essa un significato dal testo stesso. Benché l'obiettivo finale sia ancora lontano da raggiungere, esistono sistemi che possono effettuare alcuni tipi di analisi con un certo grado di successo. Il ruolo di questi sistemi nel Text Mining è tipicamente quello di intervenire nelle attività di estrazione di significato, e quindi di informazione, dal testo;
- *Estrazione delle informazioni*: è il processo che consente di ottenere dati strutturati da un documento non strutturato in linguaggio naturale. L'obiettivo è l'estrazione di termini specifici dal testo aventi una particolare semantica predefinita. Esempi in questo senso possono essere offerti dall'estrazione di un insieme di keyword tematiche da un articolo di ricerca, dall'individuazione di *esperti* su un argomento (*expert finding*), dalla costruzione di una rete di citazioni tra articoli e/o libri, ecc.

Al termine di questa fase, per ogni documento si ottengono informazioni (meta-informazioni aggiuntive), che possono ad esempio essere memorizzate in un dizionario o in una ontologia.

- *Data Mining*: la fase di Data Mining vera e propria si con-

cretizza nell'applicazione di algoritmi ai documenti, rappresentati dal loro contenuto e/o da informazioni estratte dal contenuto stesso mediante applicazione di tecniche di NLP e di *information extraction*. Gli algoritmi utilizzati dipendono evidentemente dagli obiettivi dell'attività di mining, obiettivi che possono essere di classificazione di documenti, così come di estrazione di topic (argomenti trattati), di misura del livello di similitudine tra documenti e del conseguente clustering, di stima della rilevanza di un documento, di analisi dell'andamento temporale dei contenuti dei documenti.

Un caso esemplare, molto diffuso, di classificazione è rappresentato dall'implementazione di filtri anti-spam nei sistemi di e-mail: tali filtri permettono di effettuare una prima discriminazione automatica tra messaggi (presumibilmente) significativi e messaggi (presumibilmente) non significativi, o spam. La classificazione viene effettuata sulla base del contenuto dei messaggi e di una rappresentazione delle caratteristiche dei messaggi spam ottenuta sulla base di esempi: tale caratterizzazione è dinamica, e può essere modificata dall'utente indicando esempi di messaggi spam e non spam. Un altro esempio, più complesso, di classificazione di documenti è fornito dai sistemi di *sentiment analysis*: questi, dato un tema (rappresentato ad esempio da una query), classificano i documenti che trattano di tale argomento sulla base del tipo di giudizio che ne viene fornito, secondo uno schema di classificazione predefinito, che potrà essere ad esempio *giudizio favorevole*, *giudizio sfavorevole*, *giudizio neutro*, *giudizio non espresso*. È facile rendersi conto della appetibilità di tale tipo di sistemi, applicati ad esempio sul contenuto di collezioni di blog, ma anche di reti sociali (come Facebook© o Twitter©) per tutte le attività di monitoraggio di un marchio o di un prodotto in ambito di marketing, industriale o politico, così come di caratterizzazione delle preferenze dei singoli utenti, sulla base dei giudizi espressi.

L'estrazione di topic permette di caratterizzare un documento in modo automatico individuando l'argomento o gli argomenti sviluppati al suo interno ed eventualmente fornendo una misura di quanto ciascuno di essi venga trattato. Un argomento è rappresentato da insiemi di termini (tipicamente pesati per rilevanza e quindi visti come distribuzioni di probabilità dei termini nei documenti trattanti l'argomento in questione) e può essere, più raramente, predefinito o, più spesso, non predefinito: nel primo caso, il task può essere considerato come equivalente ad una classificazione sulla base di una misura di similitudine tra il documento in esame e i documenti *ideali* dati dalla definizione degli argomenti. Nel secondo caso siamo in un ambito non supervisionato di estrazione di un insieme di argomenti trattati, rappresentati ancora da insiemi di termini *pesati* per rilevanza. Ciò fornisce uno strumento di sintesi utile nella ricerca di documenti di interesse in una collezione.

L'estrazione di topic fornisce inoltre un utile supporto alla definizione di misure di similitudine tra documenti. I metodi più immediati si basano sul contenuto dei documenti stessi (occorrenze dei termini), utilizzando misure di tipo geometrico, nel qual caso i documenti sono modellati come punti in uno spazio di dimensione opportuna, o misure derivate dalla statistica e dalla teoria dell'informazione, in base alle quali i documenti sono visti come distribuzioni di probabilità dei termini: in entrambi i casi, è necessario fare riferimento a modelli di documenti definiti su spazi di grande dimensione. L'estrazione di topic consente di rappresentare un documento all'interno di uno spazio (quello delle topic, definite o estratte che siano) molto più piccolo rispetto a quello dei termini, con tutti i vantaggi derivanti dalla conseguente riduzione di dimensionalità.

La stima della rilevanza di un documento rappresenta un aspetto fondamentale per tutte le attività di ricerca di documenti di interesse all'interno di una collezione. Data la grande disponibilità di informazione accessibile, la collezione di riferimento

avrà tipicamente dimensione molto estesa (il Web è un caso tipico), il che comporterà che l'insieme dei documenti plausibilmente di interesse rispetto ad una specifica esigenza informativa possa essere talmente esteso da rendere impossibile il suo esame completo, documento per documento. È quindi necessario avere a disposizione tecniche che consentano di effettuare una stima della rilevanza di ogni documento, sia in riferimento all'argomento di interesse che in assoluto. Mentre la rilevanza rispetto ad una query viene tipicamente stimata sulla base del contenuto del documento stesso (essenzialmente misurando la similitudine tra tale contenuto e il contenuto di un documento fittizio corrispondente all'argomento di interesse), la sua rilevanza in assoluto richiede, per essere valutata, l'esame di informazioni di tipo diverso, come ad esempio la quantità di rimandi a tale documento da parte di altri documenti nella collezione: si valuta infatti che, approssimativamente, quanto più un documento è citato all'interno della collezione, tanto più è importante, indipendentemente dagli argomenti trattati al suo interno. Gli attuali motori di ricerca operanti sul Web, come Google©, interpretano, come è noto, i link tra pagine web come citazioni e utilizzano tale informazione per estrarre una stima della rilevanza di un documento indipendentemente dal suo contenuto.

L'evoluzione di tutti i parametri derivati dall'applicazione di tecniche di mining e riferiti a singoli documenti o a un'intera collezione, può inoltre, nel caso in cui sia disponibile informazione relativa al tempo, essere rappresentata su scala temporale, consentendo di valutare in modo dinamico l'andamento dell'informazione di interesse, sia essa la rilevanza di uno o più documenti, l'interesse verso una particolare tematica, i giudizi espressi nei confronti di un argomento. In questo caso, l'analisi di serie storiche permette di osservare e modellare l'andamento nel tempo delle informazioni studiate ed eventualmente di costruire modelli predittivi dell'andamento futuro di tali informazioni.

Per meglio descrivere il processo e le tecniche che si possono

applicare si ritiene opportuno proporre l'esempio di costruzione di un prototipo[11] per un sistema di Text Mining che consente di analizzare la letteratura biomedica, nel campo della genetica, allo scopo di individuare le eventuali interazioni tra geni.

L'uso del Text Mining in questa specifica area applicativa risulta particolarmente importante poiché la conoscenza che si estrae analizzando le pubblicazioni specialistiche può essere considerata una fondamentale sorgente di informazioni che il ricercatore utilizza per interpretare e comprendere meglio i risultati sperimentali ottenuti usando la tecnologia DNA *microarray*[12].

L'approccio che viene utilizzato si pone l'obiettivo di produrre una rappresentazione strutturata dell'informazione testuale predefinendo entità e relazioni da ricercare nei testi (ad esempio le entità possono essere proteine e farmaci e le relazioni attivazione, inibizione, ecc.). Si adottano tecniche che utilizzano sia l'analisi sintattica che l'analisi semantica. Alla fine del processo viene generato un database che può essere analizzato utilizzando le tecniche di data mining.

Come primo passo, attraverso un sistema IR, può essere effettuata una selezione tra i documenti disponibili (Pubmed[13]) per ottenere l'insieme completo delle pubblicazioni che trattano di ciclo cellulare utilizzando uno specifico insieme di parole chiave (*cell cycle* OR *cell proliferation* OR *cell death* OR *oncogenes* OR *tumor suppressor*, ecc.).

Una volta delimitato l'insieme di documenti voluto si passa alla loro cosiddetta preparazione, che può essere realizzata attraverso tre fasi:

- Individuazione delle diverse parti del documento, che consente di separare l'informazione testuale dalla metainfor-

11 Progetto MedMole. <http://medmole.cineca.it/>.
12 <http://it.wikipedia.org/wiki/Microarray> .
13 <http://www.ncbi.nlm.nih.gov/pubmed>.

mazione (organismo di appartenenza, data e rivista di pubblicazione, tipo di pubblicazione, paese, ecc.);
- *Analisi grammaticale* sulla parte testuale dei documenti così individuata;
- Applicazione di tecniche di *estrazione dell'informazione.*

In particolare, attraverso l'*analisi grammaticale* di un testo, si selezionano tutti i sostantivi, che rappresentano generalmente i termini con maggiore contenuto semantico. Questi costituiscono l'insieme di parole chiave che caratterizza e descrive ciascun documento e sulla base del quale verrà giudicato il grado di somiglianza tra i documenti. Tutte le altre parti del discorso (aggettivi, verbi e nomi propri) vengono mantenute come informazioni aggiuntive, così come la metainformazione che era stata estratta durante la prima fase. Poiché nell'ambito della ricerca biologica è molto importante poter identificare i nomi dei geni e poiché l'analisi grammaticale li classifica, genericamente, come nomi propri, è necessario analizzare ulteriormente il testo per estrarre questo tipo di informazione. Con la fase di *estrazione dell'informazione* ci si pone l'obiettivo di estrarre termini specifici (e predefiniti) dal testo, utilizzando un dizionario che, nel caso specifico, contiene i nomi ufficiali dei geni ed i relativi *alias*.

Il dizionario non è altro che una tabella (*file*) di due colonne: la prima, *GENE*, contiene il nome ufficiale; la seconda, *ALIAS*, un *alias* dello stesso gene; il nome del gene è ripetuto tante volte per quanti sono gli *alias* esistenti. Lo strumento che si utilizza cerca nel documento la presenza di qualsiasi occorrenza inserita nella colonna *ALIAS* e memorizza il corrispondente nome presente nella colonna *GENE* associandogli l'identificatore del documento analizzato. Si ottiene così una lista di nomi di geni, referenziati in ogni documento, che viene integrata con la lista di parole chiave precedentemente individuate.

Alla fine di questa fase ogni documento è descritto in formato sintetico da una lista di parole chiave e da una lista di nomi di

geni. Questo formato consente di rappresentare l'informazione in una matrice binaria che contiene, sulle righe, ciascun documento, sulle colonne, ciascun sostantivo, e all'interno i valori 1 o 0 ad indicare rispettivamente la presenza o l'assenza di una determinata parola chiave in un determinato documento (la metainformazione è rappresentata in maniera analoga).

Siamo così giunti alla fase di *data mining* vera e propria che si concretizza nell'applicazione di un algoritmo di *clustering*. Poiché l'informazione disponibile è di tipo qualitativo, si può scegliere un algoritmo partitivo basato sull'analisi relazionale.

Questo metodo consente di confrontare tutte le coppie di documenti e di calcolare, per ciascuna di esse, un indice di somiglianza basato sul numero di co-occorrenze delle parole chiave. I *cluster* vengono formati in modo da massimizzare la somiglianza complessiva dei documenti raggruppati e minimizzare la somiglianza dei documenti che vengono separati. Su questo processo si può intervenire agendo attraverso alcuni parametri (la soglia di somiglianza ed il sistema di ponderazione per l'assegnazione di pesi diversi agli attributi scelti per la formazione dei cluster).

Supponiamo di essere interessati alle eventuali interazioni tra due geni, BRCA1 e BRCA2 e selezioniamo tutti i documenti che contengono almeno uno di questi geni. Può essere lanciata un'analisi di *data mining* che, in tempo reale, confronta le parole chiave dei documenti selezionati (in quanto questi sono gli attributi descrittivi che abbiamo deciso di usare per il *clustering*) e li raggruppa secondo i *pattern* individuati. La visualizzazione dei risultati rende evidente che i documenti selezionati tendono a trattare i due geni separatamente, come era prevedibile, dato che un gene è implicato nel tumore al polmone e l'altro nel tumore al seno e che le interrelazioni dovrebbero essere minime. Viceversa in alcuni casi viene evidenziato che, quando compare BRCA1, spesso compare anche BRCA2, quindi tra questi due geni può esserci una interrelazione.

Gli stessi passi possono essere generalizzati e, quindi, essere fatti su un qualsivoglia altro tipo di documenti posto che si sia in possesso di un dizionario significativo per il settore nel cui ambito si svolge la ricerca.

Bibliografia

GIUDICI, P., *Data mining e statistica*, in «Statistica & Società», vol. 3, n. 1, 2004, pp. 3-7

STERGIOU, C., SIGANOS, D., *Neural Networks* <http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html>

Sitografia

<http://medmole.cineca.it/>
<http://dms.irb.hr/tutorial/tut_dtrees.php>
<http://it.wikipedia.org/wiki/Microarray>
<http://it.wikipedia.org/wiki/Rete_neurale>
<http://www.cineca.it>
<http://www.ncbi.nlm.nih.gov/pubmed>

Sui motori di ricerca

PAOLO FERRAGINA*

Esistono due paradigmi principali per cercare le informazioni all'interno di un'ampia collezione di documenti, quale il Web: la *navigazione* della collezione e la ricerca per parole chiave. I browser (quali Explorer, Chrome, Firefox, solo per citare i più diffusi) sono strumenti che realizzano il primo paradigma, e sono utilissimi se l'utente conosce l'indirizzo URL della pagina desiderata o la collocazione del documento all'interno di un'ontologia[1], ma inutili se si desidera reperire i documenti attraverso il loro contenuto senza conoscerne l'indirizzo. I motori di ricerca sono software sofisticati che permettono di realizzare il secondo paradigma recuperando efficientemente dal Web i documenti che sono *pertinenti* a una interrogazione formulata da un utente.

Questo compito è però assai complesso, per una serie di difficoltà legate alla dimensione del Web, alla natura variegata dei suoi contenuti, alla sua struttura interconnessa e fortemente dinamica, alla tipologia delle interrogazioni formulate dagli utenti e al modo con cui questi interagiscono con i risultati della ricerca. Non si può dunque prescindere da queste difficoltà quando si

* Università di Pisa, Dipartimento di Informatica.

1 Per ontologia si intende una rappresentazione formale di un insieme di concetti e delle relazioni che tra questi intercorrono. Nel caso specifico si fa riferimento all'uso sempre più diffuso delle ontologie nel Web, in quanto strumenti capaci di aggiungere una dimensione semantica all'organizzazione dei documenti e quindi al loro recupero.

vogliano analizzare i moderni motori di ricerca e valutare i risultati che sono stati ottenuti sia in ambito accademico sia in ambito industriale in questi vent'anni della loro storia.

Il Web è una rete sconfinata di *documenti*, oggi non solo pagine ma anche file disponibili in vari formati (dal foglio excel, al pdf o ePub, al file video o audio, ecc.) interconnessi tra loro, il cui numero è cresciuto in maniera vertiginosa nell'ultimo ventennio: si è partiti dalla singola pagina creata da Tim Berners-Lee al CERN nel 1991, per arrivare alle migliaia di miliardi di pagine indicizzate da Google oggi (Alpert, Hajaj, 2008)[2]. Questi documenti sono fortemente eterogenei, per lo stile variegato e a volte malizioso con il quale gli utenti li compongono (così da risultare più rilevanti nei motori di ricerca), e per le lingue utilizzate, che sono alcune centinaia e che vedono oggi quelle asiatiche sempre più preponderanti (rendendo ancora più sofisticata l'analisi dei documenti). A tutto ciò si aggiunge il fatto che il Web è fortemente dinamico sia nei contenuti delle pagine che nella sua struttura interconnessa, rendendo dunque ogni sua rappresentazione effimera.

Accanto a questi elementi di criticità connessi con la natura del Web, bisogna considerare le *necessità informative* degli utenti codificate nelle poche parole chiave che compongono le loro interrogazioni. Queste necessità sono variabili nel tempo e soggettive e possono avere essenzialmente tre finalità non mutuamente esclusive:

- *informazionale*, se l'interrogazione è stata effettuata per ricevere informazioni su qualcosa (ad esempio: Carlo Magno, la storia di Roma, la finale mondiale del 1982, ecc.);
- *navigazionale*, se l'interrogazione ha come obiettivo quello di trovare un sito dal quale far partire un'esplorazione

[2] Cfr. JESSE ALPERT, NISSAN HAJAJ, *We knew the Web was big*, Official Google Blog, luglio 2008. <http://googleblog.blogspot.it/2008/07/we-knew-Web-was-big.html>.

del Web (ad esempio Alitalia, per trovare alcuni voli);
– *transazionale*, se l'obiettivo ultimo dell'interrogazione è quello di trovare un sito per comprare qualcosa (ad esempio Sistina, perché voglio comprare un biglietto per uno spettacolo teatrale).

L'altro aspetto che rende ancora più arduo il compito dei motori di ricerca è legato alla tipologia delle interrogazioni poste dagli utenti e alla loro interazione con i risultati prodotti dalla ricerca. Infatti il modello *bag of words* (ossia, insieme di termini, o keyword) che si è affermato nel corso di questi anni è quanto mai povero dal punto di vista espressivo e, quindi, rende al quanto difficile per un utente non esperto il compito di tradurre le proprie necessità informative in una sequenza di parole chiave. Così più dell'80% delle interrogazioni risulta mal specificato o polisemico, essendo composto da meno di tre parole. Inoltre più dell'85% degli utenti guarda soltanto la prima schermata di risultati (tipicamente composta da 10 pagine Web) e quindi impone ai motori di ricerca di riuscire a selezionare tra le milioni di pagine che potenzialmente potrebbero fornire una risposta pertinente all'interrogazione dell'utente quelle 10 che probabilmente sono le più adeguate partendo soltanto dalle poche keyword specificate nella interrogazione.

Per capire come i progettisti dei motori di ricerca siano riusciti a migliorare significativamente le prestazioni di questi strumenti, in efficienza e in efficacia, nel seguito del capitolo si traccerà una breve cronistoria dell'evoluzione di questi sofisticati software, poi si dettaglieranno i principali moduli che li costituiscono, indicando alcune delle loro più interessanti soluzioni algoritmiche. In ultimo si accennerà ad alcune recenti evoluzioni nell'ambito dell'Information Retrieval (IR) e di come queste potrebbero avere un impatto positivo sui motori di ricerca di futura generazione.

Breve storia dei motori di ricerca

Nella storia dei motori di ricerca si possono individuare essenzialmente quattro generazioni.

La prima nasce intorno al 1993 e include alcuni pionieristici motori di ricerca per il nascente Web quali Wanderer e ALIWEB. Questi sfruttavano soltanto le meta-informazioni associate alle pagine Web dai loro autori. La struttura algoritmica di questi software prevedeva la scansione dell'intero archivio di pagine e l'assenza di qualunque ordinamento dei risultati, visto che il numero delle pagine esistenti e, quindi, pertinenti per una interrogazione era relativamente piccolo sia per le potenze di calcolo dei sistemi dell'epoca sia per l'analisi *a occhio* degli utenti. La repentina crescita della dimensione del Web rese presto inefficace questo approccio alla ricerca basato su scansione.

La seconda generazione coincide con la nascita di Altavista (ma anche Lycos, Excite, e altri) e si colloca intorno al periodo 1993-97. Questi motori utilizzavano il contenuto testuale della pagina e basavano l'ordinamento dei risultati di una ricerca sulla frequenza delle parole. I risultati furono eccellenti finché i documenti disponibili sul Web erano pochi e di elevata qualità. Con l'uso sempre più diffuso del Web, specialmente in ambito commerciale, e la conoscenza del funzionamento dei motori di ricerca, molti utenti cominciarono a costruire pagine per influenzare i risultati delle ricerche (spamming). La tecnica utilizzata era quella di riempire la pagina di numerose parole che non c'entravano nulla con il suo contenuto e che erano relative alle interrogazioni più frequenti degli utenti, scrivendole nello stesso colore dello sfondo della pagina, in modo che il lettore non sene accorgesse (ma il motore sì!). Questo inficiava il meccanismo di calcolo della rilevanza basato sul contenuto esclusivamente testuale delle pagine, rendendo spesso inutilizzabili questi motori sulle interrogazioni frequenti.

La terza generazione coincide con la nascita di Google e si

colloca quindi intorno al 1998. A questa generazione si ascrivono anche Ask Jeeves, Bing e Yahoo! che costituiscono i motori di ricerca che hanno dominato, pressoché incontrastati, lo scenario delle ricerche sul Web per tutto il decennio successivo. Nella prima versione di Google, la rilevanza di una pagina dipendeva dal suo contenuto, come per Altavista, ma anche da ciò che altre pagine *scrivevano* di lei (i cosiddetti testi *àncora*[3]) e, soprattutto, da quanto queste pagine erano a loro volta rilevanti. Si trattava quindi di una definizione *ricorsiva* ma sorprendentemente ben fondata dal punto di vista matematico (denominata *PageRank*).

In seguito anche questo meccanismo di rilevanza è stato messo in difficoltà da alcune tecniche di spamming[4], la più famosa delle quali prende il nome di *Google bombing*. Un esempio famoso è quello della interrogazione *miserable failure*, che portava Google a restituire in prima posizione la pagina dell'ex Presidente degli Stati Uniti George W. Bush.

La quarta e ultima generazione vede oggi impegnati i due principali protagonisti del settore – Bing e Google[5] – più una moltitudine di altri motori di ricerca cosiddetti *semantici* utilizzati da piccole comunità di utenti (si vedano p.e. Blekko, DuckDuckGo, Hakia): essi sono particolarmente interessanti dal punto di vista tecnologico in quanto offrono alcune caratteristiche innovative sia per quanto riguarda la composizione e l'interpretazione delle interrogazioni, sia per l'analisi del contenuto dei documenti presenti sul Web. Forse sorprende osservare che i do-

[3] Si tratta di quelle porzioni di testo sottolineate e, solitamente, di colore azzurro che si trovano nelle pagine Web e che descrivono gli hyper-link. Spesso i testi àncora sono estesi con un certo numero di parole che precedono e seguono le suddette.

[4] È del 25 Giugno 2012 la notizia che Google dichiara di scoprire circa diecimila siti dannosi al giorno e questi includono sia siti costruiti maliziosamente per ingannare i motori di ricerca sia siti oggetto di attacco.

[5] Yahoo! è uscito dalla scena della *Web Search* nel 2009 con l'accordo firmato con Microsoft per la fornitura del servizio di *search* ora basato su Bing.

cumenti indicizzati da tutti questi motori di ricerca costituiscono solo una piccola parte dell'intero Web e sono inoltre abbastanza diversi. Ciò è importante dal punto di vista degli utenti perché vuol dire che cercare su motori diversi equivale a trovare informazioni significativamente diverse e, quindi, avere un quadro del Web più ampio e variegato. Questo fa anche sì che oggi abbia senso interrogare più motori di ricerca (o servirsi di *metamotori*) per trovare le informazioni di cui si ha bisogno.

È importante osservare che i motori di ricerca moderni sfruttano non soltanto le informazioni contenute nei documenti (contenuto ed eventuali hyper-link), ma anche varie sorgenti informative (quali news, blog, previsioni meteo, mappe, ecc.) e tutte quelle tracce sociali e di comportamento che gli utenti disseminano sul Web quando lo navigano, scrivono le loro email, chattano con i loro amici in qualche social network, acquistano dei prodotti, commentano dei post/tweet, fanno delle foto o video geo-referenziati. Queste tracce sono importantissime sia individualmente, in quanto consentono ai motori di ricerca di individuare i bisogni degli utenti e quindi personalizzare i risultati delle loro ricerche, sia in forma aggregata (e quindi anonima) perché permettono di derivare automaticamente relazioni tra milioni di termini e tag (le cosiddette *folksonomies*) e di riconoscere neologismi che possono così essere utilizzati per disambiguare interrogazioni o perfezionare le stesse. Questo tema sarà trattato più diffusamente nell'ultima sezione del capitolo.

La struttura di un motore di ricerca

In letteratura esistono diversi studi sui motori di ricerca, tra i quali è possibile citare il testo di (Witten et alii, 2007)[6], che con-

[6] Cfr. IAN H. WITTEN, MARCO GORI, TERESA NUMERICO, *Web Dragons*, Morgan Kaufmann Publishers, 2007.

tiene un'analisi generale delle loro caratteristiche e delle implicazioni derivanti dal loro impiego, pur senza entrare in dettagli algoritmici, e (Witten et alii, 1999)[7] e (Manning et alii, 2008)[8] che possono essere considerati dei riferimenti algoritmici fondamentali sull'organizzazione di grandissimi insiemi di dati.

Un motore di ricerca consiste di cinque moduli principali che interagiscono, come mostrato nella Figura 1.

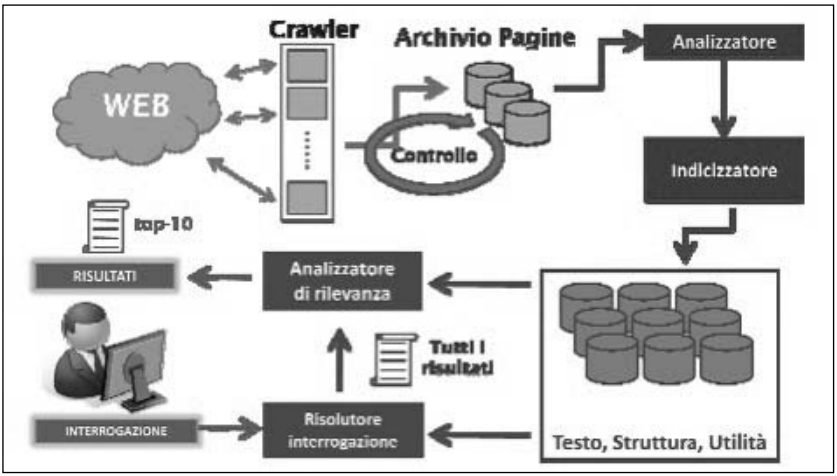


Figura 1. La struttura e il funzionamento di un motore di ricerca.

Si noti che questi moduli possono essere raggruppati in due classi: quelli che operano sui documenti estratti dal Web (quali *Crawler*, Analizzatore e Indicizzatore) e quelli che operano sulle interrogazioni poste dagli utenti (quali Risolutore e Analizzatore di rilevanza). Questi due gruppi di software interagiscono at-

7 Cfr. IAN H. WITTEN, ALISTAIR MOFFAT, TIMOTHY C. BELL, *Managing Gigabytes*, Morgan Kaufmann Publishers, 1999.

8 Cfr. CHRISTOPHER D. MANNING, PRABHAKAR RAGHAVAN, HINRICH SCHÜTZE, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

traverso la struttura dati prodotta dai primi e utilizzata dai secondi. Essa contiene vari tipi di informazioni relative al testo e alla struttura dei documenti indicizzati, così come informazioni di varia utilità quali: le parole del titolo, i font, la lingua, le meta-informazioni associate al documento/pagina dal suo autore, il numero di termini di una pagina, il numero di click eseguiti sul documento/pagina per navigazione o ricerca, ecc. Google dichiara di utilizzare oltre 200 informazioni di utilità (dette in gergo *features*) per caratterizzare una pagina Web, e di occupare ben oltre 100 milioni di Gigabyte per memorizzare il suo indice[9].

Il *crawler* (o *spider*) è quel modulo che raccoglie i documenti dal Web secondo una certa politica di visita del grafo sottostante, che è diversa da motore a motore, e necessaria, in quanto il Web, come detto, è sconfinato e dinamico. Il modulo di *analisi di documenti* esamina quelli catturati dal *crawler* stabilendo il linguaggio utilizzato (italiano, inglese, ecc.), eliminando segni di interpunzione o altri caratteri superflui, ed estraendo diverse altre informazioni testuali, strutturali (ossia gli hyper-link) e linguistiche in essi contenute. Queste informazioni vengono poi utilizzate dal modulo *indicizzatore* per costruire la struttura dati (o *indice*) che permetterà poi al motore di ricerca di trovare velocemente i documenti contenenti le parole chiave specificate dagli utenti nelle loro interrogazioni. L'esecuzione di questi tre moduli sarà continua e ripetuta con una frequenza che dipende dalla potenza di calcolo messa in gioco[10] e dalle caratteristiche dei documenti da indicizzare[11]. L'obiettivo sarà sempre e comunque

9 <http://www.google.com/insidesearch/playground/underthehood.html>.

10 Anche se le diverse società sono riluttanti a fornire dati precisi, si stima che per ciascun motore operino centinaia di migliaia di computer raccolti in sottoreti dedicate a funzioni diverse e non sempre completamente distinte tra loro.

11 I siti dinamici, quali quelli che espongono *news* e social network, dovranno essere visitati più frequentemente rispetto agli altri.

quello di offrire all'utente un *indice del Web* il più possibilmente completo e aggiornato.

Quando un utente compone un'interrogazione sotto forma di una sequenza di *parole chiave*[12], il motore di ricerca esegue il modulo *risolutore dell'interrogazione* per recuperare dall'indice tutti i documenti Web che contengono queste parole e quindi sono potenzialmente interessanti per quella interrogazione. D'altra parte questi documenti potrebbero essere milioni e l'interrogazione potrebbe celare, come già detto, diversi *bisogni informativi* che dipendono dall'utente che l'ha formulata e dalle sue intenzioni del momento. Per cui non è sufficiente recuperare tutti i documenti *interessanti*, occorre selezionare tra questi quelli più *rilevanti* per l'interrogazione stessa. Questo compito viene assolto dal modulo *analizzatore di rilevanza* che ordina i risultati in accordo a una serie di criteri sofisticati che variano da motore a motore e che sono a tutt'oggi in larga parte segreti. È proprio su quest'ultimo modulo e sugli strumenti per l'ausilio alla composizione e affinamento delle ricerche che si gioca ancor oggi la sfida scientifica e industriale in questo settore. Quanto migliore è il criterio di rilevanza, tanto più il motore di ricerca sarà in grado di soddisfare i bisogni dell'utente nelle sue prime 10 risposte.

Va da sé che il motore di ricerca potrebbe richiedere all'utente di precisare meglio la sua interrogazione mediante l'aggiunta di altre parole chiave, ma ciò raramente accade: statistiche recenti mostrano che più dell'80% delle interrogazioni è costituito da al massimo due termini e la media è di circa 2.5 termini per

[12] Come accennato, questo approccio è detto *bag of words* a sottolineare che ciò che conta è l'insieme delle parole costituenti una interrogazione e non il loro ordine. Tuttavia il lettore potrà verificare sul suo motore preferito che l'ordine delle parole ha un impatto sui risultati restituiti. I motori offrono anche delle opzioni che consentono di rendere più precisa l'interrogazione; queste però sono poco utilizzate dagli utenti che preferiscono comunque la (limitante) semplicità della sequenza di parole chiave.

interrogazione. Non si tratta solo di pigrizia a comporre interrogazioni selettive, ma anche di difficoltà da parte degli utenti a trovare le parole adeguate per esprimere le loro necessità di informazione. Ciò spiega la continua ricerca verso la realizzazione di strumenti che riescano automaticamente a *capire* i bisogni degli utenti, celati dietro le loro brevi interrogazioni, e il significato delle varie pagine e documenti presenti sul Web.

La situazione non sembra molto diversa da quella descritta in una ricerca pubblicata più di dieci anni fa dal Corriere della Sera (Aprile 2001): su un campione di 856 navigatori italiani tra i venticinque e i cinquantacinque anni che utilizzavano Internet regolarmente, il 33% di questi trovava sempre serie difficoltà nell'uso dei motori di ricerca dell'epoca e il 28% trovava difficoltà solo alcune volte. Tutto ciò induceva il giornalista a concludere che l'uso dei motori di ricerca «*genera stress, frustrazione e senso di smarrimento nel mare del Web. […] E quasi un italiano su tre sogna un motore di ricerca automatico e intelligente che non agisca solo per parole chiave*» (Sottocorona, 2001)[13]. Probabilmente l'alfabetizzazione Web ha modificato queste percentuali, riducendo la frazione di quegli utenti che hanno *sempre serie difficoltà*, ma il sogno di un motore di ricerca *intelligente* sicuramente permane!

Le pagine che seguono descriveranno le caratteristiche salienti delle funzioni di un motore di ricerca notando però che, come è facile immaginare, non tutti gli algoritmi impiegati sono pubblicamente noti. Quindi questa descrizione, seppur parziale, permetterà comunque al lettore di intuire la pregevolezza delle soluzioni algoritmiche progettate fino a oggi per risolvere le difficoltà insite nella *Web Search*.

---

[13] CHIARA SOTTOCORONA, *Sei navigatori su 10 non sanno cercare nel Web*, in «Corriere della Sera», 11 aprile 2001, p. 27. <http://archiviostorico.corriere.it/2001/aprile/12/Sei_navigatori_non_sanno_cercare_co_0_0104124467.shtml>.

Il Crawling

Nel linguaggio informatico il termine *crawling* indica la raccolta dei documenti[14] dal grafo del Web, senza altro scopo che quello di renderli disponibili ai successivi algoritmi di analisi, catalogazione e *indicizzazione*. Un programma di *crawling* è detto in gergo *crawler* (nuotatore), o anche *spider* (ragno) o *robot*[15]. È importante avere chiara la fondamentale differenza tra un browser (per esempio Internet Explorer, Chrome, Opera) e un *crawler*: mentre il primo richiede il recupero di specifiche pagine indicate da un utente, il *crawler* richiede una serie di pagine con un metodo completamente automatico.

In dettaglio, un algoritmo di *crawling* impiega due strutture di dati dette *coda* e *dizionario* con lo stesso significato che hanno questi termini nel linguaggio dei trasporti (una coda di auto) e della linguistica (un dizionario di termini). Una coda *Q* è una serie di elementi allineati in attesa di un servizio. L'elemento che si trova in testa a *Q* è il primo a ricevere tale servizio, e perché ciò avvenga l'elemento viene tolto dalla coda facendo emergere in testa ad essa l'elemento successivo. Per inserire in *Q* un nuo-

[14] Si ricorda l'uso del termine *documento* come generalizzazione del termine *pagina*, per sottolineare la varietà di formati di file che possono esistere sul Web e quindi sono potenzialmente oggetto di raccolta da parte di un *crawler* e di analisi/indicizzazione da parte di un qualunque altro modulo software che compone un motore di ricerca. Nel seguito i due termini saranno interscambiabili.

[15] Questi termini sono fuorvianti: il *crawler* è un programma che risiede in un computer del motore di ricerca e richiede che gli siano inviate le pagine Web tramite il protocollo di trasmissione impiegato dalla rete, senza spostarsi da nessuna parte (in caso contrario potrebbe essere visto come virus informatico), così come l'utente che naviga in Internet rimane sulla propria sedia in attesa che gli siano mostrate le informazioni richieste: più che un navigatore è un turista che consulta cataloghi di paesi lontani sul divano di un'agenzia di viaggi.

vo elemento, questo si pone in fondo alla coda e sarà servito dopo tutti quelli ivi contenuti. Un dizionario *D* è un insieme di elementi (non necessariamente termini di una lingua) che possono essere gestiti in modo più flessibile rispetto alla coda: ciò che qui interessa è che vi sia un metodo rapido per stabilire se un elemento è già presente nel dizionario.

Un *crawler* impiega la coda *Q* per contenere gli indirizzi di documenti del Web da esaminare e due dizionari D_{url} e D_{doc} rispettivamente per gli indirizzi dei documenti già esaminati e i *file* che contengono quei documenti. All'inizio delle operazioni D_{url} e D_{doc} sono vuoti, mentre Q contiene un insieme di indirizzi a partire dai quali il *crawler* inizierà la sua esplorazione del Web. Come è facile immaginare la scelta di questi indirizzi è cruciale per raggiungere documenti di rilevanza generale in tempi ragionevoli: tipicamente si utilizzano gli indirizzi di *portali*, ossia pagine Web che contengono una lista nutrita di risorse presenti su Internet (ad esempio DMOZ, Yahoo), di siti governativi, educativi (ad esempio Wikipedia, Freebase) o di social network, e quindi pagine ricche di link verso altre pagine o risorse importanti del Web.

Il *crawler* opera fintanto che vi sono indirizzi in *Q* da esplorare. Ad ogni passo, esso estrae da *Q* l'indirizzo *U* di un documento Web e verifica che questo non sia stato già visitato (e quindi il suo indirizzo sia presente nel dizionario D_{url})[16]. Se il documento *U* è nuovo, allora il suo contenuto viene analizzato dal *crawler* alla ricerca di link. Questi vengono inseriti in *Q* per un futuro esame se non sono già contenuti in D_{url} o nella stessa *Q*; nel primo caso si tratta di documenti già esplorati e memorizzati in D_{doc}, nel secondo si tratta di documenti già in coda per essere esplorati. Si noti che i proprietari di un sito Web possono

[16] Va da sé che in alcuni casi la visita frequente di alcuni tipi di pagine sarebbe consigliabile, si pensi alle news o ai social network (quali p.e. Twitter e Facebook), che cambiano frequentemente i loro contenuti.

vietare di estrarre informazioni dall'indirizzo *U*, basta loro porre nel sito un file *robots.txt* che specifica al suo interno le pagine a cui è negato il diritto di accesso. Questo può essere necessario per vari motivi, quelli più frequenti sono di protezione delle proprie informazioni e limitazione del numero di accessi al sito che, inevitabilmente, impattano sulla sua banda di accesso e quindi sulla velocità di consultazione.

Va da sé che la condizione di terminazione del processo di *crawling* (ossia *Q* vuota) è difficile da raggiungere vista l'enorme dimensione del Web. In ogni caso, anche se ciò si verificasse a seguito di una inappropriata scelta dei documenti iniziali, il processo di *crawling* si riattiverebbe ri-esplorando il Web e quindi aggiornando possibilmente il contenuto dei documenti in D_{doc}.

Solitamente è il *crawler* che fissa una dimensione massima per D_{url} e interrompe l'esplorazione del Web non appena questa viene raggiunta. In particolare un *crawler* è progettato per ottimizzare tre parametri: il numero *N* di documenti Web gestibili prima che i suoi algoritmi e le sue strutture dati vengano *sopraffatte* dalla dimensione di D_{url}; la velocità *S* con cui è in grado di scaricare documenti dal Web, che oggi raggiunge picchi di migliaia di documenti al secondo; infine la quantità di *risorse computazionali* (tempo, spazio di memoria e disco) utilizzate per portare a termine le operazioni. Chiaramente più grandi sono *N* e *S*, maggiore sarà il costo di mantenimento delle varie strutture dati coinvolte nel processo di *crawling*; di contro, più efficiente è la gestione di queste strutture dati, minore sarà la quantità di risorse computazionali utilizzate e quindi il consumo di energia, problema estremamente rilevante visto l'altissimo numero di computer utilizzati per il funzionamento di questi motori, e maggiore sarà il numero *N* di documenti esplorati a parità di tempo. Studi recenti hanno dimostrato che il Web si rinnova del 30% ogni anno, per cui l'ottimizzazione di questi tre parametri è cruciale e ha conseguenze sul bilanciamento tra freschezza dei contenuti e ampiezza dell'esplorazione.

Vi sono anche altri aspetti che un progettista di *crawler* deve tenere in considerazione: per esempio la riduzione dell'interferenza con il sito esaminato per evitare che questo risulti intasato dalle continue richieste di documenti dei *crawler* a scapito del servizio offerto agli utenti; la scelta dei documenti da ri-esaminare più frequentemente come news, tweet, blog, ecc.; l'uso di tecniche di *calcolo distribuito*[17] e di *resistenza ai guasti*[18] per garantire che il *crawler* non si interrompa mai nel suo funzionamento e raggiunga elevati valori di *S* e di *N*.

L'indicizzazione dei documenti e la ricerca dei risultati

I documenti raccolti dal *crawler* vengono esaminati da un imponente insieme di algoritmi che estraggono da essi alcune informazioni utilizzate successivamente per rispondere efficientemente ed efficacemente alle interrogazioni poste dagli utenti, mediante la costruzione di grandi e sofisticate strutture dati dette *indici*, o più propriamente *liste invertite*[19]. (Zobel, Moffat, 2006)[20] è un articolo di riferimento sulle tecniche di indicizzazione, in particolare sull'impiego delle liste invertite.

---

17 Il calcolo distribuito permette di suddividere un problema in molti compiti, ciascuno dei quali è assegnato e risolto da ogni singolo computer facente parte di un sistema distribuito, ovvero di un insieme di computer autonomi che comunicano tramite una rete.

18 In un sistema distribuito un eventuale guasto ad un computer non compromette l'intera rete.

19 L'aggettivo *invertita* si riferisce al fatto che nei documenti l'ordine delle occorrenze dei termini (dove per occorrenze si intende il numero delle volte in cui un termine ricorre all'interno di un documento) è quello dato dalla sequenza testuale, mentre in queste *posting list* l'ordine è quello che hanno i termini stessi nel dizionario.

20 Cfr. JUSTIN ZOBEL, ALISTAIR MOFFAT, *Inverted Files for Text Search Engines*, in «ACM Computing Surveys», vol. 38, n. 2, 2006, pp. 1-56.

Una lista invertita consta di due parti: il dizionario dei termini, estratti dai documenti raccolti dal *crawler*, e una serie di cosiddette *posting list*, una per termine, contenenti le occorrenze di quel termine più altre informazioni che indicano la *rilevanza* di ogni sua occorrenza.

Più nel dettaglio, un motore di ricerca non indicizza direttamente i documenti raccolti dal *crawler*, ma indicizza una loro versione *estesa* che è ottenuta corredando il documento *d* con un insieme di altre informazioni raccolte nel Web e ad esso correlate. Tra queste informazioni vi sono in particolare i cosiddetti *testi àncora*, cioè porzioni di testo estratte da altri documenti e che circondano i link che puntano a *d*, di cui costituiscono quindi una sorta di descrizione terza presumibilmente affidabile. I motori di ricerca danno molta importanza ai testi àncora perché consentono di ampliare i risultati di una interrogazione recuperando documenti che non contengono le parole richieste ma sono ad essa correlate. Poniamo per esempio che un documento *d* contenga le immagini di varie specie di insetti ma non contenga la parola *insetti* nel suo corpo testuale (o addirittura non contenga alcun testo). È però possibile che un appassionato entomologo realizzi una propria pagina Web con un link a *d* del tipo *belle immagini di insetti*. Tale pezzo di testo è un'àncora per *d*: le parole *belle, immagini, insetti* in esso contenute vengono aggiunte a quelle reperite in *d* e sono considerate fortemente caratterizzanti per tale pagina. Sfortunatamente, come spesso accade nel Web, a un uso pregevole di queste informazioni se ne accompagna un uso malizioso. Nel 1999 effettuando l'interrogazione *more evil than Satan* (più diabolico di Satana) Google restituiva come primo risultato la pagina di Microsoft, probabilmente a seguito della creazione di molte pagine con link a quella di Microsoft e aventi testi àncora contenenti le parole *evil* e *Satan*. Questa situazione imbarazzante fu risolta da Google in poche settimane, ma un incidente simile accadde poi nel novembre del 2003 con l'interrogazione *miserable failure* (fallimento miserevole) e la restitu-

zione nella prima posizione di Google della pagina del presidente George W. Bush. Questo tipo di attacchi ha preso il nome di *Google bombing*, ed è stato ripetuto in diversi luoghi e in diverse lingue sfruttando l'importanza che i motori di ricerca attribuiscono ai testi àncora.

I documenti *espansi* vengono poi analizzati dal motore di ricerca per costruire il *dizionario dei termini*. Si noti che un termine non è solo una parola o, in genere, una sequenza di caratteri alfabetici, ma può essere qualunque componente di un'interrogazione come un numero (730 o 800156156), un'abbreviazione (*e-ticket*), un modello dei nostri oggetti preferiti (N95, B52, Z4), un pezzo di ricambio (BH0241140500), il codice di un corso universitario (AA006), ecc. Il *dizionario dei termini* è dunque molto ampio[21] e deve essere organizzato in modo che si possa rispondere velocemente alle interrogazioni future sul suo contenuto, visto che la verifica di esistenza delle parole chiave di un'interrogazione all'interno di questo dizionario costituisce il primo passo che il motore di ricerca deve eseguire per poter risolvere quella interrogazione. Non è infatti immaginabile che a ogni ricerca di un utente corrisponda una scansione lineare del dizionario, è cruciale quindi disporre di algoritmi e strutture di dati adeguate a gestire efficientemente in tempo e spazio questa grande mole di sequenze di caratteri.

Il dizionario costituisce solo una parte dell'indice costruito dal motore di ricerca, la seconda parte, sicuramente quella più

[21] Numerosi risultati sperimentali hanno dimostrato che il numero n di termini distinti di un documento d segue una legge matematica che ha la forma $n = k\,|d\,|^{\alpha}$, con k pari a qualche decina, $|d\,|$ uguale al numero di parole del documento, e α uguale circa a 0.5. La dimensione attuale del Web indicizzato dai motori di ricerca è di diversi miliardi di documenti, ciascuno con almeno qualche centinaio di termini da cui $n > 10\text{x}10^6 = 10^7$. Quindi il dizionario può contenere decine di milioni di termini distinti, ciascuno di lunghezza arbitraria.

complessa e di dimensioni maggiori, è data dall'insieme delle *posting list*. La *posting list* di un termine *t* contiene la lista dei documenti ove *t* occorre, più tutta una messe di informazioni che indicano, per esempio, la *posizione* e la *rilevanza* di ogni occorrenza nel documento e che verranno utilizzate a tempo di interrogazione per selezionare le pagine più pertinenti con essa. Per descrivere più dettagliatamente il contenuto e la funzione delle *posting list* si consideri la Tabella 1 che riporta un esempio di lista invertita relativamente a un'ipotetica collezione di documenti in italiano.

Dizionario dei termini		**Posting lists**
termine	*#occ*	*<docID: list di posizioni nel documento>*$^+$
...	...	...
calciatore	3	<2: 5,41> <4: 1,10,13,25> <5: 13>
calcio	3	<3: 2,12> <5: 1> <6: 3>
calciopoli	1	<56: 5,41,100,103>
calciare	2	<2: 35> <100: 10, 77, 2000, 2005>
...	...	...

Tabella 1. Un esempio di Lista Invertita costruita su documenti in lingua italiana. La prima colonna riporta i termini che occorrono in quei documenti, ordinati alfabeticamente; la seconda colonna indica il numero di documenti che contengono il termine corrispondente; l'ultima colonna contiene le *posting list* di ogni termine rappresentate secondo il formato indicato in cima alla colonna stessa e descritto del testo. Si noti che il numero di elementi di ogni *posting list* è indicato nella colonna *#occ*. La corrispondenza tra *docID* e indirizzo URL del documento non è indicata in figura, ma memorizzata opportunamente dal motore di ricerca mediante un'altra tabella.

Il dizionario dei termini è ordinato alfabeticamente; questo è importante sia per velocizzare le ricerche delle parole chiave[22],

[22] Molte strutture dati per la ricerca efficiente in dizionari di termini, ad esempio i cosiddetti *trie*, si servono di dizionari ordinati. In generale l'ordinamento dei termini consente di eseguire una ricerca binaria (o dicoto-

sia per avere subito disponibili i termini che condividono un certo prefisso (e che sono contigui nell'ordine), sia per realizzare tecniche di compressione che consentono di ridurre lo spazio occupato dal dizionario mediante la rimozione di prefissi comuni tra i termini (in gergo *Front-Coding compression*).

A ogni termine è associato il numero di documenti che lo contengono, il *docID* (assegnato durante il processo di *crawling*) e le posizioni (in gergo *posting*) in questi documenti ove esso appare. Queste ultime sono solitamente espresse come distanza dall'inizio del documento, misurata *in parole*. La Tabella 1 adotta una particolare codifica per i *posting* di un termine *t*, ossia *<did: p1, p2, p3, ..., pk>*, ove *did* è il *docID* di un documento contenente il termine *t*, e *p1 ...pk* sono le posizioni delle *k* occorrenze di *t* in quel documento. Sia i *docID* che i *posting* sono ordinati in modo crescente, questo per facilitare la realizzazione di particolari tecniche di compressione di interi, e quindi ridurre lo spazio disco occupato. Studi recenti hanno dimostrato che lo spazio occupato dai *docID* è circa 1% dello spazio occupato dal contenuto testuale delle pagine; mentre i *posting* possono richiedere fino a 3-4 volte di più; la compressione dei documenti dipende chiaramente dal loro contenuto ma può arrivare fino a un 10-15% della dimensione originale. Si tratta di risultati estremamente significativi ottenuti nel corso degli ultimi anni dalla ricerca in questo settore. Ciononostante, lo spazio occupato da un motore di ricerca è significativo, e richiede diverse migliaia di dischi, dal momento che i motori indicizzano miliardi di docu-

mica), che consiste nel confrontare la stringa cercata con quella che occupa la posizione centrale nel dizionario; e in base al risultato del confronto alfabetico, si restringe la ricerca in una delle due metà risultanti, ripetendo questa operazione in modo ricorsivo. Inconsapevolmente questo algoritmo formalizza quello che solitamente si esegue su un elenco telefonico nel cercare il nome di un abbonato. L'aspetto interessante di questo approccio è che 20 confronti sono sufficienti per cercare tra un milione di termini!

menti e quindi necessitano di svariati Terabyte per memorizzare i loro contenuti e tutte le strutture dati in forma compressa[23].

La memorizzazione delle posizioni di un termine nei documenti che lo contengono incide significativamente sullo spazio totale richiesto dall'indice, pertanto la scelta di mantenere questa informazione dipende dalla tipologia di interrogazioni che il motore di ricerca è in grado di soddisfare. Nel caso di interrogazioni booleane sui documenti che contengono uno o più termini, sarebbero sufficienti i soli *docID* dei documenti; mentre interrogazioni sull'esistenza di frasi (cioè di più termini consecutivi in un ordine dato) richiedono di conoscere le posizioni dei termini in quei documenti. Alcuni motori usano queste posizioni anche per stimare la rilevanza di un documento rispetto a un'interrogazione, sulla base della distanza tra i termini dell'interrogazione in quel documento.

Va in ultimo osservato che l'ordine dei *docID* è importante anche per eseguire efficientemente un'operazione molto richiesta: individuare i documenti che contengono più parole chiave. Si assuma che un utente abbia formulato un'interrogazione con due termini *t1* e *t2* (l'estensione a un numero maggiore di due è immediata). Il motore innanzitutto cerca nel dizionario i due termini: se almeno uno di essi è assente l'interrogazione non restituisce alcun risultato[24], altrimenti esso recupera le *posting list* relative ai due termini. Siano *t1 = calciatore* e *t2 = calcio*, le liste dei *docID* per essi sarebbero *L1* = 2, 4, 5 e *L2* = 3, 5, 6. Il problema è ora quello di individuare i documenti che contengono sia

[23] È stato già menzionato che Google dichiara di occupare oltre 100 milioni di Gigabyte, equivalenti a 1.000 Terabyte, o anche a 100 Petabyte.

[24] Un utente esperto potrebbe osservare che Google alcune volte restituisce documenti che non contengono tutti i termini specificati nella interrogazione posta. Ciò accade quando il numero di risultati è piccolo e dunque il motore di ricerca amplia lo spazio dei risultati ammettendo che alcune parole chiave siano assenti. Questo prende il nome di *Soft-AND* o *Fuzzy-AND*.

t1 sia *t2*, cioè gli elementi comuni alle due liste *L1* e *L2*. Un approccio possibile potrebbe essere quello che consiste nel confrontare tutti gli elementi di *L1* con quelli di *L2*. Detti *n1 e n2* il numero di elementi delle due liste, è facile rendersi conto che questo algoritmo richiederebbe *n1* x *n2* confronti e questo numero sarebbe troppo grande in presenza di termini frequenti, impedendo dunque al motore di ricerca di rispondere in milli-secondi come in fondo avviene quotidianamente. Infatti i motori adottano un algoritmo che sfrutta l'ordinamento dei *docID*[25] e richiede un tempo proporzionale a *n1 + n2*, quindi lineare nella dimensione delle due liste. L'idea è abbastanza semplice e si serve di due indici *i1* e *i2* che si *muovono* su *L1* e *L2*, partendo dal valore 1 (e quindi dal primo elemento delle liste). A ogni passo l'algoritmo confronta gli elementi in posizione *i1* e *i2* delle liste *L1 eL2*, rispettivamente. Se questi sono uguali, allora è stato scoperto un elemento comune; altrimenti, si fa avanzare l'indice che punta all'elemento più piccolo fra i due. Nell'esempio, il primo passo farebbe avanzare *i1* variando il suo valore da 1 a 2, siccome il primo elemento di *L1* (ossia 2) è più piccolo del primo elemento di L2 (ossia 3). A ogni passo si confrontano due elementi delle liste, dopodiché almeno uno degli indici avanza. Siccome gli elementi in totale sono *n1 + n2*, questo è anche il numero di possibili *avanzamenti* dei due indici e quindi il numero di possibili confronti. Si tratta di un miglioramento significativo rispetto all'algoritmo che eseguiva un confronto tutti-contro-tutti. D'altra parte di meglio non si può fare, visto che in ogni caso occorre esaminare tutti gli *n1 + n2* elementi per poter trovare quelli comuni tra le due liste.

[25] In verità ciò che accade nella pratica è leggermente più sofisticato e tiene conto anche del fatto che all'utente in prima istanza interessano i 10 documenti più rilevanti, quindi non è necessario calcolare tutti i possibili risultati prima di visualizzare i top-10 all'utente che ha posto l'interrogazione.

Valutazione della rilevanza di un documento

La sezione precedente ha illustrato come un motore di ricerca identifica i documenti che contengono le parole chiave di un'interrogazione. Questi possono risultare così numerosi da non consentire a un utente un'analisi diretta, pertanto i motori di ricerca sono *costretti* a identificare automaticamente i documenti più *pertinenti* a quella interrogazione sulla base di opportuni *criteri di rilevanza* che tengono conto sia del contenuto testuale del documento, sia della sua *posizione* nel grafo del Web, sia ancora del *bisogno di informazione* che si cela dietro l'interrogazione dell'utente. Si tratta evidentemente di un problema complesso se non impossibile da risolvere correttamente e completamente!

Ciononostante, per soddisfare al meglio le interrogazioni degli utenti, gli algoritmisti sono stati chiamati a progettare soluzioni che sono diventate sempre più sofisticate e che permettono oggi di calcolare efficientemente un'approssimazione *quantitativa* della rilevanza di un documento. Questa misura viene poi utilizzata dal motore di ricerca per ordinare i suoi risultati, definendo così la *prima pagina* delle top-10 risposte all'interrogazione dell'utente. Questa fase prende il nome di *ranking* e costituisce oggi il punto principale di distinzione tra i più importanti motori di ricerca, tanto che su questo aspetto si concentrano i maggiori segreti di realizzazione e i continui studi portati avanti dalle società del settore e da gran parte della comunità scientifica internazionale. Non è azzardato affermare che uno degli ingredienti principali che hanno permesso a Google di raggiungere un'enorme popolarità è proprio l'algoritmo di *page ranking* impiegato, che ai tempi della sua ideazione (ca. 1998) forniva elenchi di pagine ordinati meglio di quelli degli altri motori di ricerca, in particolare Altavista[26].

[26] Al punto che, a tutt'oggi, la pagina di Google presenta il pulsante *Mi sen-*

Le pagine che seguono dettaglieranno le due principali misure di rilevanza adottate oggi sulle pagine Web, premettendo alcune considerazioni che permetteranno di comprenderne i motivi ispiratori.

Un peso basato sul contenuto testuale: il TF-IDF

È naturale pensare che la rilevanza di un termine *t* per un documento *d* dipenda dalla frequenza (*Term Frequency*) TF[*t*,*d*] con cui *t* occorre in *d* e quindi dal *peso* che l'autore di quel documento ha voluto attribuire a quel termine ripetendolo più volte nel testo. D'altra parte considerare la sola frequenza è fuorviante perché ad esempio gli articoli e le preposizioni occorrono numerose volte nei testi senza caratterizzarli in alcun modo. Quindi è necessario introdurre un fattore correttivo che tenga conto della *capacità di discriminazione* di un termine e sia praticamente nullo nel caso di elementi linguistici secondari. La situazione è però complicata dal fatto che un termine apparentemente significativo, come per esempio *insetti*, può risultare discriminante se la raccolta di documenti indicizzata dal motore è limitata a testi di Informatica (ove l'apparizione del termine *insetti* è inusuale e probabilmente rilevante) oppure può non esserlo se i testi della collezione trattano di entomologia (e quindi l'occorrenza del termine è ovvia, dunque irrilevante). È perciò cruciale considerare anche la rarità di un termine nella collezione, misurandola come rapporto tra il numero *ND* di documenti totali presenti in essa e il numero *N*[*t*] di documenti contenenti il termine *t*. Tanto più il termine *t* è raro quanto più il rapporto *ND/N[t]* è grande, e quindi *t* risulta potenzialmente discriminante per i documenti in cui appare. Solitamente il rapporto non vie-

to fortunato che rimanda l'utente immediatamente al primo risultato senza visualizzare tutti gli altri.

ne utilizzato direttamente per stimare la *capacità di discriminazione* di *t*, ma viene mitigato applicando una funzione logaritmica[27]. Si definisce così il parametro $IDF[t] = log_2 (ND/N[t])$, dove IDF indica *Inverse Document Frequency*, che risulta così poco sensibile a piccole variazioni nel valore di *N*[*t*]. D'altra parte non è detto che un termine raro sia rilevante perché potrebbe corrispondere a una parola desueta o digitata in modo scorretto.

Pertanto le misure di frequenza diretta e inversa sono combinate per formare il cosiddetto *peso testuale* TF-IDF, già proposto alla fine degli anni Sessanta, dato dalla formula:

$$W[t,d] = TF[t,d] \times IDF[t]$$

Si noti che se *t* è per esempio un articolo, esso appare probabilmente in quasi tutti i documenti della raccolta rendendo il rapporto *ND/N[t]* vicino a uno e quindi il suo logaritmo vicino a zero, ovvero IDF[*t*] e *W*[*t,d*]praticamente nulli. Allo stesso modo un termine digitato scorrettamente avrà un valore piccolo per TF, e quindi piccolo sarà il valore di *W*[*t,d*].

Numerosi studi linguistici hanno corroborato la validità empirica della formula precedente che è ora alla base di un qualunque sistema di Information Retrieval. A tal proposito, si può far riferimento ai testi di (Manning et alii, 2008)[28] e (Baeza-Yates, Ribeiro-Neto, 2010)[29], che descrivono le basi generali del-

[27] Il logaritmo in base 2 di un numero x è quell'esponente che occorre dare a 2 per ottenere x. Quindi il logaritmo in base due di 8 è 3, siccome $2^3 = 8$, di 16 è 4, di 32 è 5, e così via. Quindi il suo utilizzo nel calcolo della rilevanza di un dato termine in una collezione di documenti consente di diminuire l'effetto di differenze significative nelle frequenze dei termini stessi.

[28] Cfr. MANNING, C.D., et alii, *op. cit.*

[29] Cfr. RICARDO A. BAEZA-YATES, BERTHIER RIBEIRO-NETO, *Modern Information Retrieval*, ed. 2, Massachusetts, Addison-Wesley, 2010.

l'Information Retrieval, e al testo di (Chakrabarti, 2003)[30], relativo al reperimento e all'analisi dei dati nel Web.

I motori di ricerca di *seconda generazione*, come Altavista, adottavano il peso TF-IDF come parametro preminente per valutare l'importanza di una pagina Web e ordinare di conseguenza i risultati di un'interrogazione.

Questo approccio risultò efficace finché l'accesso al Web era riservato ad agenzie governative e università, e quindi a pagine controllate nei contenuti. A partire dalla metà degli anni Novanta il Web si è aperto a tutta la comunità mondiale diventando un enorme *bazar commerciale*, in cui apparire nei risultati di un motore di ricerca voleva dire essere nella *vetrina del mondo*. Tutto ciò indusse alcune società a costruire pagine Web *truccate* che oltre alle proprie offerte commerciali contenevano, opportunamente occultate, parole chiave tipiche delle interrogazioni più frequenti di tutti gli altri utenti, allo scopo di promuovere arbitrariamente la rilevanza di quelle pagine anche in altri contesti.

Risultò dunque evidente che il peso testuale non poteva essere utilizzato da solo per valutare l'importanza di una pagina, ma occorreva tener conto di altri elementi propri del grafo del Web. A partire così dalla metà degli anni Novanta numerose proposte si susseguirono in ambito accademico e industriale per sfruttare i collegamenti presenti tra le pagine, interpretandoli come un *voto di rilevanza* espresso dall'autore di una pagina *p* verso quelle puntate dai link in uscita da *p*. Due tecniche di *ranking* dettero origine ai cosiddetti motori di ricerca di *terza generazione*: la prima, denominata *PageRank*, fu introdotta da Larry Page e Sergey Brin fondatori di Google; la seconda, denominata HITS (*Hyperlink Induced Topic Search*), fu introdotta da Jon Kleinberg nei laboratori IBM.

[30] Cfr. SOUMEN CHAKRABARTI, *Mining the Web: discovering knowledge from hypertext data*, Morgan Kaufmann Publishers, 2003.

Nel *PageRank* si assegna a ciascuna pagina una rilevanza indipendente dal suo contenuto testuale e dall'interrogazione posta dall'utente, ma dipendente dalla *centralità* della pagina nel grafo del Web. In HITS la rilevanza è invece assegnata in funzione di un sottografo del Web definito a partire dall'interrogazione posta dall'utente. Le tecniche di *PageRank* e HITS sono entrambe definite ricorsivamente perché la rilevanza di una pagina dipende dalla rilevanza delle pagine che puntano ad (o sono puntate da) essa e comportano calcoli su matrici di grandi dimensioni derivate dalla struttura del Web. I motori di ultima generazione combinano le due misure suddette con una serie di altre informazioni *usage based*, tipicamente legate all'attività di navigazione, ricerca e interazione (leggi uso dei social network) che gli utenti realizzano quotidianamente sul Web.

L'obiettivo ultimo è quello di *profilare l'utente* inferendo tutta una serie di informazioni utili per identificare i bisogni di informazione che si celano dietro le (poche) parole chiave di una sua interrogazione, e quindi ordinare meglio i risultati recuperati dal motore di ricerca.

Un peso basato sui link: il 'PageRank'

L'algoritmo del *PageRank*, che ha segnato, dal punto di vista scientifico l'inizio dell'era di Google, è stato proposto nello storico articolo di (Brin, Page, 1998)[31]. Siccome il *PageRank* sfrutta le connessioni presenti nel grafo del Web e siccome è più naturale parlare di pagine aventi hyper-link, piuttosto che documenti, nel seguito di questa sezione si utilizzerà il termine *pagina* in luogo di *documento*, sottointendendo comunque che quan-

[31] Cfr. SERGEY BRIN, LARRY PAGE, *The anatomy of a large-scale hypertextual Web search engine*, in Computer Networks and ISDN Systems, vol. 30, n. 1-7, Elsevier Science Publisher, 1998, pp. 107-117.

to sarà detto si estende naturalmente dalle pagine HTML a un qualunque tipo di documento purché dotato/dotabile di link.

Il *PageRank* ordina le pagine in funzione della loro *popolarità* nel grafo del Web misurata in base al numero e alla provenienza degli archi entranti in ogni pagina, cioè dei link che puntano a essa. Esistono diverse interpretazioni del *PageRank*, alcune più informali ma intuitive, altre più matematiche. In generale il *PageRank* di una pagina *p* è espresso da un valore *PR*(*p*), calcolato in prima approssimazione come segue. Dette $p_1,\ldots, p_k$ le pagine che hanno almeno un link verso *p*, e detto $N(p_i)$ il numero di pagine puntate da ogni pagina p_i (cioè il numero di archi uscenti da p_i nel grafo del Web), la formula di base per il calcolo di *PR*(*p*) è la seguente:

$$\mathrm{PR}(p) = \sum_{i=1\ldots k} (\mathrm{PR}(p_i)/\mathrm{N}(p_i)).$$

Solo le pagine che puntano a *p* contribuiscono al valore di *PR(p)* e il contributo di ciascuna è dato dal proprio *PageRank* diviso per il numero di archi uscenti da essa. È come se il *PageRank* di p_i venisse suddiviso equamente tra tutte le $N(p_i)$ pagine a cui p_i punta; pertanto *p* riceve una frazione pari a $1/N(p_i)$ di quel valore.

Ad esempio, si consideri nella Figura 2 la pagina *P* che ha un *PageRank* pari a 53. Questo valore è ottenuto sommando il contributo proveniente dalla pagina *P1* pari a 50=100/2 (siccome *P1* ha due link uscenti e *PR*(*P1*)=100) e il contributo della pagina *P2* pari a 3=9/3 (siccome *P2* ha tre link uscenti e *PR*(*P2*)=9). È evidente che questa formula è *ricorsiva*, visto che il *PageRank* di una pagina è definito in funzione del *PageRank* di altre pagine. Tale formula inoltre può indurre un calcolo *ciclico*: infatti, se aggiungiamo un link tra la pagina *P* e la pagina *P1* (o *P2*), allora il *PageRank* di quest'ultima pagina sarebbe a sua volta influenzato dal *PageRank* di *P*. Il grafo del Web è ricco di cicli e quindi induce molte dipendenze cicliche della definizione di *PR*. Que-

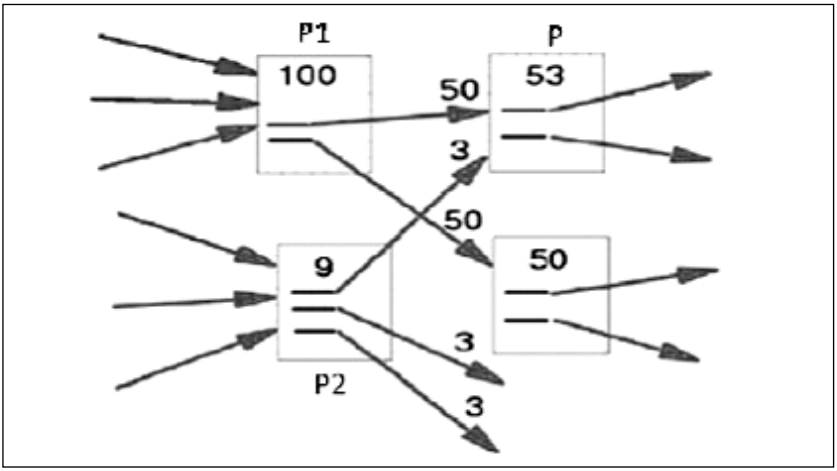


Figura 2. Una porzione di grafo del Web con quattro pagine, per ciascuna sono indicati i link uscenti ed entranti, e il valore del *PageRank*. Inoltre, per ogni link si indica il contributo al *PageRank* proveniente dalla pagina sorgente del link stesso.

sta struttura matematica però pone dei problemi di *stabilità* nella formula, in quanto è naturale chiedersi: (1) con quali valori sia opportuno inizializzare i *PR* prima di iniziare il calcolo; (2) come garantire la convergenza del calcolo dei *PR* a un numero ben preciso vista la ciclicità e ricorsività nella loro definizione.

In effetti è facile convincersi che se una pagina non ha link entranti il suo *PageRank* non cambierà mai, risultando quindi cruciale l'assegnazione iniziale di questo. Di contro, se una pagina non ha link uscenti questa si comporterà come una sorta di *pozzo* accumulando il *PageRank* che proviene dai suoi archi entranti, senza redistriburlo a nessuno. D'altra parte la formula precedente è promettente in quanto se una pagina rilevante p_i punta a p ne incrementa la sua rilevanza di una quantità che suddivide equamente $PR(p_i)$ tra tutte le pagine da puntate da p_i. Quindi una pagina p per risultare rilevante deve essere puntata o da poche pagine, esse stesse molto rilevanti (deve essere *legata* a una eli-

te), oppure da moltissime pagine poco rilevanti (deve essere popolare).

Così, per riuscire a risolvere i problemi tecnici legati alla formula precedente, i due fondatori di Google (Brin e Page) suggerirono di considerare una formula leggermente diversa dalla precedente che introduceva un fattore correttivo il quale attribuiva un *PageRank* di partenza costante a tutte le pagine pari a *(1-c)/n*, dove c è un parametro che verrà definito nel seguito e *n* è il numero di pagine indicizzate dal motore di ricerca:

$$PR(p) = c \sum_{i=1\ldots k} (PR(p_i)/N(p_i)) + (1-c)/n.$$

Essi ponevano c=0.85 perché questo valore offriva i risultati sperimentali migliori. Al limite se si ponesse c=0 tutte le pagine avrebbero pari rilevanza $1/n$; se si ponesse c=1, la rilevanza di una pagina dipenderebbe completamente dalla struttura del grafo del Web e quindi equivarrebbe alla prima formula introdotta all'inizio di questa sezione, con i problemi tecnici già discussi. La scelta c=0.85 attribuisce, ragionevolmente, rilevanza maggiore al contributo che deriva dalla struttura del Web.

La discussione precedente ha evidenziato l'aspetto *sociale* del *PageRank*, quello cioè legato all'interpretazione dei link come un *voto di fiducia* o *manifestazione di interesse* da parte di una pagina verso un'altra, per cui una pagina è rilevante *nella società del Web* se popolare, e quindi molte altre pagine la puntano, o se appartenente a una elite, poche la puntano ma queste sono esse stesse rilevanti. In letteratura esistono altre interpretazioni più matematiche della formula del *PageRank*, più utili per studiare le sue proprietà e per derivare un algoritmo efficiente di calcolo.

Qui si menzionerà, per limiti di spazio, solo l'interpretazione detta del *navigatore casuale del Web*. Il valore $PR(p)$ lo si può interpretare anche come la frequenza con cui un utente può raggiungere p camminando a caso sul grafo del Web ed eseguendo il seguente algoritmo: a ogni passo il navigatore si trova su una

pagina *p*, a questo punto egli sceglie con probabilità *c* se seguire i link presenti in *p*, e con probabilità (1-*c*) se saltare a una qualunque pagina indicizzata dal motore di ricerca (passo di *teletrasporto*); successivamente, se il primo caso occorre, allora il navigatore sceglierà uno dei link uscenti da *p* con probabilità uniforme $1/N(p)$ (ovvero con la medesima probabilità per ciascun link di essere scelto) raggiungendo così un'altra pagina Web, nel secondo caso un'altra pagina è stata già raggiunta dal salto casuale. Questo processo di esplorazione del grafo del Web viene ripetuto indefinitamente, grazie al passo di teletrasporto e indipendentemente dalla struttura del Web. Si può comunque dimostrare, facendo uso di una teoria matematica ben nota (quella delle cosiddette Catene di Markov[32]) che al limite $PR(p)$ converge alla frequenza con cui il *navigatore casuale del Web* visiterà la pagina *p* durante la sua percolazione del Web. Questo risultato ha un valenza duplice: da una parte dimostra che $PR(p)$ è una valida quantificazione della popolarità della pagina *p*, dall'altra, che la formula su descritta ammette una soluzione unica e che questa non dipende dai valori con cui i $PR()$ sono inizializzati.

Si pone a questo punto il problema di calcolare i *PR* in modo efficiente, sicuramente un compito non banale visto che i motori di ricerca moderni indicizzano diversi miliardi di pagine (n quindi supera abbondantemente il valore 10^9) e quindi operano su grafi che hanno dimensioni enormi, tali cioè da richiedere l'impiego di migliaia di computer per poter essere manipolati in breve tempo.

Ma è d'uopo osservare che non interessa il valore finale dei $PR(p)$ quanto piuttosto solo l'ordinamento tra essi: se $PR(p1) > PR(p2)$ allora *p1* è più importante di *p2*. Pertanto quando il va-

[32] Una catena di Markov è un processo stocastico caratterizzato da un insieme di stati nel quale la probabilità di transizione che determina il passaggio ad uno stato del sistema dipende unicamente dallo stato del sistema immediatamente precedente.

lore delle varie componenti di *PR* risulta sufficientemente stabile, il calcolo suddetto viene interrotto. Prove sperimentali hanno dimostrato che un centinaio di iterazioni sono solitamente sufficienti, rendendo il calcolo praticabile in tempi ragionevoli se si dispone di data-center con qualche migliaio di computer.

Per concludere, due osservazioni sono dovute; la prima riguarda l'uso di *PR*(*p*) nella risoluzione di una interrogazione, la seconda, più sofisticata,riguarda la *tipologia di link* coinvolti nel calcolo del *PageRank*.

Il *PageRank* induce un ordinamento delle pagine che è funzione solo della struttura del grafo e non dipende quindi in alcun modo dalla interrogazione posta dall'utente (si dice per questo che la rilevanza delle pagine è *query independent*). Pertanto il *PageRank* può essere calcolato durante la fase di indicizzazione delle pagine e memorizzato con esse, per poi essere recuperato a tempo di interrogazione per ordinare le pagine che contengono i termini della ricerca. I dettagli d'uso del *PageRank* in Google sono sconosciuti ma, come menzionato nelle pagine precedenti, Google combina questo peso con il TF-IDF e con alcune centinaia di altri parametri estratti automaticamente e manualmente dal Web.

Per quanto riguarda i link, la formula precedente del *PageRank* li considera tutti equivalenti e quindi assume che il *navigatore casuale* possa percorrere uno qualunque dei link che escono da una pagina e possa teletrasportarsi a una qualunque delle pagine del Web indicizzate dal motore. D'altra parte i link non sono tutti *uguali*, nel senso che si possono solitamente classificare in quattro grandi gruppi: *topical link* vs *templatic link*, *egoistic link* vs *altruistic link*. I primi collegano pagine che trattano dello stesso argomento mentre i secondi sono spesso parte di schemi HTML applicati per progettare siti Web e collegano pagine solitamente dissimili (per esempio, collegano i documenti sulla privacy di un sito o i contatti o la sezione ove spedire una email); i terzi sono spesso il risultato di un'azione di spam che

porta a collegare pagine dello stesso autore per *forzare maliziosamente* il calcolo del *PR*, mentre gli ultimi sono i link più pregiati per la significatività del *PR* e sono quelli inseriti da autori di pagine Web con lo scopo di indicare pagine autorevoli su un certo argomento. Chiaramente il primo e il quarto gruppo di link sono quelli più interessanti ai fini del calcolo della rilevanza di una pagina, e i motori di ricerca moderni hanno sviluppato tecniche molto sofisticate per poterli identificare correttamente attingendo al già menzionato *usage* del Web da parte degli utenti. In questo caso le *toolbar* gratuite dei motori Google e Bing svolgono un ruolo importante nella raccolta di tali informazioni.

Altre funzioni di un motore di ricerca

Tra le altre operazioni che un motore di ricerca è chiamato a eseguire ha grande importanza il modo con cui si presentano i risultati all'utente. Il dizionario D_{doc}, che contiene le pagine raccolte dal *crawler*, è utilizzato per restituire brevi frammenti testuali, noti come *snippet*, che riportano il contesto di occorrenza dei termini dell'interrogazione in ogni pagina dei risultati e offrono la possibilità di visualizzare la copia originale della pagina recuperata dal *crawler* che nel frattempo potrebbe essere stata modificata o potrebbe addirittura essere scomparsa dal Web. Queste due funzionalità permettono all'utente di valutare la significatività dei risultati o di trovarne alcuni non più disponibili sul Web ma comunque interessanti per la sua interrogazione.

Le risposte dei motori di ricerca sono a volte corrotte ad arte con sofisticate tecniche di spam che promuovono fraudolentemente la rilevanza di alcune pagine per farle apparire tra le prime nella risposta, o alterano la risposta stessa in modi più subdoli. Così per esempio una tecnica di attacco chiamata *cloaking* impiega copie di pagine presumibilmente rilevanti per gli utenti (per esempio tratte da Wikipedia) per mascherarne altre con con-

tenuti del tutto diversi. Se la pagina artefatta è rilevante per l'interrogazione di un utente, il motore di ricerca visualizzerà uno *snippet* appropriato e interessante per essa, ma l'utente, cliccando sul link contenuto nello *snippet*, scaricherà una pagina irrilevante se non offensiva nei contenuti. Una discussione sui metodi di spam non può entrare in questo breve capitolo, ma si faccia riferimento allo studio di (Fetterly, 2007)[33], che contiene una recente rassegna sulle tecniche di spamming per ingannare i *crawler.*

È bene, infatti, sapere che questi fenomeni hanno un'estensione inaspettatamente grande poiché si stima che più del 20% del Web sia costituito da pagine artefatte che mettono a repentaglio la reputazione e l'utilità dei motori di ricerca. In tutte le loro fasi di funzionamento questi adottano quindi algoritmi antispam molto sofisticati per prevenire il recupero, l'indicizzazione ed eventualmente la restituzione di pagine artefatte. È ovvio che questi algoritmi vengono solo parzialmente svelati!

Recentemente l'obiettivo dei motori di ricerca si è spostato verso l'individuazione dell'*intento* che si cela dietro l'interrogazione posta dall'utente, oltre che sulla sua composizione puramente *sintattica*. Ciò spiega il moltiplicarsi di metodi diversi per presentare adeguatamente le risposte sullo schermo (iniziati con l'esperienza del motore Vivisimo), per integrare diverse sorgenti di informazione (news, Wikipedia, immagini, video, blogs, prodotti, ecc.), o per fornire suggerimenti alla composizione delle interrogazioni (ad esempio, Google Suggest). A ciò si aggiunga che gli utenti, da *attori attivi* del processo di ricerca, stanno purtroppo diventando sempre più *spettatori passivi*: pubblicità, suggerimenti, meteo, amici connessi, news personalizzate, ecc., sono tutte *informazioni* che abbiamo probabilmente indicato come *interessanti* in qualche scheda personale o che i motori han-

[33] Cfr. DENNIS FETTERLY, *Adversarial Information Retrieval: the manipulation of Web content*, in «ACM Computing Reviews», 2007.

no stabilito che sono per noi *potenzialmente interessanti*: ciò alla luce dei nostri comportamenti di ricerca sul Web che i motori sottopongono a un continuo scrutinio. Tutte queste informazioni appaiono già sullo schermo quando si legge l'email, sulle pagine personali di iGoogle o myYahoo!, sulla pagina correntemente visitata sul Web, o anche sul nostro navigatore satellitare, senza che un utente ne abbia fatto esplicita richiesta.

Molte altre caratteristiche dei motori di ricerca meriterebbero di essere studiate attentamente sia dal punto di vista algoritmico sia da quello dell'organizzazione generale e dell'impiego.

D'altra parte le pagine precedenti possono risultare preziose non soltanto per chi voglia addentrarsi nei meandri algoritmici dei motori di ricerca, ma anche per coloro che progettano le pagine Web, visto che più dell'85% degli utenti giunge a una pagina attraverso un motore di ricerca guardando soltanto ai primi 10 risultati e più del 33% di questi utenti ritiene che i primi 10 risultati siano *il posto migliore dove spendere i propri soldi*. Tra le indicazioni utili si ricavano le seguenti: il titolo della pagina e i metatag devono essere brevi e devono contenere tutte le parole chiave che descrivono quella pagina Web e che potenziali clienti utilizzeranno per cercarla; il primo paragrafo gioca un ruolo determinante, ed è opportuno che le parole chiave figurino con una certa frequenza nel corpo dell'intera pagina; per assicurare un buon *PageRank* alla pagina occorre assicurare che pagine rilevanti puntino a essa, e che questi link contengano un testo àncora con le parole chiave che la descrivono opportunamente (tra queste pagine troviamo quelle di siti governativi, università, Wikipedia, news, ecc.). Si osserva che link provenienti da pagine autorevoli sono cruciali anche per garantire che la pagina sia recuperata dal *crawler* del motore di ricerca. Per concludere, si suggerisce l'uso di alcuni strumenti che consentono di valutare la buona composizione di una pagina Web: Google, per esempio, ha messo a disposizione degli utenti lo strumento *Sitemaps*; Overture invece offre *Word Tracker* che consente di scoprire

quali interrogazioni sono state frequentemente eseguite negli ultimi mesi dagli utenti su un determinato argomento. In tal senso, è chi scrive a doversi adeguare al linguaggio utilizzato dagli utenti e non viceversa.

Verso una ricerca 'semantica'

Oggi si considera che i motori di ricerca siano ancora nella loro infanzia, sebbene abbiano già conosciuto grandi evoluzioni e i loro algoritmi complessi permettano di impiegarli con una certa facilità. Il limite consiste nell'indicare documenti rilevanti selezionati essenzialmente tra quelli che contengono le parole chiave specificate dall'utente. Ma ben altro si potrebbe desiderare!

L'importanza di muoversi verso la costruzione di un motore di ricerca capace di *interpretare* le richieste degli utenti al di là del semplice esame delle parole chiave è stata da tempo indicata con grande determinazione dallo stesso Berners-Lee, l'inventore del Web. In tal senso si parla di *Web semantico* come prossima forma di impiego della rete. Le direzioni verso cui esso evolve e le ricerche che lo interessano vengono discusse in (Baeza-Yates et alii, 2008)[34], (Microyannidis, 2007)[35], (Weikum et alii, 2012)[36] e (Spaniol et alii, 2012)[37].

---

[34] Cfr. Ricardo A. Baeza-Yates, Massimiliano Ciaramita, Peter Mika, Hugo Zaragoza, *Towards semantic search*, in Natural Language to Information Systems, in «Lecture Notes in Computer Science (LNCS)», vol. 5039, 2008, pp. 4-11.

[35] Cfr. Alexander Microyannidis, *Toward a social semantic Web*, in «Computer», vol. 40, n. 11, IEEE Computer Society Press, 2007, pp. 113-115.

[36] Cfr. Gerhard Weikum, Johannes Hoffart, Ndapandula Nakashole, Marc Spaniol, Fabian M. Suchanek, Mohamed Amir Yosef, *Big Data Methods for Computational Linguistics*, in «IEEE Data Engineering Bulletin», vol. 35, n. 3, 2012, pp. 46-55.

[37] Cfr. Marc Spaniol, András A. Benczúr, Zsolt Viharos, Gerhard

Ingredienti fondamentali saranno l'uso di tecniche algoritmiche per l'elaborazione di testi in linguaggio naturale e per l'apprendimento automatico: campi in cui già esistono importanti esperienze sviluppate in ambiti diversi, che iniziano a essere dirette verso la ricerca sulla rete. Per quanto riguarda l'organizzazione dei dati per il nuovo Web qualche passo significativo è stato fatto ma è difficile per ora valutarne la portata. Il linguaggio RDF (*Resource Description Framework*)[38] consente di aggiungere ai dati contenuti nelle pagine un'informazione che ne descrive il *significato* in termini molto liberi, che per ora si concentrano nell'indicare l'appartenenza di un dato a una classe predefinita e le sue relazioni con altri dati. Non si pretende che un computer *comprenda* una richiesta ma quanto meno possa classificarla e metterla in relazione con dati contenuti in altre pagine che non contengono le parole chiave della richiesta originale.

Alcuni sforzi recenti stanno offrendo metodi automatici per aggiungere *semantica* alle pagine Web. Una prima sorgente di queste informazioni è il cosiddetto *Web 2.0* comprendente un processo di *tagging* che porta milioni di utenti ad associare termini o frasi alle proprie immagini, video, pagine e ogni possibile file che occorre sul Web. Questi termini stanno dando vita, attraverso un vero e proprio *linguaggio* parallelo, alle cosiddette *folksonomy*, ossia tassonomie pubbliche di concetti emergenti in una comunità. Tra queste ricordiamo: Flickr, Technorati, YouTube, Del.icio.us, Panoramio, CiteULike, Last.fm, ESP Game, ecc. Nati con l'obiettivo di *classificare* i propri oggetti per facilitarne il loro recupero personale, questi sistemi di tagging assurgono a un ruolo sempre più importante nei motori di ricerca, dimostrando la loro efficacia nel migliorare le ricerche, individuare lo spam, creare nuove modalità di comunicazione e analisi dei da-

---

WEIKUM, *Big Web Analytics: Toward a Virtual Web Observatory*, in «ERCIM News», vol. 1, n. 89, 2012, pp. 23-24.

38 <http://www.w3.org/RDF/>.

ti, identificare nuovi soggetti che caratterizzano un settore.

D'altra parte il *tagging manuale* è particolarmente costoso, per cui alcuni gruppi di ricerca stanno investigando la possibilità di automatizzare questo processo arricchendo i documenti con *annotazioni strutturate* il cui obiettivo precipuo è quello di identificare entità linguistiche preminenti per il documento in input (nomi, città, società, eventi, elementi linguistici, e molto altro ancora) e di assegnare loro una *interpretazione* che a oggi viene codificata mediante un link a una pagina di Wikipedia che descrive quella entità[39]. La scelta di Wikipedia risiede nel fatto che questo catalogo è oramai disponibile in numerose lingue ed è di dimensioni significative, soprattutto per l'Inglese per cui si contano circa 4 milioni di articoli (le pagine in Italiano sono circa 1 milione).

Per comprendere l'utilità di questa annotazione per i motori di ricerca, si consideri la frase *Diego Maradona ha vinto contro il Messico* che fu il titolo di una notizia battuta durante l'ultima Coppa del Mondo di calcio. Questi annotatori identificherebbero *Diego Maradona* e *Messico* come entità significative per quel testo e le legherebbero alle pagine di Wikipedia che parlano del famoso ex-calciatore e della squadra di calcio del Messico. Quindi un motore di ricerca potrebbe identificare quella notizia come rilevante se l'utente cercasse informazioni sulla squadra di calcio del Messico, oppure non restituirla se l'intenzione dell'utente fosse quella di fare un viaggio in Messico o se fosse interessato a notizie storiche sulla civiltà dei Maya.

Questo tipo di annotazione è anche utile per scoprire la similarità tra testi anche se questi non condividono alcun termine, superando così le difficoltà che sono insite nel modello *bag of words*. La ragione di tutto ciò risiede nel fatto che l'annotazione

[39] Tra gli annotatori noti ricordiamo Tagme (<http://acube.di.unipi.it/tagme>) e Wikipedia Miner (<http://wikipedia-miner.cms.waikato.ac.nz>).

associa entità a pagine di Wikipedia le quali *vivono* all'interno di un grafo i cui collegamenti sono gli hyper-link definiti dagli autori di quelle pagine. Questi link sono abbastanza autorevoli, per cui se tra due nodi esiste un link o un cammino di pochi archi che li connette, allora è ragionevole attendersi che le due pagine corrispondenti siano in qualche modo correlate tra loro. In una parola, le due pagine sono *simili*. Ne consegue quindi che sfruttando l'*annotazione strutturata* di due documenti e calcolando i cammini che connettono le loro entità nel grafo di Wikipedia, si potrebbe stabilire con buona approssimazione se le due frasi stanno *parlando* dello stesso argomento. Ciò sarebbe impossibile se si utilizzasse la tecnica classica del *bag of words*. Non sorprende dunque il recente annuncio da parte di Google della creazione di un *knowledge graph* di circa 500 milioni di entità e più di 3.5 miliardi di connessioni. Secondo quanto riportato da comunicazioni ufficiali, questo grafo è stato creato analizzando una serie di archivi di dominio pubblico quali Wikipedia, Freebase, CIA World Factbook e molti altri ancora non svelati. Google dichiara di usare il *knowledge graph* in tre modi diversi: per perfezionare i risultati prodotti da una interrogazione, conoscendo i *significati* delle sue parole chiave; per trovare inaspettate relazioni tra entità (siano esse l'interrogazione e i potenziali documenti pertinenti ad essa); per consentire agli utenti di navigare lo *spazio dei risultati* correlati a una interrogazione. Gli utenti dovranno dunque abituarsi all'uso intensivo dei grafi come modello della conoscenza e gli informatici dovranno perfezionare gli algoritmi per l'elaborazione di grafi di enormi dimensioni!

Un'altra sorgente di informazione efficace nel migliorare il processo di ricerca e analisi del Web è rappresentata dall'insieme delle ricerche eseguite dagli utenti. Queste informazioni sono catturate dai motori di ricerca, memorizzate in cosiddetti *Query Log* e poi successivamente elaborate al fine di estrarre relazioni di *vicinanza semantica* tra interrogazioni e/o pagine del Web. Si considerino due interrogazioni *q1* e *q2* formulate da molti uten-

ti, i quali hanno poi raggiunto una stessa pagina *p* che appariva nei risultati restituiti dal motore di ricerca per quelle due interrogazioni. Ciò significa che *p* ha probabilmente a che fare con gli argomenti di *q1* e *q2* e potrebbe essere quindi etichettata con le loro parole chiave, arricchendo potenzialmente il vocabolario di termini estratti da *p* in fase di indicizzazione o rendendo quei termini più rilevanti per *p* nelle ricerche future. Allo stesso modo, il click condiviso sulla pagina *p* potrebbe far dedurre che *q1* e *q2* sono *correlate* e quindi potrebbero l'una costituire un valido suggerimento per l'altra. Ad esempio le due interrogazioni *iTunes* e *iPod* restituiscono su Google, nel momento in cui scriviamo, la pagina <http://www.apple.com/itunes/> come primo risultato. Quindi si può ipotizzare che numerosi utenti raggiungeranno questa pagina determinando un *link semantico* tra i due termini.

D'altra parte è evidente che gli utenti possono scegliere diverse pagine tra quelle restituite dal motore di ricerca a partire da una stessa interrogazione *q*. In questo caso può accadere che queste pagine siano tutte correlate con *q*, e quindi dovrebbero essere considerate *simili* tra loro; oppure che *q* sia una interrogazione *polisemica* e quindi le pagine raggiunte successivamente dagli utenti possono avere diversi significati ed essere così suddivise in gruppi di pagine *simili*, ciascuno di essi rilevante per una particolare *interpretazione* di *q*. Si pensi ad esempio all'interrogazione *eclipse*: questa potrebbe indicare il termine inglese per *eclissi* (lunare o solare), oppure un sistema di sviluppo software o un modello di aereo o di una macchina o di un autoradio, ecc. Quindi nel *Query Log* si avranno per esempio utenti che hanno richiesto pagine di Wikipedia per documentarsi sulle eclissi lunari o solari o su pagine del progetto *eclipse.org* o su pagine che descrivono il jet Eclipse 500 o ancora su pagine della Mitsubishi. In ogni caso l'analisi di queste interrogazioni e di questi click (chiamati *usage del Web* nelle sezioni precedenti), della struttura del Web, del contenuto di quelle pagine così come di altre informazioni, porterà all'identificazione di gruppi di pagine simili e

alla caratterizzazione di *q* come interrogazione polisemica. Pertanto *q* dovrà essere *trattata* dal motore di ricerca con cautela, nel senso che i primi risultati visualizzati per essa dovranno tener conto della sua polisemia e quindi essere allo stesso tempo diversificati e completi, nel senso che dovranno rappresentare nel miglior modo possibile le diverse sfaccettature semantiche racchiuse nei termini che costituiscono *q*.

Dal punto di vista algoritmico tutte le relazioni estratte dai *Query Log* e dalle altre sorgenti di informazione su citate (Web, contenuto, *click* di navigazione,…) vengono modellate ancora una volta mediante un grafo di enormi dimensioni sia in termini di numero di nodi (pagine e interrogazioni) sia in termini di archi che li connettono (relazioni *sintattiche* o *semantiche*). L'analisi strutturale di questo grafo ha consentito in questi ultimi anni di estrarre numerose informazioni utili su come il costituirsi di folksonomy abbia influenzato le ricerche sul Web, sulle comunità di utenti e sui loro tendenziali interessi e sulle pagine rilevanti in quanto raggiunte frequentemente dagli utenti dei motori di ricerca.

Va da sé che gli approcci di assegnazione automatica di significato agli oggetti del Web sono proni a errori incontrollati. Ma la *conoscenza collettiva* generata da questo gigantesco processo che coinvolge quotidianamente miliardi di utenti sembra sufficiente a ridurre il manifestarsi di questi errori; ciò limita la necessità di un intervento correttivo e rende il processo di estrazione di conoscenza potenzialmente scalabile alla dimensione corrente del Web.

Bibliografia

ALPERT, J., HAJAJ, N., *We knew the Web was big*, Official Google Blog, luglio 2008 <http://googleblog.blogspot.it/2008/07/we-knew-Web-was-big.html>

BAEZA-YATES, R.A., CIARAMITA, M., MIKA, P., ZARAGOZA, H., *Towards semantic search*, in Natural Language to Information Systems, in «Lecture Notes in Computer Science (LNCS)», vol. 5039, 2008, pp. 4-11

BAEZA-YATES, R.A., RIBEIRO-NETO, B., *Modern Information Retrieval*, ed. 2, Massachusetts, Addison-Wesley, 2010

BRIN, S., PAGE, L., *The anatomy of a large-scale hypertextual Web search engine*, in Computer Networks and ISDN Systems, vol. 30, n. 1-7, Elsevier Science Publisher, 1998, pp. 107-117

CHAKRABARTI, S., *Mining the Web: discovering knowledge from hypertext data*, Morgan Kaufmann Publishers, 2003

FERRAGINA, P., MANZINI, G., *On compressing the textual Web*, in Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM), New York, 4-6 febbraio 2010, pp. 391-400

FERRAGINA, P., SCAIELLA, U., *Fast and accurate annotation of short texts with Wikipedia pages*, in «IEEE Software», vol. 29, n. 1, 2012, pp. 70-75

FETTERLY, D., *Adversarial Information Retrieval: the manipulation of Web content*, in «ACM Computing Reviews», 2007

MANNING, C.D., RAGHAVAN, P., SCHUTZE, H., *Introduction to Information Retrieval*, Cambridge University Press, 2008

MICROYANNIDIS, A., *Toward a social semantic Web*, in «Computer», vol. 40, n. 11, IEEE Computer Society Press, 2007, pp. 113-115

SOTTOCORONA, C., *Sei navigatori su 10 non sanno cercare nel Web*, in «Corriere della Sera», 11 aprile 2001, p. 27 <http://archiviostorico.corriere.it/2001/aprile/12/Sei_navigatori_non_sanno_cercare_co_0_0104124467.shtml>

SPANIOL, M., BENCZÚR, A.A., VIHAROS, Z., WEIKUM, G., *Big Web Analytics: Toward a Virtual Web Observatory*, in «ERCIM News», vol. 1, n. 89, 2012, pp. 23-24

WEIKUM, G., HOFFART, J., NAKASHOLE, N., SPANIOL, M., SUCHANEK, F.M., AMIR YOSEF, M.A., *Big Data Methods for Computational Linguistics*, in «IEEE Data Engineering Bulletin», vol. 35, n. 3, 2012, pp. 46-55

WITTEN, I.H., MOFFAT, A., BELL, T.C., *Managing Gigabytes*, Morgan Kaufmann Publishers, 1999

WITTEN, I.H., GORI, M., NUMERICO, T., *Web Dragons*, Morgan Kaufmann Publishers, 2007

ZOBEL, J., MOFFAT, A., *Inverted Files for Text Search Engines*, in «ACM Computing Surveys», vol. 38, n. 2, 2006, pp. 1-56

Sitografia

<http://www.google.com/insidesearch/playground/underthehood.html>
<http://wikipedia-miner.cms.waikato.ac.nz>
<http://acube.di.unipi.it/tagme>

Classificazioni bibliografiche

MAURO GUERRINI*

Classificazione deriva etimologicamente dal sostantivo latino *classis*, termine che si riferisce alle cinque categorie in cui fu divisa la città di Roma, in base al patrimonio fondiario; dalla stessa radice deriva l'espressione *classi sociali*. Nel tempo gli uomini hanno tentato di organizzare e disporre in un qualche ordine le conoscenze acquisite. Classificare significa raggruppare oggetti o concetti che presentano aspetti comuni, distinguendoli da oggetti e concetti che non possiedono queste caratteristiche. Gli elementi in comune in una classe costituiscono caratteristiche o *principi di divisione*. Scrive (Chan, 1981):

> *Classificare, definito in senso ampio, è l'atto di organizzare l'universo del sapere in qualche ordine logico sistematico. È stata considerata l'attività più importante della mente umana. Classificare consiste nel processo dicotomico di distinguere cose o oggetti che possiedono certe proprietà o caratteristiche da quelle che non le possiedono, e successivamente, di raggruppare in una classe cose o oggetti che hanno proprietà o caratteristiche comuni*[1].

* Università di Firenze, Dipartimento di Storia, Archeologia, Geografia, Arte e Spettacolo.

1 LOIS MAI CHAN, *Cataloging and classification: an introduction*, New York, McGraw-Hill, 1981, p. 209.

Classificare significa, quindi, suddividere un insieme in due o più parti secondo un determinato criterio fino ad arrivare a *concetti individuali*. Per illustrare questo processo di suddivisione che procede dal *generale* al *particolare* si ricorre a una rappresentazione nota come *Albero di Porfirio*, che bene esemplifica il metodo di biforcazione – o dicotomia – degli schemi classificatori gerarchici. Porfirio ritiene che il genere sia «*ciò a cui è subordinata la specie*» e la specie «*ciò che è subordinato al genere*». Genere e specie sono termini relativi, un genere posto su di un nodo alto dell'albero definisce la specie sottostante, la quale diventa a sua volta genere della specie sottostante, e così via.

Schema *per* classificazione e schema *di* classificazione

(Ranganathan, 1965) opera una distinzione fra *schema 'per' classificazione* e *schema 'di' classificazione*[2]. Il primo si riferisce alla sistemazione di concetti in una struttura ordinata e riguarda i filosofi e gli scienziati che studiano come organizzare logicamente l'universo della conoscenza; il secondo si riferisce all'organizzazione delle informazioni in un catalogo o in una bibliografia, nonché all'organizzazione dei libri sugli scaffali di una biblioteca e delle risorse bibliografiche in alcuni siti web, e riguarda i bibliotecari che organizzano la conoscenza trasmessa dalle risorse bibliografiche stesse. Fra la classificazione in ambito filosofico o scientifico e la classificazione in ambito biblioteconomico vi è una *netta separazione* e, al contempo, una *relazione diretta*, in quanto la maggior parte degli schemi di classificazione ideati da bibliotecari si ispira a sistemi elaborati da filosofi (Aristotele, Platone, Tommaso d'Aquino, Bacone, Kant, He-

[2] SHIYALI RAMAMRITA RANGANATHAN, *A descriptive account of Colon classification*, Bangalore, Sarada Ranganathan Endowment for Library Science, 1965, p. A3.

gel, ecc.) e da scienziati (Linneo, Darwin, ecc.): per esempio, la Classificazione Decimale Dewey (DDC) e la Classificazione Decimale Universale (CDU) si riferiscono più o meno direttamente al sistema di Bacone (e di Hegel) e la Classificazione espansiva di Cutter a quello di Darwin.

Gli schemi di classificazione possono considerarsi grandi *mappe del sapere*; non sono sistemazioni teoriche dello scibile, seppure molte di esse si riferiscano a principi filosofici o finalizzati al lavoro catalografico e bibliografico: loro scopo è sistemare logicamente le informazioni, evitando l'ordine casuale dei sistemi d'indicizzazione alfabetica; sono costruiti sulla base della *literary warrant*, della garanzia letteraria o, meglio, della garanzia bibliografica, ovvero della produzione editoriale su un argomento: se la produzione su un soggetto aumenta, lo spazio sulla mappa per quella voce si allarga oppure ottiene una collocazione più idonea; se diminuisce, la voce si comprime o addirittura scompare. Infatti la *«classificazione non si occupa del mondo, ma dei documenti che ne trattano (principio della garanzia bibliografica [literary warrant]); perciò non è una classificazione del sapere, ma del sapere come s'incarna nei documenti»* (Crocetti, Fagiolini, 2001)[3]. L'operazione che descrive un'opera nei termini del suo contenuto concettuale individua e definisce l'argomento trattato principalmente, il *tema di base*,

> *quell'oggetto unitario di conoscenza al quale sono riferibili i singoli temi particolari discussi nel documento e al quale sono correlate nel testo tutte le informazioni fornite intenzionalmente dall'autore, essendo stata proprio la volontà di comunicare nozioni dirette e specifiche su quell'argomento di conoscenza il*

[3] LUIGI CROCETTI, ALBAROSA FAGIOLINI, *Classificazione Decimale Dewey*, ed. aggiornata a DDC 21, Roma, Associazione italiana biblioteche, 2001, pp. 10-11.

> *motivo fondamentale della produzione intellettuale dell'intero documento*,

come afferma la *Guida all'indicizzazione per soggetto* redatta dal GRIS, Gruppo di Ricerca sull'Indicizzazione per Soggetto dell'Associazione Italiana Biblioteche (AIB, 1996)[4]. Vi sono alcune eccezioni come, per esempio, la poesia, la letteratura drammatica, la narrativa che non sono convenzionalmente organizzate per soggetto bensì per forma letteraria.

La tecnica classificatoria

La classificazione è un'attività concreta, è una tecnica catalografica. Per ottenere un buon risultato è necessario che il catalogatore la contestualizzi per:

1. tipo di biblioteca (p.e., generale, specializzata);
2. tipo di risorsa bibliografica;
3. tipo di destinatario a cui la biblioteca si rivolge.

Bohdan S. Wynar precisa la tecnica classificatoria utilizzata per le risorse:

1. classificare l'opera prima secondo il soggetto, poi tramite la forma in cui il soggetto è rappresentato, a eccezione delle classi generali e della letteratura;
2. classificare un'opera dove sarà utilizzata in prevalenza;
3. porre l'opera nella divisione di soggetto più specifica che la contiene, piuttosto che con l'argomento generale (principio di specificità). Se le biblioteche di una certa grandezza assegnassero una sola notazione alle opere sulla storia della Francia, senza dividerle per periodi di tempo e per

[4] ASSOCIAZIONE ITALIANA BIBLIOTECHE (AIB), *Guida all'indicizzazione per soggetto*, Roma, 1996, p. 13.

luoghi, il risultato sarebbe l'affollamento di molti volumi sotto un solo numero;
4. quando l'opera concerne due o tre soggetti, collocarla con il soggetto predominante o con quello trattato per primo. Quando l'opera concerne più di tre aspetti collocarla nella classe generale che li comprende tutti (Taylor, 2004)[5].

Schemi di classificazione

Gli schemi di classificazione gerarchici adottano una sequenza che procede da concetti generali a concetti particolari:
1. voci preliminari al soggetto;
2. soggetto;
3. voci che sviluppano il soggetto.

Il *Glossario* della Classificazione Decimale Dewey definisce le *notazioni «numeri, lettere e/o altri segni adoperati per rappresentare le divisioni principali e subordinate di uno schema di classificazione»*. Commenta (Chan, 1981): *«Tradizionalmente gli schemi di classificazione bibliotecaria tendono a enumerare tutti i soggetti e le loro suddivisioni e a fornire per essi simboli già costruiti»*[6]. Questo tipo di classificazione è detta *enumerativa*; l'esempio più importante è costituito dalla classificazione della Library of Congress.

Le notazioni possono essere costituite da insiemi numerici, da insiemi alfabetici, oppure da insiemi alfanumerici, come le *notazioni miste* usate dalla Library *of Congress Classification* (LCC) e dalla *Bibliographic Classification* (BC). Ogni notazione esprime la sequenza delle classi. Nel caso della DDC, il prolunga-

[5] ARLENE G. TAYLOR, *Wynar's Introduction to cataloging and classification*, rev. 9 ed., Westport Conn., London, Libraries Unlimited, 2004, p. 25.
[6] CHAN, L. M., *op. cit.*, pp. 210-211.

mento della frazione decimale specifica il soggetto; la flessibilità dello schema permette l'integrazione periodica di nuove voci dovute alla comparsa di opere che trattano argomenti nuovi.

Ciascuna caratteristica è chiamata *faccetta*; cuore, polmoni, fegato, ecc. definiscono e nominano la faccetta *parti del corpo umano*. I *fenomeni* di questa faccetta – cuore, polmoni, fegato – hanno in comune la caratteristica di essere tutte parti del corpo umano. Una faccetta di un'area concettuale consiste di un certo numero di fenomeni, all'interno di una disciplina, che condividono alcune caratteristiche. Gli elementi coordinati su ciascun livello o fase di divisione formano una *serie* (in inglese *array*), un *ordinamento* di classi coordinate basato sul principio di divisione: per esempio *Letteratura americana, Letteratura inglese, Letteratura tedesca*, etc. La classe non presenta sempre un ordine precostituito o naturale delle faccette. La letteratura può essere divisa per la lingua e poi per il periodo, oppure per gli aspetti formali (poesia, dramma, ecc.) e poi per la lingua e il periodo o, ancora, per il periodo, per gli aspetti formali e infine per la lingua; il diritto può essere diviso per branche e, quindi, per paese (p.e., diritto penale italiano, diritto penale francese, diritto civile italiano, ecc.), o viceversa (diritto italiano penale, diritto italiano civile, diritto francese penale, ecc.). Per essere coerente, ciascun sistema di classificazione determina l'ordine di *disposizione* delle faccette, chiamato *ordine di citazione*. La notazione (il simbolo di classificazione) viene *sintetizzata*; in altre parole, la notazione viene costruita tramite la combinazione di più elementi, derivati dalle Tavole e dalle Tavole ausiliarie della classificazione (tavole che presentano aspetti ricorrenti validi per tutte o alcune classi elencate nelle Tavole principali), per formare numeri complessi idonei a mostrare pienamente tutte le caratteristiche o *faccette* di un soggetto complesso, seguendo l'ordine di citazione prescritto dalle Tavole mediante note. *Sintesi multipla* indica l'aggiunta di due o più faccette, l'una dietro l'altra, allo stesso numero base. La teoria della classificazione tende a valorizzare l'*a-*

nalisi per faccette e la *sintesi*, cioè l'analisi e la frammentazione di un soggetto nelle sue parti componenti e il riassemblaggio di queste parti come richiesto dall'opera che viene rappresentata. Invece di enumerare tutti i soggetti in una struttura gerarchica, la teoria moderna suggerisce che uno schema di classificazione potrebbe identificare i componenti base dei soggetti, elencando sotto ciascuna disciplina o classe principale i concetti base o elementi *isolati* secondo certe caratteristiche o *faccette*. In aggiunta, divisioni ricorrenti come quelle formali, geografiche o cronologiche sono elencate separatamente per l'applicazione a tutte le classi (tavole ausiliarie o complementari).

L'atto del classificare

L'atto del classificare consiste essenzialmente nella *sintesi*, cioè nell'identificazione o nel raggruppamento degli elementi componenti che costituiscono il soggetto dell'opera da classificare. I componenti sono raggruppati secondo un ordine predeterminato, detto *formula di citazione*, prescritto per quella classe specifica. Un sistema basato su questi principi è detto *classificazione a faccette* o *analitico-sintetica*, un esempio della quale è costituito dalla Classificazione Colon (CC) di Ranganathan.

Un sistema di classificazione è sempre pre-coordinato (risponde a una struttura delineata) a differenza di un sistema di soggettazione verbale, inteso come insieme di voci indice o soggetti da attribuire alle risorse bibliografiche che può essere pre- o post-coordinato (nel secondo caso ha una struttura meno rigida, costruita sulla base della letteratura indicizzata).

Le tavole di una classificazione possono coprire l'intero universo bibliografico (della documentazione prodotta dall'uomo) – si parla di classificazione *generale*; o riguardare la documentazione prodotta su un settore particolare della conoscenza (una di-

sciplina o una sua parte) – e in tal caso si parla di classificazione *settoriale* o *specialistica*. Si chiamano *gerarchiche* le classificazioni che procedono dal generale al particolare, *non gerarchiche* le classificazioni caratterizzate da schemi che hanno la facoltà di crescere su se stessi in modo da:

1. mantenere sostanzialmente invariato l'impianto classificatorio;
2. disporre di sviluppi e di articolazioni capaci di ospitare e di specificare adeguatamente raccolte documentarie via via più estese, come la *Classificazione espansiva* di Cutter.

Sistemi di classificazione bibliografica

Numerosi sono i sistemi di classificazione bibliografica, elaborati nel tempo, per l'ordinamento dei libri sugli scaffali di una biblioteca e delle risorse elettroniche su siti web e per l'organizzazione logica delle registrazioni bibliografiche di un catalogo di biblioteca. I sistemi moderni nascono nella seconda metà dell'Ottocento e si sviluppano nella prima metà del Novecento e hanno finalità prevalentemente o esclusivamente catalografiche. I principali sistemi di classificazione sono: la *Expansive Classification* (EC), elaborata da Charles Ammi Cutter (1837-1903); la *Dewey Decimal Classification* (DDC), ideata da Melvil Dewey (1851-1931) e pubblicata per la prima volta nel 1876; la *Classification Decimale Universelle* (CDU), derivata dalla DDC e pubblicata nel 1905 da Paul Otlet e Henri La Fontaine, aggiornata periodicamente e diffusa soprattutto in ambito scientifico; la *Library of Congress Classification* (LCC), elaborata all'interno di quella grande biblioteca e utilizzata da altri istituti per l'ampiezza delle sue voci; la *Bibliographic Classification* (BC) alla quale Henry Evelyn Bliss (1870-1955) comincia a lavorare all'inizio del Novecento e che è stata pubblicata nel 1935 col titolo *A System of bibliographic classification* (un primo schema era appar-

so nel 1910 su *Library journal* dell'ALA[7]), seguita da una ristampa l'anno successivo e pubblicata in veste definitiva in quattro volumi fra il 1940 e il 1953 e da una successiva seconda edizione nel 1986; la *Colon Classification* (CC) di S.R. Ranganathan, la cui prima edizione è apparsa nel 1933, classificazione ritenuta fra le più interessanti per il suo impianto teorico. Altri sistemi sono stati usati localmente, come la *Bibliotecno-Bibliograficeskaja Klassifikacija* (BBK) della Biblioteca Lenin di Mosca, adottata dal 1959 e abbandonata alla fine del secolo XX.

La classificazione nel web

Il web è costituito di un numero incommensurato di risorse documentarie, che sarebbero del tutto inutili se non fosse possibile accedervi in modo selettivo. La prima forma di organizzazione del sapere del web è stata rappresentata dai motori di ricerca di prima generazione, che consentivano l'interrogazione delle parole contenute nelle risorse testuali. Ciò, dal punto di vista della classificazione, è un sistema primitivo, perché l'indice delle parole restituito dal motore di ricerca consentiva di creare una classe di risorse documentarie per ogni parola indicizzata e di presentare di conseguenza l'insieme delle risorse che contenevano le parole scelte come stringa d'interrogazione. L'insieme delle risorse documentarie che venivano elencate in risposta a una interrogazione costituiva una classe di documenti costruita a posteriori, cioè su richiesta. Con questo tipo di ricerca, come in quello degli attuali motori di ricerca, non esisteva e non esiste alcun controllo terminologico, sicché l'interrogazione deve essere ripetuta per tutte le forme, le varianti e i sinonimi con i quali può esprimersi l'oggetto della ricerca.

7 *American Library Association.*

I primi strumenti a utilizzare le tassonomie, sistemi di classificazione in forma verbale (non utilizzano notazioni ma espressioni verbali), sono state le *web directory* (elenchi di siti); le tassonomie, tuttavia, sono alla base di un numero significativo di repertori e motori di ricerca. *«I grandi repertori di Internet per argomenti (directory, subject gateway, virtual reference desk ecc.), come Yahoo! e Open directory, utilizzano principalmente questo modello di organizzazione»* (Gnoli et alii, 2006)[8].

Le web directory servono per ottenere rapidamente notizie, selezionate da esperti dell'informazione, che scelgono le risorse sulla base dell'affidabilità e dell'autorevolezza delle fonti. Si tratta di indici nei quali vengono descritte le migliori risorse su una tematica, la cui consultazione può costituire di per sé informazione importante e, insieme, una base di partenza per una ricerca approfondita. Le web directory, per il notevole impiego di tempo che il lavoro di scelta richiede, non sono aggiornate *in tempo reale*; esse garantiscono tuttavia un certo livello di qualità. Per questo motivo, per disporre di una bibliografia aggiornata, è opportuno interrogare altre fonti di rete. Navigare in una web directory è molto semplice: si presenta come un elenco ramificato, con una struttura ad albero; per esempio, all'interno della voce generale *Scienze sociali* si trova, in ordine alfabetico, *Archeologia*, *Filosofia*, *Geografia*, *Pedagogia*, *Psicologia*, *Scienze Umane*; all'interno di *Filosofia* troveremo i siti di *Estetica*, *Logica*, *Metafisica*, ecc. La struttura generale del sito non è sempre intuitiva, perciò alcune web directory sono dotate di un motore di ricerca interno, che consente di individuare le risorse rilevanti, tramite una ricerca per parola chiave, tra i siti selezionati. Le web directory prevedono, inoltre, la possibilità che un termine si tro-

8 CLAUDIO GNOLI, VITTORIO MARINO, LUCA ROSATI, *Organizzare la conoscenza. Dalle biblioteche all'architettura dell'informazione per il web*, Milano, Hops Tecniche nuove, 2006, p. 44.

vi in più punti della struttura; i termini di rinvio si possono distinguere perché terminano con il carattere @. Per esempio, nella directory di Yahoo, *Law* (Giurisprudenza) si trova all'interno della classe *Government* (Funzione pubblica) e ha un rinvio *Law@* nella classe *Social Sciences* (Scienze sociali).

Tra le web directory generaliste più importanti si segnalano *l'Open Directory Project* (DMOZ)[9] che, a cura di volontari, recupera risorse di diverse discipline, e *Yahoo! Directory*[10], una web directory *storica*, ovvero presente da molto tempo sul web.

Le classificazioni sono sistematicamente utilizzate nel mondo del web in numerosi database e motori di ricerca, nonché nell'organizzazione dei siti web. Per esempio, il *BUBL*, *BUlletin Board for Libraries* è una directory che presenta le informazioni organizzate secondo la Classificazione Decimale Dewey. Il sito *Bielefeld Academic Search Engine* (BASE), invece, è un sito generalista che fornisce accesso libero a oltre 38 milioni di risorse; integra la ricerca mediante un motore, con la possibilità di navigare tra le risorse secondo lo schema della Classificazione Dewey.

Le web directory generaliste sono sempre più in disuso (BUBL non è più aggiornato da aprile 2011), mentre quelle disciplinari sono ampliamente utilizzate. Tra queste ultime possiamo considerare i *virtual reference desk* (VRD), i quali sono in genere, specializzati per disciplina – come la *Guida dell'utente nel labirinto dell'informazione*, predisposta dall'Università La Sapienza di Roma per le diverse facoltà[11] – o per ambiti d'interesse specifici – come la *Rassegna degli strumenti informatici per lo studio dell'antichità classica* dell'Università di Bologna[12]. Un

9 <http://www.dmoz.org/>.
10 <http://dir.yahoo.com/>.
11 <http://www.uniroma1.it/vrd>.
12 <http://www.rassegna.unibo.it>.

esempio rappresentativo di VRD generale è il WWW Virtual Library[13], il primo importante *Virtual Reference Desk* generale, nato nel 1991.

Il limite dei *virtual reference desk* è la mancanza di personalizzazione, che costituisce invece l'elemento cardine del servizio di reference in presenza; l'interazione umana rappresenta il valore sostanziale del reference, poiché nessuna macchina, nessun automatismo e nessuna lista preconfezionata può sostituire, almeno finora, il processo d'interazione tra le due reti neurali del bibliotecario e del ricercatore.

La classificazione dei siti web secondo uno schema appositamente creato, o in base a una classificazione bibliografica nota, richiede un'enorme quantità di risorse umane per essere svolta con efficacia e per essere costantemente aggiornata. I problemi sono maggiori nel caso di siti generalisti, perché le competenze nei vari settori disciplinari devono essere suddivise tra vari collaboratori. D'altra parte:

> *le fonti d'informazioni diffuse ed eterogenee costituiscono un corpus di oggetti enorme, mutevole, indefinito da classificare senza un'autorità centrale, rivolti a un gruppo di utenti eterogeneo e in continuo aumento. Questa situazione richiede strategie di classificazioni nuove e diverse. Il web di oggi si adatta perfettamente a questo quadro. Sul web, la tendenza è la scalabilità, la flessibilità, la fluidità e la semplicità per soddisfare le crescenti richieste dei milioni di persone con un background culturale e sociale diverso e che operano in tutte le parti del mondo. In queste circostanze, gli schemi di classificazione tradizionali e precisi diventano costosi (da creare e da mantenere) e probabilmente perdono la capacità di corrispondere al modo in cui l'utente pensa e organizza il mondo. Le folksonomie for-*

[13] <http://www.vlib.org>.

niscono un approccio che risolve i problemi delle classificazioni adatte al web (Quintarelli, 2005)[14].

Il termine *folksonomia* deriva dall'unione delle parole *folk* e tassonomia, per mettere in evidenza come la tassonomia, l'uso di stringhe verbali per la classificazione delle risorse, venga applicata direttamente dagli utenti del web. Gli utenti sono moltissimi e in poco tempo ciascuno può svolgere un'enorme quantità di lavoro, anche con poco sforzo.

Alcuni siti, come Flickr[15], che serve per la condivisione di fotografie, ospitano progetti di catalogazione sociale, ovvero sollecitano la collaborazione dei visitatori per l'aggiunta di parole chiave o etichette (tag) e per la descrizione delle fotografie di importanti enti, come la *Smithsonian Institution*, la *Library of Congress* (entrambe di Washington, D.C.) e la *New York Public Library*.

La classificazione sociale

La classificazione sociale non è un sistema che interessa solo le immagini; può essere estesa a molti altri oggetti del web: per esempio, alla condivisione dei bookmark (segnalibri), per la quale esistono servizi di *social book marking* che consentono di etichettare, gestire e condividere indirizzi web – per esempio Delicious[16] e Connotea[17] – o alla condivisione di citazioni bibliografiche (e relativi link) – per esempio Citeulike[18] o Mendeley[19].

---

14 Cfr. EMANUELE QUINTARELLI, *Folksonomies: power to the people*, intervento all'incontro ISKO Italia-UniMIB, 24 giugno 2005, in *ISKO Italia. Documenti*. <http://www.iskoi.org/doc/folksonomies.htm>.

15 <http://www.flickr.com>.

16 <http://www.delicious.com>.

17 <http://www.connotea.org>.

18 <http://www.citeulike.org/>.

19 <http://www.mendeley.com/>.

La classificazione sociale mediante tagging collaborativo (*folksonomy* o folksonomia) può esser considerata come un'alternativa ai modelli di classificazione tradizionali predisposti in ambito bibliotecario e come uno strumento complementare da utilizzare accanto ad essi. È tuttavia molto importante distinguere tra l'autenticità e l'autorevolezza delle immagini o delle registrazioni bibliografiche, che provengono da prestigiose istituzioni, e la qualità dei metadati relativi agli oggetti digitali, che possono essere aggiunti da chiunque, e quindi possono non sempre essere pertinenti o di valore universale, ovvero condivisi.

(Howarth, 2011) ha evidenziato come l'assegnazione di un ruolo così importante agli utenti possa essere considerato come la naturale evoluzione del concetto di centralità dell'utente che ha sempre caratterizzato gli obiettivi del catalogo, a partire da Antonio Panizzi:

> *La catalogazione sociale – un'attività che incoraggia gli utenti a taggare, valutare, recensire e commentare le raccolte delle biblioteche, e che e che comprende anche il tagging degli utenti direttamente sulle registrazioni del catalogo per facilitare l'individuazione delle risorse – è considerata generalmente come una tendenza recente; […] al contrario […] il coinvolgimento degli utenti nella descrizione e nell'aggiunta di termini del linguaggio naturale rientra nel processo logico di sviluppo dei codici e degli standard di catalogazione*[20].

[20] LYNNE C. HOWARTH, *Da «un magnifico errore» a «una comunità d'interazione dinamica». I codici di catalogazione angloamericani e l'evoluzione della catalogazione sociale, Lectio Magistralis in Biblioteconomia, Firenze, Università degli studi di Firenze, 23 marzo 2011*, Fiesole, Firenze, Casalini Libri, 2011, p. 29.

Bibliografia

ASSOCIAZIONE ITALIANA BIBLIOTECHE (AIB), *Guida all'indicizzazione per soggetto*, Roma, 1996

CHAN, L.M., *Cataloging and classification: an introduction*, New York, McGraw-Hill, 1981

CROCETTI, L., FAGIOLINI, A., *Classificazione Decimale Dewey*, ed. agg. a DDC 21, Roma, Associazione italiana biblioteche, 2001

GNOLI, C., MARINO, V., ROSATI, L., *Organizzare la conoscenza. Dalle biblioteche all'architettura dell'informazione per il web*, Milano, Hops Tecniche nuove, 2006

HOWARTH, L.C., *Da «un magnifico errore» a «una comunità d'interazione dinamica». I codici di catalogazione angloamericani e l'evoluzione della catalogazione sociale, Lectio Magistralis in Biblioteconomia, Firenze, Università degli studi di Firenze, 23 marzo 2011*, Fiesole, Firenze, Casalini Libri, 2011

QUINTARELLI, E., *Folksonomies: power to the people*, intervento all'incontro ISKO Italia-UniMIB, 24 Giugno 2005, in *ISKO Italia. Documenti* <http://www.iskoi.org/doc/folksonomies.htm>

RANGANATHAN, S.R., *A descriptive account of Colon classification*, Bangalore, Sarada Ranganathan Endowment for Library Science, 1965

TAYLOR, A.G., *Wynar's Introduction to cataloging and classification*, rev. ed. 9, Westport Conn., London, Libraries Unlimited, 2004

Sitografia

<http://dir.yahoo.com/>
<http://www.citeulike.org/>
<http://www.connotea.org>
<http://www.delicious.com>
<http://www.dmoz.org/>
<http://www.flickr.com>
<http://www.mendeley.com/>
<http://www.rassegna.unibo.it>
<http://www.uniroma1.it/vrd>
<http://www.vlib.org>

Tassonomie e thesauri

ANTONIETTA FOLINO*

1. Introduzione

La gestione delle informazioni e della conoscenza relative ad uno o più domini e gli scambi comunicativi tra gli attori che operano al loro interno non possono prescindere da un utilizzo condiviso della terminologia[1], rendendo necessaria la definizione di strumenti e risorse che ne consentano un'organizzazione il più possibile coerente e non ambigua.

Si tratta di lessici, glossari, tassonomie, soggettari, sistemi di classificazione, thesauri, mappe concettuali, ontologie, ecc., che, dal punto di vista strutturale, si differenziano gli uni dagli altri sostanzialmente per la presenza o meno di relazioni di tipo semantico tra i concetti in essi rappresentati e per il diverso grado di formalismo che caratterizza la modellizzazione dell'informazione, come mostrato in Figura 1.

* Università della Calabria, Dipartimento di Lingue e Scienze dell'Educazione.

1 *«The word terminology refers to at least three different concepts: a. The principles and conceptual bases that govern the study of terms; b. The guidelines used in terminographic work; c. The set of terms of a particular special subject»*.
MARIA TERESA CABRÉ, *Terminology: theory, methods and applications*, Sager J.C. (ed.), DeCesaris J.A. (traduzione di), Philadelphia PA, John Benjamins, 1998, p. 33.

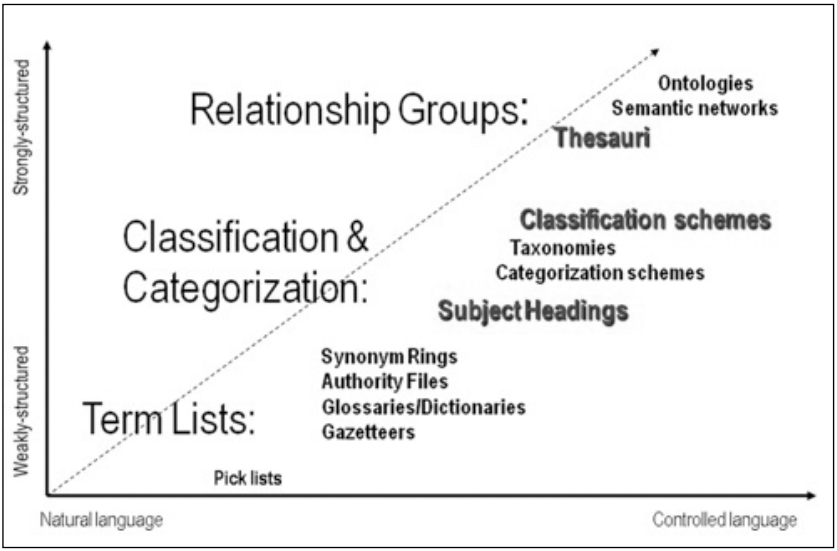


Figura 1. Controllo terminologico e strutturazione[2].

Si va infatti dai lessici specialistici, che consistono in liste non strutturate di termini afferenti ad un settore della conoscenza[3], ai glossari, che aggiungono a ciascuna voce la relativa definizione[4],

2 MARCIA LEI ZENG, ATHENA SALABA, *Toward an International Sharing and Use of Subject Authority Data*, FRBR Workshop, OCLC, 2005.

3 Un lessico specialistico può essere definito come un insieme di termini utilizzati in modo consensuale e convenzionale dagli individui che operano in uno stesso ambito al fine di scambiare e divulgare informazioni e conoscenze in maniera precisa, univoca e concisa.
Cfr. HELLMUT RIEDIGER, *Cos'è la terminologia e come si fa un glossario*, 2012.
<http://www.term-minator.it/corso/doc/mod3_termino_glossa.pdf>.

4 Se costruiti in forma di database terminologici, all'interno dei quali ciascun termine è descritto da un'apposita scheda, i glossari prevedono anche l'inserimento di relazioni tra concetti (sinonimia, iperonimia-iponimia, ecc.) sulla base di un albero concettuale che ne definisce la struttura.

ai sistemi di classificazione[5], ai soggettari[6], ai thesauri, che, pur se con modalità diverse, integrano la terminologia con relazioni semantiche tra i concetti, fino alle mappe concettuali e alle ontologie, che aggiungono un alto livello di formalismo attraverso la definizione di restrizioni nella partecipazione dei concetti alle relazioni, la rappresentazione in linguaggi interpretabili dalle macchine, l'esplicitazione della natura delle relazioni, ecc.

Il principio sottostante a tali risorse, seppur ottenibile a diversi livelli e con differenti modalità, è il controllo terminologico: a ciascun termine è attribuibile il solo significato valido per il dominio di interesse e l'interpretazione non deve essere soggetta ad ambiguità e incomprensione. Tali significati, insieme all'uso di ciascuna voce, sono condivisi e compresi dalla comunità di utenti che li impiega nelle situazioni comunicative in cui è coinvolta. Nella misura in cui rappresentano la conoscenza di dominio, invece, gran parte di queste risorse rientra in quelli che vengono comunemente definiti *Knowledge Organization Systems* (KOS)[7].

---

5 Rispetto ai thesauri, i sistemi di classificazione hanno spesso la pretesa di rappresentare l'intero scibile, mentre dal punto di vista strutturale, non presentano la relazione associativa e prevedono un sistema di notazione obbligatorio.

6 Rispetto ad un thesaurus un soggettario fornisce le regole sintattiche per la costruzione della stringa di soggetto, ovvero di «*una sequenza ordinata di termini, che rappresenta il soggetto di un documento*» (Cfr. ALBERTO CHETI, *Manuale ipertestuale di analisi concettuale*, 1996, <http://biblioteche.unibo.it/manuals/html_1/HOME.HTML>), da assegnare ai documenti, soprattutto in una logica di pre-coordinazione, nella quale la combinazione dei concetti rappresentativi del contenuto di un documento avviene già al momento dell'indicizzazione. Per tali motivi, un soggettario viene espressamente costruito per esigenze di indicizzazione, per cui anche la sua struttura viene definita in relazione all'insieme di documenti da indicizzare.

7 *«The term knowledge organization systems is intended to encompass all types of schemes for organizing information and promoting knowledge*

Ci soffermeremo essenzialmente sulla descrizione delle caratteristiche e delle funzionalità di tassonomie e thesauri, in ragione della loro riscoperta e rivalutazione nell'ambito della ricerca sul Web, dopo una fase di declino che li ha considerati quasi obsoleti di fronte alle emergenti tecnologie del Web Semantico.

Se, infatti, in quanto strumenti cardine delle scienze documentali e bibliotecarie, assolvevano a funzioni di ordinamento, organizzazione e recupero dell'informazione in ambienti prettamente cartacei, oggi la loro definizione risulta indispensabile nel mondo del web e dei documenti digitali, con la conseguente acquisizione di nuove connotazioni e funzionalità e con una rivalutazione e valorizzazione delle loro potenzialità.

Ci occuperemo in un primo momento delle tassonomie fornendo una presentazione delle caratteristiche che le contraddistinguono per poi passare alla trattazione più approfondita dei thesauri, presentandone il contesto normativo, le funzionalità, le modalità di realizzazione e gli impieghi. Sebbene infatti le norme più recenti in materia di vocabolari controllati abbiano ampliato i propri interessi non limitandosi ai soli thesauri, le indicazioni più precise e complete continuano ad essere fornite solo per questi ultimi, per cui anche in questo caso, l'attenzione sarà focalizzata maggiormente sui thesauri facendo riferimento alle tassonomie laddove si ritiene interessante effettuare dei confronti.

---

management. […]. Knowledge organization systems also include highly structured vocabularies, such as thesauri, and less traditional schemes, such as semantic networks and ontologies».

GAIL HODGE, *Systems of Knowledge Organization for Digital libraries. Beyond traditional authority files*, 2000.

<http://www.clir.org/pubs/reports/pub91/contents.html>.

2. Tassonomie

In senso tradizionale, l'idea di tassonomia è legata alle discipline scientifiche e alla classificazione dicotomica degli organismi nelle scienze biologiche[8]. Nell'ambito della documentazione, invece, il termine acquista un'accezione più ampia, riferendosi all'organizzazione sistematica di un soggetto o dominio. Secondo la definizione fornita dallo standard americano ANSI/NISO Z39-19:2005[9], infatti, per tassonomia si intende: *«a controlled vocabulary consisting of preferred terms, all of which are connected in a hierarchy or polyhierarchy»*. Le tassonomie sono impiegate in modo massiccio nell'organizzazione dell'informazione in ambienti digitali. Spesso, infatti, vengono citati repertori quali Yahoo! o Open directory come casi esemplificativi del loro impiego (Gnoli et alii, 2006)[10]. Tuttavia, proprio a causa di questa ampia diffusione, spesso si assiste ad un utilizzo indiscriminato del termine tassonomia, indicando con esso strutture di concetti che non rispettano le caratteristiche distintive di tale strumento e il rigore con il quale dovrebbe essere costruito per garantire il controllo terminologico e l'organizzazione della conoscenza. In taluni casi, ad esempio, per tassonomia si intende l'insieme delle categorie (o faccette) che costituiscono il menù di navigazione di un sito web, mentre il significato tradizionalmente accettato nelle scienze del libro è quello di sistema di classificazione[11].

---

8 Si pensi ai botanici Konrad Gesner e Karl von Linné.

9 ANSI/NISO Z39-19:2005, *Guidelines for the construction, format, and management of monolingual controlled vocabularies*, p. 18.

10 Claudio Gnoli, Vittorio Marino, Luca Rosati, *Organizzare la conoscenza: dalle biblioteche all'architettura dell'informazione per il web*, Milano, Tecniche Nuove, 2006, p. 44.

11 Cfr. Gail Rayburn, *Taxonomies and Thesauri*, 2011. <http://www.llrx.com/system/files?file=taxonomiesthesauri.pdf>.

La sola relazione semantica che si inserisce tra i concetti di una tassonomia è, quindi, quella gerarchica, che ne determina la tipica organizzazione ad albero che rende visibili i rapporti tra sovra- e sotto-ordinati. Si predilige la monogerarchia, quindi una collocazione unica per ciascun concetto. Sebbene i termini tassonomia e thesaurus siano spesso utilizzati indifferentemente come se fossero sinonimi, tra le due tipologie di vocabolari controllati esistono differenze significative: rispetto ad un thesaurus, infatti, una tassonomia non prevede la relazione di equivalenza, essendo costituita, come da definizione sopra riportata, da soli termini preferiti (non vi si ritrovano dunque sinonimi, quasi sinonimi, ecc. che costituirebbero il vocabolario d'accesso[12]), né la relazione associativa per l'esplicitazione di rapporti semantici diversi da quelli gerarchici. Inoltre, le relazioni non sono esplicitate per mezzo di sigle standard, come avviene invece per i thesauri. Per quanto riguarda la struttura, è possibile prevedere, così come per i thesauri, l'inserimento di faccette (vedi par. 3.4) che consentano di separare in maniera sistematica le gerarchie soprattutto ai fini della navigazione sul Web. Non prevedono, inoltre, un sistema di notazione, spesso presente nei thesauri, e i criteri che guidano la definizione delle relazioni sono meno rigorosi rispetto a quelli che le regolano in questi ultimi. Le principali differenze sono riassunte nella Figura 2, estratta dalla citata ANSI/NISO Z39-19:2005.

Per quanto riguarda le funzionalità alle quali assolvono, le tassonomie condividono sostanzialmente quelle proprie dei thesauri, quindi indicizzazione, recupero di informazione, organizzazione della conoscenza, descritte dettagliatamente nel seguito

12 *«Costituito sia dai termini preferiti che dai termini non preferiti, cioè dai termini che non possono essere utilizzati per l'indicizzazione e che rimandano a termini preferiti»*.
SERAFINA SPINELLI, *Introduzione all'indicizzazione*, 2006.
<http://biocfarm.unibo.it/~spinelli/indicizzazione/>.

del capitolo. Tuttavia, in ragione delle differenze pocanzi illustrate, le potenzialità di un thesaurus nella maggior parte dei contesti sono indubbiamente maggiori potendo sfruttare relazioni altre rispetto a quella gerarchica[13]. La navigazione in ambienti digitali resta la funzione principale che le tassonomie adempiono.

Property	List	Synonym Ring	Taxonomy	Thesaurus
Types of Terms				
Preferred terms	Yes	No	Yes	Yes
Entry terms	No	Yes	No	Yes
Candidate terms	No	No	No	Optional
Provisional terms	No	No	No	Optional
Deleted terms	No	No	No	Optional
Relationships	No	Yes	Yes	Yes
Equivalence		Yes	No	Yes
Hierarchy		No	Yes	Yes
Part/Whole		No	Yes	Yes
IsA		No	Yes	Yes
HasA		No	Yes	Yes
Classification		No	Optional	Optional
Related terms		No	No	Yes
Facet		No	No	Optional
Notes	No	No	Optional	Optional
Scope note			No	Optional
History note			No	Optional
Other notes			No	Optional

Figura 2. Tassonomie e Thesauri[14].

[13] Basti pensare alla funzione di mediazione tra indicizzatore e utente che il thesaurus svolge grazie alla relazione di equivalenza o all'estensione dei risultati delle ricerche per mezzo di quella associative.

[14] ANSI/NISO, *op. cit.*, p. 135.

3. Thesauri

3.1 *Definizione e struttura di un thesaurus*

La definizione di thesaurus fornita dalla recente norma ISO 25964-1:2011 che ne regola i principi di costruzione e gestione è *«controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms»*[15]. Per vocabolario controllato si intende un insieme di termini il cui significato è chiaro e non ambiguo nello specifico contesto o dominio per il quale il thesaurus è costruito. Tale controllo si esercita principalmente attraverso la strutturazione dei concetti sulla base di relazioni semantiche rese esplicite. Nel caso specifico, si tratta di tre grandi tipologie standardizzate di relazioni: di equivalenza, gerarchiche e associative.

La relazione di equivalenza consente la gestione della sinonimia e della quasi-sinonimia e delle varianti linguistiche. I termini che ai fini del thesaurus rappresentano il medesimo concetto fanno parte del cosiddetto gruppo di equivalenza[16]. Ad uno di questi termini viene attribuito lo status di preferito, mentre i restanti, in qualità di non preferiti, gli sono legati per mezzo di rinvii. Tale relazione viene esplicitata per mezzo delle sigle USE, che rimanda dal/dai termini non preferiti a quello preferito, e UF (*Used for*), che viene definita nel senso opposto. Nell'ambito della stessa relazione di equivalenza è possibile distinguere tra sinonimia assoluta e sinonimia relativa: nel primo caso essa esiste indipendentemente dall'area semantica, dalla specificità del

[15] ISO 25964-1:2011, Information and documentation – *Thesauri and interoperability with other vocabularies*, Part 1: *Thesauri for information retrieval*, p. 12.

[16] SERAFINA SPINELLI, *Introduzione ai thesauri*, 2005. <http://biocfarm.unibo.it/~spinelli/indicizzazione/thesauri.htm>.

thesaurus e dalla scelta del termine preferito e interessa varianti ortografiche[17], sigle e acronimi e relative forme sciolte[18], preferenze linguistiche[19], termini specialistici e termini utilizzati nel linguaggio comune[20], nomi comuni e nomi commerciali[21], prestiti e relative traduzioni[22], ecc.; in presenza di quasi sinonimi, invece, si parla si sinonimia relativa, in quanto la relazione è valida solo ai fini e nel contesto per il quale il thesaurus viene realizzato. La possibilità di inserire tale relazione dipende dalla copertura semantica del thesaurus e dal grado di discriminazione necessario in fase di indicizzazione e di ricerca. La si ritrova, ad esempio, in thesauri non specialistici o per tematiche che si discostano dal focus principale del vocabolario controllato[23]. Rientrano in questa tipologie di sinonimia anche gli antonimi[24], che, sempre in base alle esigenze di specificità del thesaurus, possono essere ritenuti equivalenti sebbene esprimano significato opposto, poiché il tema di un documento potrebbe essere espresso per mezzo di entrambe le forme linguistiche, e termini per i quali la relazione di sinonimia è valida solo in alcuni contesti d'u-

17 Es. Database – Data-base – Data Base.

18 Per quanto concerne la scelta del termine preferito, la normativa prevede che l'acronimo venga privilegiato qualora sia più utilizzato e conosciuto della relativa forma sciolta (es. UNICEF). Se, invece, il suo utilizzo causerebbe ambiguità, è preferibile stabilire un rinvio verso il suo scioglimento (es. CC – Corrente Continua, Conto Corrente, ecc.).

19 Es. Flusso Termico - Flusso di calore.

20 Es. Cefalea – Mal di testa.

21 Es. Penna – Bic.

22 Es. Computer – Elaboratore elettronico.

23 Una relazione di quasi sinonimia potrebbe essere definita tra Legge, Decreto Legge, Decreto Legislativo in un thesaurus non relativo a discipline giuridiche. Nel caso di un thesaurus specialistico in questo dominio, ovviamente, tra i tre concetti non potrebbe sussistere alcuna relazione di equivalenza.

24 Es. Tolleranza – Intolleranza.

so[25]. Ulteriori casi in cui è possibile ricorrere alla relazione di equivalenza consistono nel rinvio da termini più specifici ad un termine sovraordinato[26] e nella *compound equivalence*[27], nella quale, in presenza di un concetto complesso, è preferibile utilizzare i concetti semplici che risultano dalla sua scomposizione. Anche in questi casi, la scelta dipende dal contenuto dei documenti da indicizzare e, di conseguenza, dai termini potenzialmente impiegabili il loro recupero.

L'attribuzione dello status di termine preferito dipende da un insieme di fattori, tra i quali principalmente le finalità del thesaurus, soprattutto nel caso della quasi sinonimia, e gli utenti ai quali il thesaurus è rivolto e che lo utilizzeranno per finalità di recupero dell'informazione o come strumento di organizzazione della conoscenza[28]. Così come per la scelta dei termini che devono essere inseriti in un thesaurus (si veda il par. 3.5), anche in questo caso, l'autorevolezza della fonte dalla quale un termine è stato estratto, il parere di esperti di dominio e la frequenza d'uso rappresentano un supporto nella scelta del termine preferito. La relazione di equivalenza si rivela di fondamentale importanza per indirizzare nel corretto utilizzo della terminologia nei domini di realizzazione dei thesauri, limitando l'ambiguità e l'incoerenza negli scambi comunicativi tra gli attori che vi operano.

La relazione gerarchica esprime un rapporto di subordinazione/sovraordinazione tra concetti che rappresentano una classe o

---

25 Es. Faccia – Viso.

26 Es. Rocce magmatiche, Rocce sedimentarie, Rocce metamorfiche USE Rocce.

27 Es. Glicemia USE Sangue + Glucosio.

28 Un thesaurus costruito nel dominio della medicina e rivolto ai pazienti privilegerà in qualità di termini preferiti quelli quotidianamente impiegati nel linguaggio comune. La relazione di equivalenza con i corrispondenti termini specialistici, ad esempio nel contesto di un sistema di recupero di informazione, permetterà l'incontro con gli specialisti del settore.

un insieme e concetti che rappresentano elementi, parti o individui[29]. Si distingue, infatti, tra relazione gerarchica di tipo genere-specie, esplicitata tramite le sigle BTG (*Broader Term Generic*) – NTG (*Narrower Term Generic*)[30], di tipo parte-tutto, le cui sigle sono BTP (*Broader Term Partitive*) e NTP (*Narrower Term Partitive*)[31] ed esemplificativa, BTI (*Broader Term Instantial*) e NTI (*Narrower Term Instantial*)[32]. La relazione genere-specie può essere definita solo tra concetti che appartengono alla medesima categoria (oggetti, materiali, attività, proprietà, discipline, ecc.) e che rispettano l'*all-and-some test*[33], ovvero: solo alcuni membri della classe che indica il genere rientrano in quella che indica la specie, ma tutti i membri della classe che indica la specie devono rientrare in quella che indica il genere. I concetti che possono essere interessati dalla relazione parte-tutto rientrano in categorie ben identificate, quali sistemi e organi del corpo, luoghi geografici, discipline e campi di studio, strutture sociali gerarchizzate[34]. I concetti che indicano le parti, infatti, devono appartenere in maniera esclusiva al concetto che rappresenta il tutto e con il quale esiste una relazione di questo tipo. Nei casi in cui questa condizione non si verifica è preferibile ricorrere ad una relazione di tipo associativo.

La relazione associativa, la cui sigla è RT (*Related Term*), consente la gestione delle relazioni semantiche diverse da quella gerarchica che possono essere stabilite tra due concetti. La ISO 25964-1:2011 prevede la possibilità che la natura di tale re-

29 La relazione gerarchica può essere definita su più livelli, generalmente contrassegnati da un numero progressivo, mentre nei casi in cui un concetto abbia più di un sovraordinato, si parla di poligerarchia.
30 Es. Felini NTG Gatti – Gatti BTG Felini.
31 Es. Apparato circolatorio NTP Cuore – Cuore BTP Apparato circolatorio.
32 Es. Catene Montuose NTI Alpi – Alpi BTI Catene Montuose.
33 ISO 25964-1:2011, *cit.*, p. 59.
34 *Ivi*, p. 60.

lazione venga di volta in volta esplicitata (es. causa-effetto), agevolando la comprensione della struttura thesaurale e accorciando le distanze tra questo strumento ed un'ontologia. Tuttavia non si tratta ancora di relazioni standardizzate e le eventuali operazioni di mappatura tra thesauri esistenti potrebbero risultare compromesse dalla difficoltà di stabilire delle corrispondenze. Tale tipo di relazione interessa concetti che condividono lo stesso sovraordinato o, nel caso di un thesaurus a faccette, concetti appartenenti a raggruppamenti diversi[35] e legati tra loro da un qualsiasi legame semantico e quelli introdotti dal medesimo principio di suddivisione e collocati sullo stesso livello gerarchico.

L'esercizio del controllo terminologico avviene anche tramite l'inserimento di note d'ambito o *Scope Note* (SN), ovvero campi testuali nei quali è possibile delimitare il significato di un dato termine, fornire informazioni circa il suo impiego, eventuali usi particolari, ecc.[36], e di qualificatori, che in presenza di omografi permettono di specificare l'ambito al quale il significato di ciascuno si riferisce[37].

3.2 *Evoluzione concettuale e normativa*

La maggior parte dei thesauri esistenti in letteratura è stata costruita secondo la norma ISO 2788:1986[38], se monolingue, e se-

[35] Es. Attività - prodotto (Tessitura RT Tessuto); Agente – Attività (Docente – Insegnamento); Disciplina – Oggetto di studio (Anatomia – Corpo Umano); Oggetti – Proprietà (Metalli – Malleabilità); Attività – Strumento (Incisione – Bulino); ecc.

[36] Esse possono essere ulteriormente specificate ricorrendo alle sigle DEF e HN (*History Note*) qualora si voglia fornire il significato o informazioni sull'evoluzione temporale di un dato concetto.

[37] Es. Organo (strumento musicale); Organo (corpo umano).

[38] ISO 2788:1986, Documentation – *Guidelines for the establishment and development of monolingual thesauri.*

condo la norma ISO 5964:1985[39], se multilingue, che hanno rappresentato un riferimento internazionale per lungo tempo, ovvero fino alla pubblicazione della recente ISO 25964-1:2011, che le ha sostituite[40] e la cui emanazione era particolarmente attesa dalla comunità dei professionisti dell'informazione, dal momento che le norme esistenti non rispecchiavano l'evoluzione dei thesauri ed era necessario un adattamento alle nuove esigenze di gestione dell'informazione in ambienti prettamente digitali. In particolare ci si riferisce all'aumento considerevole e costante della quantità di informazioni e di documenti disponibili sul Web, alle conseguenti accresciute opportunità di recupero degli stessi, alla diversa natura delle risorse informative e all'avvento dei motori di ricerca e del metodo di ricerca full-text.

Proprio i limiti di tale metodo rendono evidente la necessità dei thesauri nei repository di documenti digitali: la presenza dei termini che compongono un'interrogazione (o query) nel testo di un documento non garantisce che gli stessi siano rappresentativi del suo contenuto concettuale. L'attribuzione di voci indice a partire da un thesaurus, invece, aumenta la probabilità che il risultato di una ricerca sia pertinente e risponda ai bisogni informativi degli utenti. Altri limiti dipendono dal fatto che la ricerca può avvenire a partire da più di un termine o essere formulata in una lingua diversa da quella in cui sono redatti i documenti.

In Figura 3 viene riportato un quadro riassuntivo delle norme che hanno interessato i thesauri, e più in generale i vocabolari controllati, negli ultimi decenni.

---

39 ISO 5964:1985, Documentation – *Guidelines for the establishment and development of multilingual thesauri.*

40 La sostituzione ha interessato anche le norme nazionali francesi definite dall'ente di normazione AFNOR (*Association française de Normalisation*), ovvero la NF Z 47-100-1981- *Règles d'établissement des thésaurus monolingues* e la NF Z 47-101-1990 - *Principes directeurs pour l'établissement des thésaurus multilingues.*

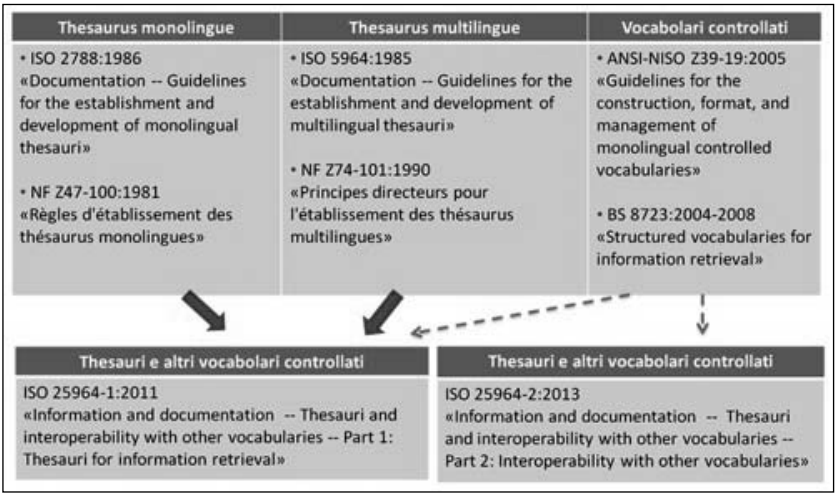


Figura 3. Thesaurus: contesto normativo

La direzione verso la quale si orienta la ISO 25964:2011, ovvero l'interoperabilità tra i vocabolari controllati, si coniuga con l'operato e gli obiettivi del *World Wide Web Consortium* (W3C)[41], concretizzatisi con la raccomandazione SKOS (*Simple Knowledge Organization Systems*)[42], finalizzata all'utilizzo dei sistemi di organizzazione della conoscenza nel Web Semantico, prevedendo sia la trasposizione di quelli esistenti sia la definizione di nuovi secondo questo formato. Il linguaggio si basa su RDF[43] e viene considerato come una tecnologia intermedia tra l'elevato formalismo logico dei linguaggi OWL[44] per la costru-

41 <http://www.w3.org/>.
42 <http://www.w3.org/2004/02/skos/>.
43 *Resource Description Framework*. <http://www.w3.org/RDF/>.
44 *Ontology Web Language*. <http://www.w3.org/TR/owl-features/>.

zione di ontologie e l'assenza o lo scarsa strutturazione delle applicazioni che attualmente caratterizzano il Web. Tra gli elementi o tag per mezzo dei quali è possibile descrivere i sistemi di organizzazione della conoscenza rientrano: *Concept, prefLabel, altLabel, broader, narrower, related, definition, scopeNote, exactMatch, collection, member, broaderTransitive, ecc.* alcune delle quali direttamente riconducibili alla struttura di un thesaurus.

Come evidenziato in figura, l'elaborazione della norma attualmente in vigore è avvenuta anche sulla base delle norme redatte in contesto inglese ed americano: la BS 8723:2004-2008[45] e la citata ANSI/NISO Z39-19:2005. Seppur non con valenza internazionale, infatti, questi documenti normativi avevano già introdotto ed anticipato alcune delle principali novità contenute nella ISO 25964[46], tra le quali l'interesse verso altre tipologie di vocabolari controllati, il formato elettronico dei thesauri, il loro

---

45 BS 8723:2004-2008, *Structured vocabularies for information retrievial – Guide.*
È costituito da cinque sezioni edite separatamente.

46 L'iter di sviluppo di una norma internazionale, infatti, viene in molti casi avviato in risposta ad esigenze e richieste provenienti da stakeholder esterni all'ente di normazione (nel caso specifico l'*International Organization for Standardization* – ISO). Tali esigenze vengono comunicate agli enti nazionali (es. UNI per l'Italia, BSI per il Regno Unito, ANSI per gli Stati Uniti, AFNOR per la Francia, ecc.), che, a loro volta, contattano l'ISO, del quale sono membri. L'iter di sviluppo degli standard è così riassunto sul sito dell'ISO (<http://www.iso.org/iso/home.html>): *«1. New standard is proposed to relevant technical committee. If proposal is accepted 2. Working group of experts start discussion to prepare a working draft. 3. 1st working draft shared with technical committee and with ISO CS. If consensus is reached within the TC 4. Draft shared with all ISO national members, who are asked to comment. If consensus is reached 5. Final draft sent to all ISO members. If standard is approved by member vote 6. ISO International Standard»*.
Cfr. anche ISABELLA FLORIO, *La normativa standardizzata per la gestione delle documentazione tra Italia e Francia*, Rubbettino Editore, 2011.

impiego per scopi di Information Retrieval (IR), l'interoperabilità tra più vocabolari attraverso la definizione di modelli[47] e formati (Calvitti, Viti, 2009)[48] (Casson, 2006)[49] (Groupe Langages documentaires de l'ADBS, 2007)[50], (Dextre Clarke, Lei Zeng, 2012)[51].

Al fine di comprendere che cosa si intende per thesaurus, quali sono le sue caratteristiche e come queste siano mutate nel tempo è opportuno riprendere le definizioni presenti nella normativa tecnica. Si riportano, quindi, sia quella prevista dalla ISO 2788:1986, nonostante la stessa non sia più in vigore, sia quella fornita dalla ISO 25964-1:2011, peraltro già riportata nel precedente paragrafo, a garanzia di completezza nella descrizione di tale strumento e a testimonianza dell'evoluzione della sua natura: *«Il thesaurus è un vocabolario di un linguaggio di indicizzazione controllato, organizzato formalmente in maniera da rende-*

47 Il modello Zthes (<http://zthes.z3950.org/>), ad esempio, è basato sul linguaggio XML - eXtensible Markup *Language* (<http://www.w3.org/XML/>) e si propone di facilitare l'interoperabilità tra applicazioni che utilizzano thesauri conformi a quanto previsto dalle norme ISO 2788 e ANSI/NISO Z39-19.

48 Cfr. TIZIANA CALVITTI, ELISABETTA VITI, *Da ISO 2788 ai nuovi standard per la costruzione e l'interoperabilità dei vocabolari controllati: un'analisi comparativa*, in «Bollettino AIB», vol. 49, n. 3, settembre 2009, pp. 307-322.

49 Cfr. EMANUELA CASSON, *Dai thesauri ai vocabolari controllati: alcune novità introdotte nell'ultima edizione dello standard ANSI/NISO Z39.19-2005*, in «AIDAinformazioni», a. 24, n. 1-2, gennaio-giugno 2006, pp. 69-77.

50 Cfr. GROUPE LANGAGES DOCUMENTAIRES DE L'ADBS, *Les normes de conception, gestion et maintenance de thésaurus: évolution récentes et perspectives*, in «Documentaliste-Sciences de l'Information», vol. 44, n. 1, 2007, pp. 66-74.

51 Cfr. STELLA G. DEXTRE CLARKE, MARCIA LEI ZENG, *From ISO 2788 to ISO 25964: The Evolution of Thesaurus Standards towards Interoperability and Data Modeling*, in «Information Standards Quarterly», vol. 24, n. 1, 2012, pp. 20-26.

re esplicite le relazioni 'a priori' fra i concetti»[52]; *«a controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms»*.

Dal confronto con la definizione della ISO 2788:1986 emerge che:

- Il thesaurus è un linguaggio controllato e strutturato, ma il suo utilizzo non si limita all'indicizzazione;
- Si distingue tra concetti e termini: questa distinzione restringe il divario tra thesauri e ontologie ed è indispensabile per l'uso in sistemi informatici. Il concetto è un'unità di pensiero indipendente dai termini impiegati per identificarlo, considerati come etichette. Prima della ISO 25964-1:2011 si parlava indistintamente di relazioni tra termini e di relazioni tra concetti. La norma chiarisce invece come solo per la relazione di equivalenza si possa parlare di relazione tra termini (essendo definita tra un termine preferito selezionato per rappresentare il concetto e i termini non preferiti considerati come ulteriori possibili etichette), mentre le due restanti relazioni thesaurali interessano i concetti;
- Si introduce la distinzione tra termini preferiti e sinonimi e quasi sinonimi.

In quest'ultimo punto consiste l'obiettivo primario del thesaurus in quanto strumento di recupero dell'informazione: ren-

[52] ISO 2788:1986, *op. cit*., p. 3.
Le relazioni paradigmatiche o *a priori* tra i concetti sono così definite dalla ISO 25964-1:2011: *«Relationship between concepts which is inherent in the concepts themselves»*. Si tratta cioè di relazioni che, contrariamente a quelle sintagmatiche, sono sempre valide indipendentemente dai contesti specifici di indicizzazione o di definizione del thesaurus. Le relazioni sintagmatiche sono perciò sconsigliate all'interno dei thesauri.

dere possibile l'incontro tra indicizzatore e utente e far sì che entrambi utilizzino lo stesso termine preferito per individuare un dato concetto. L'utente, cioè, attraverso i rinvii costituiti dai sinonimi e dai quasi sinonimi viene ricondotto verso il termine preferito scelto dall'indicizzatore recuperando l'informazione pur utilizzando una chiave di ricerca diversa da quella preferita[53]. In tal modo, inoltre, la ricerca può avvenire anche a partire dai termini appartenenti al vocabolario d'accesso.

Il thesaurus, in tal senso, funge da mediatore tra professionisti dell'informazione e utenti finali.

Una differenza fondamentale nell'impiego del thesaurus per la gestione dei documenti e delle informazioni in ambienti digitali riguarda gli utenti e il loro ruolo: da strumenti elaborati ed utilizzati esclusivamente dai professionisti dell'informazione a supporto delle pratiche di indicizzazione[54] e di ricerca in ambienti cartacei, diventano strumenti resi disponibili all'utenza generica impegnata in attività di ricerca, consultabili nella loro strutturazione, navigabili per permettere il recupero dei documenti e interpretabili dalle macchine. L'utente è quindi reso partecipe dell'organizzazione data alla conoscenza di un dato dominio e assume un ruolo attivo nell'utilizzo delle risorse termino-

---

53 In tal senso nella sezione *Terms and definitions* della norma si introducono anche le voci *Entry term* e *Search term*.

54 *«The act of describing or identifying a document in terms of its subject content»*.
UNI ISO 5963:1985, Documentazione - *Metodi per l'analisi dei documenti, la determinazione del loro soggetto e la selezione dei termini di indicizzazione*, 1985, p. 2.
«L'operazione mediante la quale si creano gli accessi al contenuto semantico del documento. Consta delle fasi di analisi concettuale e di traduzione dei concetti individuati e delle relazioni logiche individuate nei termini e nelle forme proprie del linguaggio di indicizzazione prescelto».
CHETI, A., *op. cit.*

logiche. Si riportano due estratti delle rispettive norme che testimoniano di questo mutamento:

> *La sua applicazione è limitata alle agenzie che utilizzano persone, quali indicizzatori, per analizzare i documenti ed esprimere i soggetti […]. Non è applicabile alle agenzie che utilizzano tecniche di indicizzazione completamente automatiche […]*[55].

> *Whereas in the past thesauri were designed for information professionals trained in indexing and searching, today there is a demand for vocabularies that untrained users will find to be intuitive, and for vocabularies that enable inferencing by machines*[56].

Pur non stravolgendo il concetto stesso di thesaurus e le relazioni che ne sono alla base, la nuova norma introduce ulteriori novità legate al ruolo prioritario di strumento di IR e si focalizza su aspetti precedentemente non contemplati per le motivazioni già illustrate. In particolare si registra la presenza di:

- Raccomandazioni sulla scelta dei software per la costruzione di thesauri, al fine di individuare le caratteristiche che gli stessi dovrebbero possedere per una corretta gestione di tali strumenti;
- *Modelli di dati* che definiscono in maniera astratta la struttura e la semantica di un thesaurus e che possono essere utilizzati per costruire strutture relazionali per database e formati di scambio, questi ultimi spesso basati su linguaggi di marcatura quale XML;
- Indicazioni relative all'integrazione dei thesauri in sistemi di indicizzazione e di ricerca dell'informazione (es. database bibliografici, centri di documentazione, banche dati docu-

[55] ISO 2788:1986, *op. cit.*, p. 3.
[56] ISO 25964-1:2011, *op. cit.*, p. VI.

mentali, basi di conoscenza, CMS[57], motori di ricerca, ecc.);
– Indicazioni più dettagliate sull'analisi a faccette per la costruzione di thesauri: come illustrato nel paragrafo 4.2, infatti, tale approccio si adatta meglio alle caratteristiche degli ambienti digitali;
– Nuove modalità di visualizzazione e di presentazione dei thesauri, date le potenzialità e la flessibilità garantite da ambienti web.

Quanto finora detto a proposito della norma ISO 25964 riguarda la prima delle due parti delle quali si compone: la seconda[58], pubblicata in data 4 marzo 2013, si sofferma sulla questione dell'interoperabilità tra vocabolari controllati (schemi di classificazione, ontologie, tassonomie, soggettari, terminologie, ecc.) e sulle operazioni di mappatura tra vocabolari diversi[59]. In alcune situazioni, quali ad esempio la ricerca in raccolte indicizzate con risorse diverse o l'utilizzo integrato di vocabolari controllati, è necessario stabilire una corrispondenza tra le differenti strutture concettuali. Pur nel pieno rispetto dei criteri di costruzione presenti nelle norme e pur interessando lo stesso dominio di conoscenza, infatti, tra due o più vocabolari possono esistere differenze sia tecniche dovute ai formati e ai sistemi informatici utilizzati, sia contenutistiche dovute all'utilizzo di terminologia settoriale o ad una diversa definizione delle relazioni, per cui in un thesaurus due termini possono essere considerati sino-

57 *Content Management Systems*.
58 ISO 25964-2: 2013, Information and Documentation – *Thesauri and interoperability with other vocabularies*, Part 2: *Interoperability with other vocabularies.*
59 O anche tra le diverse versioni linguistiche in thesauri multilingue. Finora tali operazioni sono state regolate dalle seguenti Linee Guida: Cfr. IFLA, WORKING GROUP ON GUIDELINES FOR MULTILINGUAL THESAURI, *Guidelines for multilingual thesauri*, IFLA, 2005.

nimi, mentre in un altro rappresentano due concetti distinti. Le operazioni di mappatura contribuiscono quindi ad un recupero più efficiente dell'informazione, poiché l'equivalenza tra voci appartenenti a più vocabolari permette di ritrovare tutte le risorse informative indicizzate tramite ciascuna di esse.

3.3 *Funzionalità di tassonomie e thesauri*

Come accennato, pur nella consapevolezza delle differenze tra le due tipologie di vocabolari controllati, tassonomie e thesauri assolvono pressoché alle medesime funzionalità.

I cambiamenti verificatisi negli ultimi decenni ai quali si è accennato hanno richiesto un'evoluzione del concetto di thesaurus e di tassonomia, ma al tempo stesso ne hanno valorizzato le funzionalità attraverso l'allargamento dei contesti d'uso e la dimostrazione delle loro potenzialità anche in ambienti digitali.

Riprendendo quando affermato a proposito delle figure interessate dall'utilizzo del thesaurus in relazione alle funzionalità, si può distinguere tra controllo terminologico, indicizzazione, supporto nella definizione dei metadati e classificazione da una parte, in quanto attività che continuano ad essere di competenza del professionista dell'informazione[60], e navigazione, ricerca ed espansione dei risultati delle ricerche dall'altra, poiché coinvolgono direttamente l'utente e/o la macchina.

---

[60] I più recenti sviluppi del Web stanno determinando anche per la pratica dell'indicizzazione una ridefinizione dei ruoli: si parla infatti di *social indexing* e di *folksonomy*, intese come forme di organizzazione della conoscenza nelle quali gli utenti modellizzano (attribuiscono parole chiave o classificano) sulla base del loro punto di vista e della loro visione di un dato dominio di conoscenza.
Cfr. OLIVIER ERTZSCHEID, GABRIEL GALLEZOT, *Etude exploratoire des pratiques d'indexation sociale comme une renégociation des espaces documentaires. Vers un nouveau big bang documentaire?*, in Document numérique et société, Charton G., Broudoux E. (a cura di), ADBS Éditions,

Controllo terminologico

Attraverso il controllo terminologico è possibile gestire l'ambiguità del linguaggio naturale e limitare il significato di un dato concetto al contesto di applicazione del thesaurus. In particolare tale funzione viene esercitata per disambiguare i concetti interessati dai fenomeni della polisemia e della sinonimia, attraverso la scelta del termine preferito, la strutturazione dei concetti per mezzo delle relazioni, che fornendo il contesto semantico di ciascuno di essi contribuiscono ad esplicitarne il significato, l'inserimento di note d'ambito, che forniscono definizioni o indicazioni riguardo all'impiego dei termini, e di qualificatori, che specificano l'ambito o la disciplina alla quale i concetti appartengono (soprattutto nei thesauri multidisciplinari).

Indicizzazione e supporto nella metadatazione

Come più volte menzionato, l'indicizzazione, intesa come l'azione di descrivere o identificare un documento nei termini del suo contenuto concettuale, è stata riconosciuta per lungo tempo come la funzionalità principale di un thesaurus, tant'è che lo stesso era definito *vocabolario di un linguaggio di indicizzazione controllato*. Seppur in contesti diversi, l'indicizzazione rimane una funzione fondamentale e i thesauri rappresentano fonti autorevoli dalle quali estrarre i concetti da attribuire ad una risorsa informativa di qualsivoglia natura al fine di permetterne la descrizione e il recupero.

2006; ZACKLAD, M., *Classification, thésaurus, ontologies, folksonomies : comparaison du point de vue de la recherche ouverte d'information (ROI)*, in CAIS/ACSI 2007, 35e Congrès annuel de l'Association Canadienne des Sciences de l'Information. Partage de l'information dans un monde fragmenté: Franchir les frontières, Montréal, 10-12 maggio 2007, Arsenault C., Dalkir, K. (a cura di), 2007; ÉLIE FRANCIS, ODILE QUESNEL, *Indéxation collaborative et folksonomies*, in «Documentaliste – Sciences de l'Information», vol. 44, n. 1, 2007, pp. 58-63.

Tassonomie e thesauri forniscono un supporto all'indicizzatore anche in virtù della loro struttura: l'organizzazione dei concetti gli consente non solo di identificare i termini, ma anche di determinare il livello di specificità con il quale si vuole rappresentare il contenuto concettuale dei documenti e che più si adatta alle caratteristiche della collezione e alle esigenze dei potenziali utenti, guidandolo verso concetti più generici, più specifici, o, solo per i thesauri, semanticamente correlati.

In ambiente digitale l'indicizzazione rientra nell'operazione di metadatazione delle risorse: gli standard di descrizione documentale, quale *Dublin Core* (DC)[61], prevedono, all'interno del set di metadati, un apposito elemento per l'inserimento di parole chiave rappresentative del contenuto delle risorse (*subject* nel DC). Anche in questo caso, come espressamente raccomandato dagli stessi standard, i vocabolari controllati possono essere utilizzati come fonte per la compilazione di metadati semantici[62] o indicizzazione semantica.

L'utilizzo di un linguaggio controllato a fini di indicizzazione, sebbene più dispendioso rispetto all'uso del linguaggio naturale o dei termini presenti nei titoli o nel testo dei documenti, contribuisce a limitare la soggettività e l'incoerenza che inevitabilmente caratterizza il lavoro di indicizzazione, soprattutto se eseguito da diversi indicizzatori. Presenta, inoltre, indubbi vantaggi in fase di ricerca data la scarsa coincidenza tra i termini utilizzati da questi ultimi e quelli utilizzati dagli utenti. Ne deriva anche un'applicazione più rigorosa del controllo terminologico favorendo in ogni situazione l'impiego dello stesso termine per rappresentare uno stesso concetto. Tale scelta è alla base dell'in-

61 <http://dublincore.org/>.

62 Relativi cioè al contenuto del documento. I metadati forniscono anche informazioni gestionali e relative alla proprietà intellettuale delle risorse informative.

dicizzazione detta *assegnata* o *per concetti*, che si contrappone a quella *derivata* o *per termini*[63].

Rappresentare il contenuto di un documento e, di conseguenza, recuperarlo in fase di ricerca, può comportare l'uso di più termini, la cui combinazione può avvenire in maniera pre- o post-coordinata[64]. Nel primo caso la modalità di combinazione è prevista e definita a priori e in fase di indicizzazione sulla base di regole di citazione che stabiliscono la sequenza secondo la quale i termini devono comparire in un'intestazione o stringa di soggetto[65] e tale rigidità potrebbe in molti casi compromettere il buon esito delle operazioni di ricerca. I contesti d'uso più frequenti sono, quindi, l'indicizzazione per soggetto e la collocazione di materiale librario. Nel secondo caso, invece, i termini vengono combinati solo al momento della ricerca e, per tale ragione, la post-coordinazione è la scelta più comune in ambiente digitale, vista la semplicità di effettuare delle ricerche utilizzando uno o più termini come chiave di accesso all'informazione. La metodologia di classificazione a faccette predilige tale approccio, che permette, attraverso la fase di sintesi, di non inserire termini eccessivamente lunghi e di non enumerarne quanti più possibile come nei sistemi tradizionali[66]. In una logica di recupero dell'informazione, sebbene un documento possa essere ritrovato anche a partire da uno solo dei descrittori attribuitigli, la

63 Cfr. SPINELLI, S., *op. cit.*

64 Cfr. CLAUDIO GNOLI, *Coordinazione, ordine di citazione e livelli integrativi in ambiente digitale*, in «Bibliotime», a. 6, n. 1, marzo 2003. <http://www.spbo.unibo.it/bibliotime/num-vi-1/gnoli.htm>.

65 Come nel caso del Nuovo Soggettario della Biblioteca Nazionale Centrale di Firenze, che prevede, oltre all'ordine di citazione, anche delle regole sintattiche che derivano da un'analisi dei ruoli svolti dai concetti contenuti nelle faccette e nelle categorie.

66 DOUGLAS TUDHOPE, CERI BINDING, *Faceted Thesauri*, in «Axiomathes», vol. 18, n. 2, giugno 2008, pp. 217-218.

post-coordinazione riduce il grado di precisione della ricerca, poiché il documento potrebbe rispondere solo parzialmente alle esigenze dell'utente[67].

Thesauri e tassonomie forniscono un supporto anche nelle tecniche di indicizzazione e di classificazione automatica[68]. Esemplificativo a tal proposito è il sistema AgroTagger[69], un estrattore di termini che attribuisce le voci indice ai nuovi documenti sulla base dei descrittori contenuti all'interno del thesaurus AGROVOC[70].

Rispetto ai tradizionali software di estrazione terminologica, basati prevalentemente su misure statistiche che calcolano la frequenza delle occorrenze dei termini all'interno dei testi, senza distinguere quelli che effettivamente sono rappresentativi del dominio di interesse da quelli relativi al linguaggio comune, gli applicativi basati su thesauri identificano all'interno dei testi stessi solo le voci che con molta probabilità ne rappresentano il contenuto concettuale. Mentre quindi nel primo caso l'indicizzazione richiede una fase di validazione e di selezione manuale, nel secondo avviene un'operazione di indicizzazione vera e propria.

Recupero dell'informazione

Sempre più frequentemente tassonomie e thesauri sono integrati in CMS, centri di documentazione, banche dati, ecc. come

---

67 Richiamo e precisione sono misure che in IR consentono di valutare l'esito delle ricerche di informazione. Il richiamo esprime il rapporto tra documenti rilevanti trovati e il totale dei documenti rilevanti esistenti, mentre la precisione indica il rapporto tra documenti rilevanti trovati e il totale dei documenti trovati. L'utilizzo di strumenti come i thesauri contribuisce ad accrescere il valore di queste misure.

68 Cfr. Olena Medelyan, Ian H. Witten, *Thesaurus Based Authomatic Keyphase Indexing*, in Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, Chapel Hill, NC, USA, ACM, 2006, pp. 296-297.

69 <http://aims.fao.org/agrotagger>.

70 <http://aims.fao.org/standards/agrovoc/about>.

strumenti di recupero dell'informazione e dei documenti.

In alcune applicazioni la struttura è resa disponibile all'utente, il quale può navigare al suo interno attraverso le relazioni precedentemente definite, al fine di recuperare le risorse informative associate a ciascuna voce durante la fase di indicizzazione (elemento *subject*). Ciascun documento può essere assegnato ad una o più entrate del thesaurus o della tassonomia, favorendo una classificazione multipla, e perciò flessibile, e garantendo più punti di accesso per il suo recupero. La navigazione del thesaurus, inoltre, permette all'utente di farsi un'idea sul contenuto dei documenti della collezione, sull'utilizzo della terminologia (in particolare in presenza di sinonimi e varianti) e sull'organizzazione concettuale del relativo dominio di conoscenza, soprattutto se non esperto. L'accesso al'informazione per mezzo del *browsing* può avvenire sia a partire dalla presentazione alfabetica delle entrate lessicali, sia da quella sistematica, ovvero dalla struttura classificatoria basata sull'organizzazione in categorie rappresentative del dominio per il quale il thesaurus o la tassonomia sono realizzati; in genere la prima modalità viene preferita di fronte ad un bisogno informativo meglio definito.

L'integrazione di una struttura thesaurale in un sistema di gestione del contenuto o in un software di ricerca e la possibilità di recuperare i documenti o le informazioni direttamente collegate ai concetti presuppone una scelta a monte in ordine alla definizione dell'architettura dell'informazione: inserire tutte le voci di un thesaurus esistente, anche se non esistono contenuti indicizzabili per mezzo di alcuni dei suoi descrittori o inserire (e integrare di volta in volta) solo le voci alle quali possono essere associate delle risorse informative. Nel primo caso il rischio è quello di effettuare delle ricerche che non producono alcun risultato, anche se disporre del thesaurus nella sua interezza permetterebbe di meglio gestire l'ampliamento della collezione documentale e di comprendere l'organizzazione concettuale del dominio; nel secondo il rischio risiede appunto nella parzialità della rap-

presentazione dei concetti, ma le ricerche produrrebbero in tutti i casi un set di risultati.

In aggiunta alla navigazione, la struttura del thesaurus può fornire un supporto all'utente nella scelta dei termini da impiegare come chiavi di ricerca e quindi nella formulazione della query.

Nei sistemi di Information Retrieval che integrano tassonomie e thesauri per migliorare i risultati ottenuti a seguito di un'interrogazione, sfruttando le relazioni semantiche definite tra termini e concetti utilizzati per indicizzare i documenti, le parole che compongono la query vengono confrontate con i termini del vocabolario controllato. Al termine di una ricerca il sistema può proporre all'utente delle possibilità di raffinamento dei risultati ottenuti attraverso la visualizzazione delle relazioni, oppure può automaticamente inserire nei risultati i documenti indicizzati con termini correlati a quelli impiegati per l'interrogazione.

Nello specifico in un thesaurus:

- La relazione di equivalenza fa sì che un documento venga recuperato anche se l'utente, per effettuare l'interrogazione, non ha utilizzato il termine preferito attribuito alla risorsa dall'indicizzatore. In tal senso i termini del gruppo di equivalenza rivestono un'importanza notevole, in quanto punti di accesso all'informazione che anticipano le possibili modalità di ricerca da parte degli utenti finali e per tali motivi è importante inserire tutti quelli potenzialmente utili a favorire l'incontro tra il professionista dell'informazione e l'utente;
- La relazione NT consente di specificare meglio la ricerca e ridurre gli item che potrebbero essere recuperati. Si parla in questo caso di *query extension*;
- Le relazioni BT e RT consentono di ampliare la ricerca nel caso vengano restituiti pochi documenti. Si parla in questo caso di *query expansion*.

La Figura 4 mostra l'utilizzo del LISA (*Library and Information Science Abstracts*) Thesaurus per il raffinamento della ricerca. L'utente può migliorare i risultati ottenuti attraverso i concetti sovra e sotto ordinati e i concetti correlati.

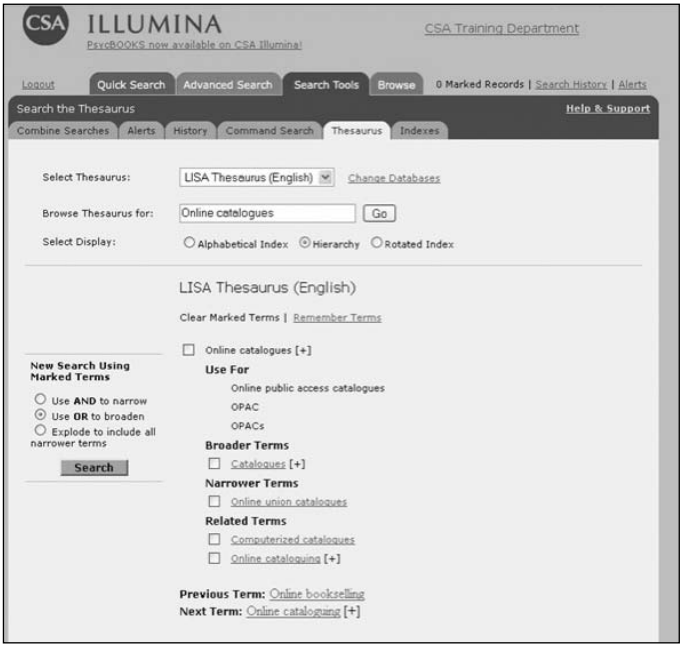


Figura 4. Thesaurus e IR[71].

Organizzazione della conoscenza di dominio

Thesauri e tassonomie possono essere progettati e realizzati anche per la rappresentazione di un dominio di conoscenza o di una disciplina: in questo senso i concetti in esse contenuti sono rappresentativi di un settore della realtà nella sua interezza piuttosto che di una collezione di documenti da indicizzare e da re-

71 <http://www.csa.com/factsheets/supplements/LISAguide.pdf>.

cuperare. Le relazioni semantiche e, laddove presente, l'organizzazione dei concetti per mezzo di categorie rappresentative, forniscono anche una classificazione del dominio di interesse, che può diventare un punto di riferimento per la sistematizzazione della conoscenza, la predisposizione di documenti, lo scambio non ambiguo di informazione, la normalizzazione e la guida nell'utilizzo della terminologia.

In questo caso, quindi, i termini da inserire all'interno del thesaurus non devono essere scelti sulla base della collezione da indicizzare (inserendo quindi solo i termini potenzialmente utilizzabili per l'indicizzazione e per il successivo recupero delle risorse informative) ma devono costituire un set il più possibile rappresentativo del dominio da rappresentare.

3.4 *Tipologie di thesauri*

È possibile distinguere tra diverse tipologie di thesauri sulla base della copertura semantica, delle lingue di compilazione e della struttura classificatoria sottostante. Rispetto al primo criterio i thesauri possono essere generali[72] (o multidisciplinari o ad ampio spettro) o speciali[73].

---

[72] Di questa categoria fanno parte ad esempio: l'UNESCO Thesaurus (<http://databases.unesco.org/thesaurus/>) che interessa le seguenti discipline: istruzione, cultura, scienze naturali, scienze umane e sociali, comunicazione e informazione, l'AAT - *Art and Architecture Thesaurus* (<http://www.getty.edu/research/tools/vocabularies/aat/index.html>) che organizza concetti relative ad oggetti, artisti, luoghi legati all'arte, all'architettura e più in generale alla cultura; l'AGROVOC thesaurus (<http://aims.fao.org/standards/agrovoc/about>) che si occupa di alimentazione, agricoltura, pesca, ambiente e altri domini correlati; ecc.

[73] Relativi a domini specifici. Ne sono esempi il MeSh – *Medical Subject Headings* (<http://www.nlm.nih.gov/mesh/>), il NASA Thesaurus (<http://data.nasa.gov/nasa-thesaurus/>), l'*Alzheimer's Disease Thesaurus* (<http://adlib.alzheimers.org/adear/alzdb/thesaurus.aspx>), lo *European Education Thesaurus* (<http://www.freethesaurus.info/redined/en/index.php>), ecc.

Anche se multidisciplinari o generali, i thesauri interessano sempre domini della conoscenza ben identificati e in numero limitato. Essi infatti, anche in ragione dei principi del controllo terminologico, non nascono con la pretesa di rappresentare l'intero scibile, contrariamente agli obiettivi propri dei sistemi di classificazione tradizionali. Ciò rappresenta, infatti, una delle principali differenze tra queste due tipologie di sistemi di organizzazione della conoscenza.

Relativamente alle lingue di compilazione si distingue tra thesauri monolingue e thesauri multilingue, interessati da due norme distinte fino all'emanazione della ISO 25964-1:2011 e ora in essa confluite essendo alcuni principi validi per entrambe le tipologie.

La predisposizione di thesauri multilingue si rende necessaria al fine di consentire l'accesso e il recupero dell'informazione a partire da risorse informative disponibili in più lingue e indipendentemente dalla lingua di indicizzazione.

Vi si stabiliscono quindi relazioni di equivalenza interlinguistica così che sia l'indicizzatore sia l'utente non si vedano imposto l'utilizzo di una lingua dominante. Le problematiche che emergono per la definizione di tale strumento dipendono principalmente dalle differenze non solo linguistiche, ma anche concettuali, che esistono tra le lingue e, conseguentemente tra le relative culture, nelle quali il thesaurus viene compilato.

Le versioni linguistiche devono essere quanto più possibile sovrapponibili; la ISO 25964-1:2011 stabilisce che ciascuna lingua debba avere lo stesso status e che ciascun concetto debba essere rappresentato in ognuna di esse. La sola distinzione possibile è quella tra lingua d'origine e lingua di destinazione: la prima rappresenta il punto di partenza dal quale promana la traduzione o la ricerca degli equivalenti nelle lingue restanti, mentre la seconda è la lingua verso la quale si indirizza l'attività di traduzione o di ricerca.

Rappresentare i concetti in ciascuna lingua significa identifi-

care il corrispettivo semanticamente più vicino di ciascuno di essi e stabilire relazioni di equivalenza interlinguistica bidirezionali. Tuttavia, così come accade all'interno di un thesaurus monolingue, i possibili livelli di equivalenza si dispongono lungo un continuum che va da quella esatta all'assenza di equivalenza. Il primo caso, che corrisponde alla sinonimia assoluta, si verifica nel momento in cui è possibile identificare termini preferiti in ciascuna delle lingue di compilazione senza differenze di tipo semantico o culturale[74]. È possibile che tra due o più concetti esista invece una relazione di equivalenza non assoluta o quasi equivalenza, del tutto simile alla quasi sinonimia nel thesaurus monolingue: tra i termini nelle diverse lingue esistono differenze di significato, spesso dovute a differenti connotazioni culturali[75]. La relazione viene comunque stabilita se si ritiene che ai fini del thesaurus esse rappresentino il relativo concetto, affidando in alcuni casi ad una nota d'ambito l'esplicitazione della diversa copertura semantica. Qualora, invece, una delle lingue possieda un termine che rappresenta un concetto sovraordinato rispetto a quello rappresentato dai termini delle altre, si parla di equivalenza parziale o *broader/narrower equivalence*. Rientra in questo caso anche la *compound equivalence*, che si verifica quando una lingua possiede uno o più equivalenti parziali, che, in combinazione, rappresentano il concetto identificato da un solo termine nella lingua d'origine[76]. Infine, la non equivalenza interessa quei casi in cui non esistono, in una o più lingue, termini rappresentativi di un concetto espresso in un'altra lingua[77]. Spesso alla ba-

74 Es. IT Sole – EN Sun – FR Soleil – DE Sonne.

75 Es. IT Educazione – EN Education.

76 Es. IT Sicurezza – EN Security + Safety.

77 Es. *Grandes-écoles* è un concetto tipico del contesto francese, in quanto indica istituti di istruzione superiore di livello universitario che mancano nel sistema scolastico italiano. Nel caso di un thesaurus multilingue il termine verrebbe inserito in quanto tale o verrebbe tradotto letteralmente, ri-

se vi sono differenze nei contesti e nelle culture dei paesi in cui vengono parlate le lingue di compilazione del thesaurus. In quest'ultimo caso gli espedienti ai quali si ricorre possono essere quello di accettare un prestito linguistico o quello di coniare un neologismo o un calco, soprattutto laddove il prestito non sarebbe di immediata comprensione per gli utenti.

Sia la ISO 25964-1:2011, sia le Linee Guida IFLA (*International Federation of Library Associations and Institutions*) suggeriscono modalità di gestione di tutti i casi in cui non è possibile stabilire una relazione di equivalenza esatta. In generale si tratta di: note d'ambito, qualificatori soprattutto in presenza di omografi, l'assegnazione dello status di termine preferito ad un termine più generico che più si avvicina al concetto espresso in un'altra lingua.

La relazione di equivalenza interessa solo i termini non preferiti: tra quelli non preferiti non è necessario stabilire alcuna relazione, anche perché il loro numero e il loro significato potrebbero cambiare significativamente da una lingua all'altra.

La trasposizione interlinguistica deve interessare tutte le relazioni del thesaurus: in tal senso si distingue tra struttura simmetrica e asimmetrica. Nel primo caso esiste una corrispondenza tra tutti i concetti contenuti nel thesaurus in tutte le lingue di compilazione e la struttura concettuale determinata dalle relazioni gerarchiche e associative è condivisa. Nel caso di asimmetria, invece, non è possibile stabilire una corrispondenza esatta perché le relazioni definite in una lingua potrebbero non essere valide per un'altra, soprattutto nel caso di lingue molto distanti tra di loro. È necessario, quindi, ricorrere a operazioni di mappatura tra le diverse versioni linguistiche parallele del thesaurus multilingue al fine di stabilire delle corrispondenze che consentano la na-

chiedendo sempre la presenza di una nota d'ambito che ne espliciti il significato e il contesto d'uso.

vigazione e l'interoperabilità al fine di non privilegiare nessuna delle lingue del thesaurus forzando la corrispondenza tra i concetti e creando strutture che non sarebbero accettate e condivise dagli utenti. La definizione delle modalità di mappatura sono specificatamente demandate alla seconda parte della ISO 25964.

La Figura 5 mostra il dettaglio del termine *Delinquenza* nel thesaurus multilingue dell'Unione Europea EuroVoc. La ricerca può avvenire a partire da ciascuna delle lingue previste dal thesaurus e per ogni termine scelto vengono mostrati gli equivalenti interlinguistici preceduti dai codice identificativi delle relative lingue, definiti dalle ISO 639-1 e ISO 639-2. In generale, nella

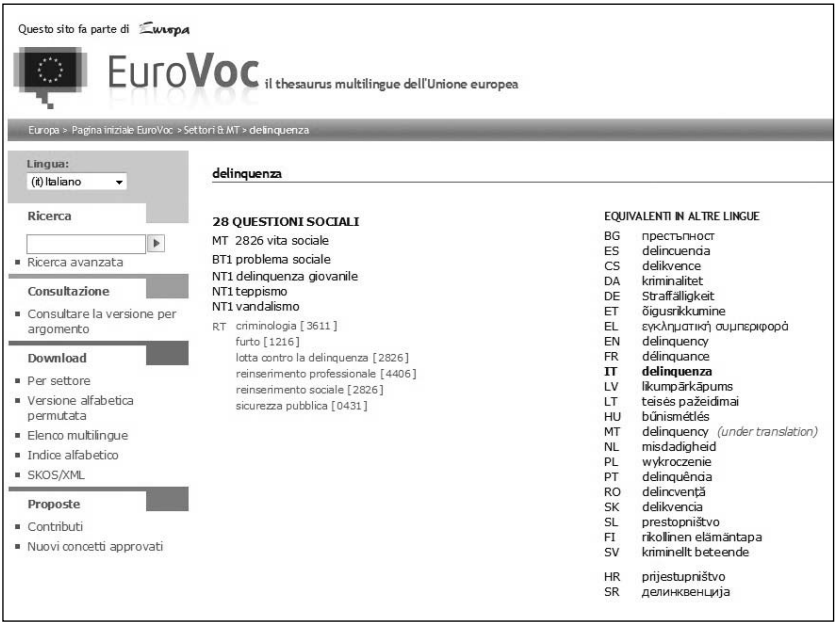


Figura 5. Thesaurus multilingue EuroVoc[78].

78 <http://eurovoc.europa.eu/drupal/?q=it>.

presentazione alfabetica viene mostrato il contesto strutturale di ciascun concetto, mentre in caso di presentazione sistematica, sarebbe opportuno poter visualizzare la struttura di due o più versioni linguistiche contemporaneamente al fine di comprenderne le corrispondenze.

Per quanto riguarda, invece, la costruzione di thesauri multilingue, la norma illustra tre approcci, che nell'ordine presentano un grado di complessità di costruzione crescente, ma consentono di ottenere risultati migliori in termini di rispetto del contesto culturale delle lingue di compilazione:

- Traduzione di un thesaurus monolingue esistente: in questo caso è alto il rischio che la lingua d'origine diventi dominante rispetto a quella di destinazione e che la struttura rispecchi poco le aspettative degli utenti;
- *Merging* di diverse versioni linguistiche: questo approccio presterebbe maggiore attenzione alle differenze linguistico-concettuali dei thesauri da integrare non assegnando a nessuna lingua un ruolo predominante. La complessità risiede nelle differenti scelte, soprattutto nel grado di specificità, compiute nei diversi thesauri;
- Costruzione ex novo del thesaurus multilingue: la costruzione simultanea fa sì che ogni lingua sia a turno d'origine e di destinazione.

3.4.1 *Thesauri e analisi a faccette*

La stretta interdipendenza tra classificazione a faccette e thesaurus è messa in evidenza in (Broughton, 2008a)[79], che afferma che *«quando si costruisce una classificazione a faccette, si prepara anche un thesaurus e per costruire un thesauro, si deve passare per una classificazione»*. In un altro studio (Broughton,

[79] BROUGHTON , V., (a), *op. cit.*, p. 13.

2008b)[80], l'autrice mette in evidenza come il valore dell'analisi a faccette in quanto supporto notevole nella realizzazione di un thesaurus non sia stato riconosciuto dagli standard o dalle linee guida esistenti, se non molto di recente. A ciò si aggiunga l'impiego spesso inappropriato del termine *faccetta*, che ha spesso ingenerato confusione circa il suo reale significato e, di conseguenza, il suo corretto utilizzo nei sistemi di classificazione e nei thesauri. Come dimostrato in (Spiteri, 1999)[81], infatti, ne vengono fornite definizioni diverse anche all'interno di quegli stessi thesauri che sono basati su tale sistema:

> *IBE, and UNICEF, for example, define facets as groups that cover related concepts. In BINDING and GENRE, facets are "gathering terms" used to arrange the hierarchical relationships amongst broader and narrower terms. ROOT and YOUTH both state that facets are fundamental categories, but do not explain what this means. In AAT, facets are homogeneous, mutually exclusive units of information that share characteristics that demonstrate their differences from each other*.

Come accennato, invece, la norma ISO 25964-1:2011 dedica un'intera sezione alla presentazione dell'analisi a faccette e all'applicazione della stessa nel processo di costruzione di un thesaurus. Vengono a tal proposito fornite le seguenti definizioni: *«Facet: Grouping of concepts of the same inherent category»*; *«Facet analysis*: *Analysis of subject areas into constituent con-*

[80] VANDA BROUGHTON (b), *A faceted classification as the basis of a faceted terminology: conversion of a classified structure to thesaurus format in the Bliss Bibliographic Classification*, ed. 2, in «Axiomathes», vol. 18, Springer, 2008, p. 196.

[81] LOUISE F. SPITERI, *The Essential Element of Faceted Thesauri*, in «Cataloging & Classification Quarterly», vol. 28, n. 4, The Haworth Press, Inc, 1999, p. 7.

cepts grouped into facets, and the subdivision of concepts into narrower concepts by specified characteristics of division»; «*Faceted classification scheme: Classification scheme in which subjects are analyzed into their constituent facets*».

Riprendendo la citazione precedente di Vanda Broghton, si può quindi affermare che applicare i principi dell'analisi a faccette alla costruzione di un thesaurus significa classificarne i concetti organizzandoli sulla base di un set di categorie precedentemente identificate e rappresentative del dominio per il quale lo stesso viene costruito e che tale strutturazione fornisce un supporto significativo nella corretta definizione delle relazioni thesaurali e quindi nella costruzione del thesaurus nella sua presentazione gerarchica[82]. Anche (Aitchison et alii, 2000)[83] sostengono che una classificazione a faccette[84] possa rappresentare un punto di partenza o una fonte per la costruzione di un thesaurus.

In ogni caso, la struttura a faccette e la visualizzazione gerarchica del thesaurus risultano complementari: in una presentazione sistematica a faccette, infatti, se si escludono le relazioni gerarchiche di tipo genere-specie, graficamente rappresentate per mezzo di rientri, non si tiene traccia delle relazioni parte-tutto e di quelle associative e di equivalenza, definite, invece, nella presentazione gerarchico-alfabetica del thesaurus.

---

82 La relazione gerarchica (BT-NT) può essere derivata dall'organizzazione dei termini in sottofaccette, nel senso che i concetti introdotti da un principio di suddivisione sono sotto-ordinati del concetto al quale il principio stesso viene applicato (ciò è valido su più livelli di strutturazione), mentre la relazione associativa riguarda i concetti introdotti dal medesimo *node label* e quindi collocati sullo stesso livello gerarchico o concetti appartenenti a diverse faccette e tra i quali si identifica una relazione semantica.

83 JEAN AITCHISON, DAVID BAWDEN, ALAN GILCHRIST, *Thesaurus Construction and use: a practical manual*, ed. 4, Londra, ASLIB, 2000, p. 69.

84 Per una presentazione delle differenze in termini di obiettivi e struttura sistema di classificazione e thesaurus si veda TUDHOPE, D., BINDING, C., *op. cit.*, pp. 211-222.

Le origini dell'analisi a faccette risalgono all'ideazione e alla pubblicazione nel 1934 della *Faceted Classification* (FC) o *Colon Classification*[85] (CC) da parte di Ranganathan, bibliotecario e matematico indiano, che propose un approccio decisamente innovativo rispetto alle tradizionali classificazioni biblioteconomiche comunemente adottate per la sistemazione del materiale librario[86]. A partire dalla propria esperienza nell'utilizzo della CDD, egli ne individuò i principali limiti, quali l'impossibilità a rappresentare tutti i temi trattati in un'opera, ad enumerare tutti i soggetti o ad accoglierne di nuovi e formulò i principi di un nuovo approccio che avrebbe consentito di superarli attraverso un sistema basato su operazioni di scomposizione e ricomposizione dei soggetti da classificare[87]. Alla base di tale principio vi è l'i-

85 Così chiamata perché utilizza il segno di punteggiatura dei due punti (in inglese Colon) come separatore tra i soggetti.

86 La Classificazione Decimale Dewey (CDD) Progettato da Melvil Dewey per l'Amherst College nel 1873, rappresentò una vera e propria rivoluzione del campo della biblioteconomia in quanto introdusse il metodo della notazione decimale che consente un *«ordinamento monodimensionale di ospitalità infinita»* (ALFREDO SERRAI, *Le classificazioni: idee e materiali per una teoria e per una storia*, Firenze, Leo S. Olschki Editore, 1970, p. 283), nel senso che la successione delle classi avviene secondo un solo principio di suddivisione per volta, ma è possibile estendere la struttura in maniera potenzialmente infinita. La Classificazione Decimale Universale (CDU), derivata dalla precedente, fu elaborata da Paul Otlet e Henri La Fontaine nel biennio 1893-1894. Rispetto alla Dewey, il suo obiettivo principale fu quello di classificare oggetti documentali piuttosto che quello di collocare il materiale bibliotecario. Tali sistemi sono ancora oggi ampliamente applicati.

87 SHIYALI RAMAMRITA RANGANATHAN, *Colon Classification, I: Schedules for Classification*, ed. 7, Gopinath M.A. (a cura di), Sarada Ranganathan Endowment for Library Science, 1989, (ed. 1, 1933), p. 3.
La classificazione del materiale bibliotecario richiede, dunque, una fase di analisi e di scomposizione del soggetto sulla base delle categorie identificate, seguita da un'attività di traduzione del linguaggio naturale in linguaggio controllato attraverso la verifica dei concetti nelle tavole di clas-

dentificazione di cinque categorie fondamentali o faccette che permettono di analizzare qualsiasi soggetto, poiché ognuna di esse ne mette in evidenza un particolare aspetto. Esse sono: *Personality*, *Matter*, *Energy*, *Space*, *Time* (PMEST)[88].

Nonostante la classificazione a faccette abbia rappresentato un'intuizione innovativa per il superamento dei limiti propri dei sistemi gerarchici, la sua applicazione in contesto bibliotecario è fin da subito risultata complessa, anche solo per il semplice fatto di dover assegnare ad un volume più collocazioni. Tuttavia, le sue potenzialità sono state riscoperte nelle pratiche di indicizzazione e di recupero dell'informazione in ambiente digitale data la virtualità degli oggetti da classificare e la necessità di sistemi multidimensionali (Marino, 2004)[89]. Lo sviluppo quindi di thesauri a faccette[90] ha ricevuto e sta ricevendo un forte impulso e l'attenzione dedicatagli dalla recente norma ISO ne è una conferma.

L'organizzazione dei contenuti nel web ha visto l'applicazione massiccia tanto di sistemi gerarchici che di sistemi a faccette, seppure non nel pieno rispetto dei principi sui quali sono fondati (Rosenfeld, Morville, 2002)[91]. I sistemi di classificazione tra-

sificazione e nel linguaggio ordinale attraverso l'attribuzione del codice di notazione a ciascuno di essi. La fase di sintesi prevede la definizione di un codice unico dato dalla ricomposizione dei singoli codici attribuiti agli aspetti nei quali il soggetto è stato analizzato.

88 *Personality*: oggetti di studio delle varie discipline; *Matter* proprietà o materiali; *Energy*: le azioni o i processi che si verificano in una disciplina; *Space*: concetti relativi allo spazio; *Time*: concetti relativi al tempo.

89 Cfr. VITTORIO MARINO, *Classificazioni per il Web. I vantaggi dell'adozione di schemi a faccette*, Associazione Italiana Biblioteche (AIB) - WEB, 2004.
<http://www.aib.it/aib/contr/marino1.htm>

90 Tra i principali thesauri a faccette rientra il già citato AAT, che rappresenta un'applicazione rigorosa ed esemplificativa dei principi alla base di tale approccio.

91 LOUIS ROSENFELD, PETER MORVILLE, *Information Architecture for the World Wide Web*, ed. 2, O'Reilly, 2002, p. 208.

dizionali a cui si è accennato, e più in generale, i sistemi gerarchici, sono basati sull'enumerazione di tutte le classi e sono caratterizzati dalla difficoltà di accogliere integrazioni, se non a condizione di modifiche consistenti dello schema di base, e da una struttura che costringe l'utente a navigare solo secondo il percorso definito. I sistemi a faccette sono, invece, come decisamente più flessibili. Questi infatti, non precludendo integrazioni successive in termini di categorie e principi di analisi, si adattano all'evoluzione e alle esigenze di aggiornamento dei contenuti sul Web, garantendo pluridimensionalità, persistenza, scalabilità e flessibilità (Rosati, 2003)[92].

Significativo, in tale ottica è anche il lavoro svolto dal *Classification Research Group (CRG)*[93], che ha accolto i principi della metodologia a faccette approfondendoli e perfezionandoli soprattutto in riferimento alla revisione del suddetto schema di faccette fondamentali di Ranganathan e alla conseguente definizione di un set di faccette più ampio e di più immediata applicabilità. Lo schema definito dal CRG è così costituito: *thing, types, parts, properties, materials, processes, activities, products, by products, patients, agents, space and time*[94].

Il punto di forza di tale schema risiede nella sua potenziale

---

92 LUCA ROSATI, *La classificazione a faccette fra Knowledge Management et Information Architecture (parte I)*, It Consult, 2003. <http://www.itconsult.it/knowledge/articoli/pdf/itc_rosati_faccette_e_KM.pdf>.
Classificazione sulla base di molteplici attributi; Cambiamenti limitati dovuti al fatto che le proprietà rappresentano attributi intrinseci dei concetti; Possibilità di aggiungere nuove faccette e nuovi principi di suddivisione; Ricerca a partire da un solo attributo da più attributi in combinazione.

93 Gruppo di ricercatori inglesi attivi nel campo della biblioteconomia e della classificazione costituitosi a Londra all'inizio degli anni 50 del secolo scorso.

94 Le definizioni di ciascuna faccetta sono tratte da BROUGHTON, V. (a), *op. cit.*, pp. 259-281.

applicazione a qualsiasi dominio oggetto di interesse, data la genericità delle categorie e, al tempo stesso, l'elevata probabilità di essere rappresentative degli aspetti di un dato ambito semantico. Data la notevole specificità di ciascun settore e le differenti finalità che possono essere alla base della costruzione di un simile strumento di classificazione, è possibile adottare anche parzialmente lo schema proposto, scartando alcune faccette, che risultano non applicabili, accorpandone delle altre o anche prevedendone alcune aggiuntive, laddove quelle iniziali non dovessero rivelarsi sufficienti per la descrizione completa del dominio.

Applicare i principi della classificazione a faccette alla costruzione di un thesaurus implica la scomposizione del dominio di interesse in categorie (faccette) rappresentative dello stesso: così come un oggetto viene analizzato nelle sue caratteristiche intrinseche, un dominio viene analizzato negli aspetti che ne ricoprono l'intero ambito semantico. All'interno di ciascuna faccetta i concetti possono essere ulteriormente organizzati attraverso principi di suddivisione o *node labels* o etichette di snodo

«Cose: comprende i concetti che sono i principali oggetti di studio per un argomento o disciplina; Parti: comprende i concetti che sono parti dei concetti della categoria delle entità; Proprietà: concetti che sono proprietà o attributi di concetti appartenenti alla categoria principale; Materiali: Raccoglie i concetti collegati a sostanze e materiali di tutti i tipi […]; Processi: raccoglie i concetti di azioni che accadono spontaneamente, non compiute da agenti umani; Attività: raccoglie i concetti di azioni condotte su di un oggetto da un agente umano; Pazienti: raccoglie i concetti che sono oggetti di azioni, […] Dovrebbe comprendere gli oggetti impiegati in fasi intermedie di processi produttivi quando i prodotti finali sono le entità primarie; Prodotti: comprende i prodotti di attività quando questi non appartengono alla categoria primaria delle entità; Prodotti intermedi: raccoglie i prodotti intermedi di attività […];Agenti/Strumenti: comprende i concetti per mezzo dei quali si compiono delle azioni […]; Spazio: raccoglie i concetti relativi a luoghi […]; Tempo: comprende i concetti legati al tempo […]».

che rappresentano caratteristiche intrinseche dei concetti stessi[95]. La Figura 6 mostra un estratto del Nuovo Soggettario della Biblioteca Nazionale Centrale di Firenze, nel quale è possibile distinguere i principi di suddivisione (tra parentesi quadre) e la strutturazione dei termini al loro interno.

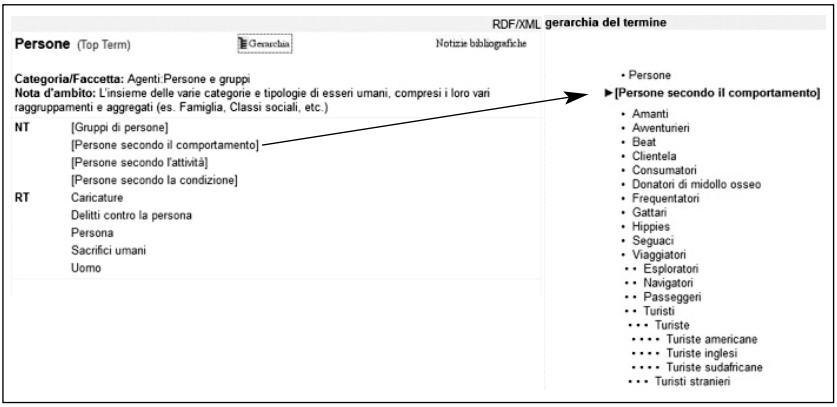


Figura 6. Nuovo Soggettario.

Tra i principi fondamentali alla base di un sistema a faccette rientrano:

– Approccio analitico – sintetico: rappresenta l'aspetto che maggiormente contraddistingue tale metodo rispetto alle classificazioni enumerative. Si susseguono due fasi: la prima orientata all'analisi del dominio di interesse o degli oggetti da classificare al fine di individuare gli aspetti secondo i quali possono essere scomposti; la seconda a sintetizzare e combinare i concetti appartenenti a più suddivisioni in stringhe più complesse[96];

95 Data la loro funzione, non sono considerati termini o concetti all'interno del thesaurus.

96 Il processo di sintesi interessa anche i codici eventualmente associati a ciascun concetto ed è in tal senso che si parla di notazione sintetica.

- Mutua esclusività: garantisce che ciascun insieme o sottoinsieme di termini venga introdotto da un sola caratteristica di suddivisione per volta e che si evitino fenomeni di sovrapposizione semantica[97];
- Ordine di elencazione delle faccette[98]: stabilisce l'ordine in base al quale le faccette devono essere presentate in un layout sistematico del thesaurus o in un menù di navigazione; va dal generale al particolare, secondo un principio di concretezza crescente (quindi, nel caso del set del CRG, da *Tempo* a *Cose*), in quanto ritenuto più intuitivo da parte degli utenti;
- Ordine di citazione standard: ordine che determina la sintassi in base alla quale i concetti appartenenti a diverse faccette dovrebbero essere inseriti all'interno della stringa di soggetto in un contesto di pre-coordinazione (Iyer, 2012)[99]; va dal particolare al generale, quindi prevede una sequenzialità inversa rispetto al principio precedente.

Per quanto riguarda, invece, l'ordinamento dei concetti all'interno dei raggruppamenti, è consigliabile inserirli in base ad un principio che non sia quello alfabetico, molto poco significativo,

97 *«[…] i termini che appartengono ad una stessa sottofaccetta come mutuamente esclusivi, ossia escludentisi a vicenda. Ciò significa che, diversamente dai termini che appartengono a differenti faccette o a differenti sottofaccette di una stessa faccetta, i termini che appartengono ad una stessa sottofaccetta non possono combinarsi tra loro, non possono dare luogo ad una sovrapposizione o intersezione di classe (p.e., Coniugi celibi). Questa proprietà (mutua esclusione) indica la fine del processo di divisione di una classe».*
BIBLIOTECA NAZIONALE CENTRALE DI FIRENZE, *Nuovo Soggettario*, Milano, Editrice Bibliografica, 2006, p. 82.

98 BROUGHTON, V. (a), *op. cit.*, p. 201.

99 HEMALATA IYER, *Classificatory Structures: Concepts, Relations and Representation*, Würzburg, Ergon Verlag, 2012, p. 128.

ma piuttosto quello cronologico, o in funzione alla complessità, della sequenzialità, ecc.

L'approccio a faccette risulta estremamente adatto ad un dominio specialistico data l'omogeneità degli oggetti da classificare, così che gli stessi possano essere analizzati a partire da caratteristiche comuni. L'adozione in contesti multidisciplinari o multi tematici risulterebbe, per tale motivo, decisamente più complessa. Lo svantaggio che anche le precedenti norme riconoscevano alla struttura a faccette, infatti, risiedeva nella separazione, attraverso il loro inserimento in faccette diverse, di concetti appartenenti ad una stessa disciplina o ambito.

3.5 *Costruzione di un thesaurus*

Così come per le funzionalità, anche le modalità di costruzione di una tassonomia non sono esplicitamente descritte nelle citate norme. Tuttavia, se si escludono le attività e le indicazioni che conducono verso la definizione di elementi propri dei thesauri, in particolare la gestione della sinonimia e dei concetti correlati, le fasi di costruzione si possono considerare comuni ai due sistemi. La raccolta dei termini, l'organizzazione gerarchica e l'eventuale identificazione delle faccette, infatti, interessano entrambi. Si farà comunque riferimento ai thesauri, essendo il relativo processo di costruzione maggiormente standardizzato.

Preliminare al processo di costruzione vero e proprio è la fase di progettazione, durante la quale, oltre allo studio di fattibilità che interessa aspetti propriamente gestionali (es. risorse umane e materiali), è importante stabilire:

- a chi si rivolge, intendendo soprattutto gli utenti finali che lo utilizzeranno come supporto nelle loro ricerche o come fonte di termini, distinguendo principalmente tra esperti di dominio e utenti comuni senza competenze specifiche nel settore;
- per quali scopi viene costruito, ovvero indicizzazione di una specifica collezione documentale, IR, organizzazione

della conoscenza, ecc., dal momento che alcune scelte dipendono strettamente dalla funzionalità alla quale lo strumento dovrà assolvere (Folino et alii, 2012)[100];
- l'eventuale integrazione in un sistema di *content management*;
- la costruzione ex novo o la traduzione/adattamento di un thesaurus esistente;
- la struttura di classificazione alla base dell'organizzazione dei concetti.

La prima fase di costruzione del thesaurus consiste nella raccolta dei termini da inserire al suo interno. Tale attività può avvenire a partire da risorse terminologiche esistenti in letteratura e relative al medesimo dominio di interesse del thesaurus, quali lessici o glossari, e dall'estrazione dei termini a partire da un corpus documentale[101] appositamente costituito.

D'accordo con la linguistica dei corpora, di cui il corpus rappresenta l'oggetto di studio e la cui finalità è quella di analizzare l'utilizzo della lingua in contesti d'uso reali (Lenci et alii, 2005)[102], il set di documenti costituito deve essere un campione rappresentativo rispetto alla popolazione di riferimento, nel sen-

[100] ANTONIETTA FOLINO, FRANCESCA IOZZI, MARIA TAVERNITI, *Gestione documentale in ambiente digitale*, in «Archivistica e Documentazione», Guarasci R. (a cura di), vol. 7, Marzi, Cosenza, Comet Editor Press, 2012, pp. 152-158.

[101] *«A collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research».* JOHN SINCLAIR, *Trust the text: language, corpus and discourse*, Londra, Routledge, 2004, p. 14.

[102] ALESSANDRO LENCI, SIMONETTA MONTEMAGNI, VITO PIRRELLI, *Testo e computer: Elementi di Linguistica Computazionale*, Roma, Carocci Editore, 2005, p. 25.

so che le informazioni ottenute dalla sua analisi devono poter essere generalizzate alla popolazione intera, ovvero alla lingua o ad una sua varietà. Per tali ragioni, deve essere sufficientemente grande[103] ed equilibrato in termini di tipologie testuali in esso contenute (es. leggi, norme, articoli scientifici, riviste di settore, opuscoli informativi, ecc.), sempre nel rispetto degli obiettivi del thesaurus. I tipi di risorse infatti si differenziano per il pubblico a cui si rivolgono e per il grado di specializzazione della terminologia impiegata, caratteristiche importanti per la selezione dei termini da inserire in un thesaurus.

I documenti non sono, quindi, creati ad hoc, ma vengono selezionati tra quelli esistenti sulla base di determinati criteri, che, nel caso specifico, variano in funzione degli obiettivi e delle caratteristiche che il thesaurus dovrà possedere, affinché possano rendere conto del reale utilizzo della terminologia di interesse. Tra i criteri più rilevanti ai fini della costruzione di un thesaurus rientrano la lingua dei documenti selezionati (corpus monolingue o multilingue) e l'ambito tematico (corpus specialistico o generale).

La scelta di costruire un corpus documentale nel dominio di interesse del thesaurus permette di non limitare i termini a quelli estratti dalla collezione di documenti da indicizzare[104], sia in ragione di un eventuale ampliamento di tale insieme, sia per prevedere il maggior numero di termini potenzialmente utilizzabili dagli utenti.

L'estrazione terminologica eseguita sul corpus può avvenire manualmente o, nella maggior parte dei casi e soprattutto in presenza di corpora di grandi dimensioni, in maniera semiautomati-

[103] Le dimensioni di un corpus si misurano in *tokens*, ovvero in parole-unità distinte.

[104] Che, quindi, si differenzia dal corpus per il non rispetto della rappresentatività statistica.

ca[105], con l'ausilio di strumenti software dedicati[106]. In quest'ultimo caso, a seguito del processo di estrazione, si ottiene una lista di candidati termini[107] dalla quale selezionare quelli effettivamente rappresentativi del dominio di conoscenza che deve essere strutturato per mezzo del thesaurus.

---

[105] Si parla di semiautomatismo poiché l'intervento umano per la validazione dei candidati termini estratti resta ancora insostituibile.

[106] Tali software si basano sull'assunto che dalla frequenza con la quale un termine occorre all'interno di un documento e dell'intero corpus, dipenda la sua rappresentatività per il dominio oggetto di analisi: quindi maggiore è il valore della frequenza, maggiore è la sua significatività. Per la lingua italiana si può fare riferimento al software T2K (text-to-Knowledge) sviluppato dall'Istituto di Linguistica Computazionale di Pisa, il quale si basa sia su misure statistiche, in particolare sulla funzione tf*idf (*term frequency***inverse document frequency)*, che calcola la frequenza di ogni *termine* all'interno di un documento (*TF = term frequency*), relazionata con la frequenza inversa del *termine* stesso all'interno del *corpus* documentale (IDF= *Inverse Document Frequency*), che su regole linguistiche. L'analisi di un testo prevede ad esempio fasi di segmentazione in parole (tokenizzazione), di attribuzione della categoria grammaticale (*Part-of-Speech Tagging*), di riconoscimento di sintagmi sintattici (*chunking*), di identificazione di parole grammaticali (articoli, congiunzioni, preposizioni, ecc.) da escludere dal processo di estrazione terminologica (stop list). Oltre al glossario terminologico si ottengono in output anche relazioni semantiche tra i termini estratti: gerarchiche e di similarità semantica. Per uno studio più approfondito di tali tematiche si rimanda a FELICE DELL'ORLETTA, ALESSANDRO LENCI, SIMONE MARCHI, SIMONETTA MONTEMAGNI, VITO PIRRELLI, GIULIA VENTURI, *Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio*, in «Aida Informazioni», a. XXVI gennaio-giugno, n. 1-2, 2008, pp. 185-206.

[107] *«Le 'parole' estratte automaticamente per mezzo di strumenti informatici, acquisiscono la dignità d'essere definite 'descrittore' oppure 'termine' solo in seguito ad un processo di validazione da parte di un esperto di dominio»*.
MARIA TAVERNITI, *Fra terminologia e documentazione: estrazione automatica di voci indice da corpora documentali della Pubblica Amministrazione*, in «AIDAinformazioni», a. XXVI gennaio-giugno, n. 1-2, 2008, p. 232.

Tale selezione dovrebbe avvenire sulla base della combinazioni di più criteri, tra i quali la frequenza d'uso[108], la conoscenza di esperti di dominio e l'autorevolezza della fonte dalla quale i termini sono stati estratti.

Successivamente alla costituzione del set di termini da inserire nel thesaurus, si passa alla normalizzazione degli stessi secondo quanto previsto dalla norma ISO 25964-1:2011. Ci si riferisce in particolare alla forma del termine, ovvero alla scelta del singolare o del plurale[109], della categoria grammaticale[110], all'inserimento di termini composti[111], ecc.

Riguardo a quest'ultimo punto, è opportuno sottolineare come in un thesaurus a faccette si protenda, laddove possibile, verso la scomposizione dei concetti complessi, soprattutto quando i termini singoli che li costituiscono possono appartenere a più faccette al fine di evitare l'inserimento di termini troppo lunghi, che aumenterebbero la complessità e i livelli gerarchici del thesaurus. La combinazione potrà avvenire, secondo il principio della sinte-

[108] La frequenza d'uso non può da sola determinare l'accettabilità o meno di un termine, soprattutto in presenza di concetti innovativi, che in quanto tali avranno poche occorrenze all'interno di un corpus documentale.

[109] In genere si adotta il plurale per quei termini che rappresentano entità concrete e quindi numerabili (es. Edifici) e il singolare in presenza di concetti astratti e non numerabili (es. Architettura).

[110] Si preferisce ad esempio il verbo sostantivato alla forma dell'infinito presente.

[111] In presenza di concetti complessi, la norma distingue tra casi in cui è preferibile mantenere e inserire nel thesaurus i termini nella loro forma composta e casi in cui è preferibile inserire il risultato della loro scomposizione. A favore della prima opzione rientrano ad esempio le seguenti motivazioni: frequenza d'uso; rischio di ambiguità; perdita o mutamento di significato; presenza di metafore. In generale, soprattutto per finalità di indicizzazione, si tiene conto del grado di specificità che si intende raggiungere, tenendo presente che i termini composti aumentano tale livello, e della capacità di discriminazione dei concetti in base al contenuto della collezione da indicizzare.

si, nell'attribuzione del soggetto al documento o, in una logica di post-coordinazione, nella formulazione di un'interrogazione.

L'organizzazione di tali termini prevede la definizione delle relazioni semantiche, così come illustrate nel paragrafo 3.1 e l'assegnazione, nel caso di thesauri a faccette, di ciascun termine alla o alle categorie di appartenenza.

La definizione della struttura del thesaurus, soprattutto in quest'ultimo caso può seguire un approccio induttivo o deduttivo: nel primo caso si procede dall'alto verso il basso definendo preliminarmente la struttura gerarchica e quindi l'insieme delle faccette e subito dopo dei principi di suddivisione per poi organizzare i termini al loro interno nella misura in cui questi vengono identificati e selezionati, mentre nel secondo si parte dal basso, ovvero dai termini scelti, la cui analisi permette di definire la struttura per mezzo della quale organizzarli. Nella maggior parte dei casi i due approcci sono complementari.

Le modalità di visualizzazione della struttura del thesaurus illustrate nella ISO 2788:1986 sono: alfabetica, sistematica, distinguendo ulteriormente tra organizzazione in settori o discipline e organizzazione a faccette, e grafica; la complementarità tra di esse veniva garantita dalla presenza della notazione. La norma, si ribadisce, si riferiva ai thesauri a stampa, per cui alcune delle indicazioni fornite e delle problematiche evidenziate non sono più valide in ambiente digitale.

Nel layout alfabetico i termini sono elencati secondo quest'ordine e per ciascuno di essi vengono evidenziate le relazioni nelle quali è coinvolto; in quello sistematico i termini sono organizzati generalmente in gerarchie (ad esempio a partire da *top term* che rappresentano i concetti con il più alto livello di genericità, quali le discipline) o in categorie nel caso di thesaurus a faccette (è possibile anche un'organizzazione mista)[112].

[112] La visualizzazione è possibile anche tramite liste permutate nelle quali i

Gli accorgimenti nella costruzione della versione a stampa riguardano ad esempio i casi di poligerarchia, per cui un termine con più di una collocazione veniva inserito completo delle relazioni nelle quali era coinvolto solo nella gerarchia principale alla quale apparteneva, prevedendo per le altre solo una forma di rinvio: il tutto per esigenze di spazio e di semplificazione della presentazione del thesaurus. Simili indicazioni non hanno ovviamente senso di esistere nei contesti digitali, per i quali già la norma britannica introduce il concetto di display informatico e di strumenti software per la costruzione dei thesauri. Ciascun termine è ad esempio inserito una volta per tutte pur moltiplicandone le possibili collocazioni: è l'utente infatti ad avere la possibilità di navigare ed esplodere le relazioni thesaurali secondo le proprie esigenze, scegliendo gli estratti che vuole rendere visibili e muovendosi, attraverso link ipertestuali, non solo da un concetto ad un altro, ma anche da un concetto ai documenti ad esso associati. Egli può navigare tramite le relazioni e passare da una forma di presentazione all'altra (multilivello, singolo termine, per categoria, per top term, per faccette, ecc.) superando la linearità delle presentazioni a stampa. Questo tipo di display viene definito *classificato esteso* (Calvitti, Viti, 2009)[113].

Da un punto di vista meramente tecnico, è opportuno accennare ai software oggi esistenti per la costruzione dei thesauri in formato elettronico, al fine di evidenziarne le caratteristiche principali che garantiscono una corretta e coerente realizzazione di tale strumento:

– Non imporre limiti al numero di termini e/o relazioni che è possibile definire;

termini composti vengono elencati in ordine alfabetico e ciascun descrittore di cui si compongono può o meno mantenere la propria posizione nella stringa (KWIC - *Keyword in context;* KWOC - *Keyword out of Context*).

[113] CALVITTI, T., VITI, E., *op. cit*., p. 314.

- Inserire in maniera automatica le relazioni inverse se le stesse sono simmetriche (es. BT/NT);
- Esportare il thesaurus in formati standard, quali XML e, nel contesto del Web semantico, anche SKOS;
- Prevedere diversi layout;
- Impedire la definizione di relazioni non corrette (che, ad esempio, coinvolgono termini non preferiti);
- Permettere l'organizzazione a faccette e quindi la gestione delle stesse e dei *node label*;
- Garantire funzioni di ricerca e di navigazione all'interno del thesaurus;
- Gestire il multilinguismo.

4 Conclusioni

Come più volte ribadito nel corso del capitolo, l'evoluzione delle tecnologie per il recupero dell'informazione e, in particolare, l'ampia diffusione delle ontologie come strumenti per eccellenza del Web Semantico, hanno determinato un ripensamento dei concetti stessi di thesaurus e tassonomia. Una simile rivalutazione ha fatto sì che gli stessi non continuassero ad essere considerati come strumenti ormai obsoleti, determinando la valorizzazione e il potenziamento delle loro funzionalità in ambiente digitale.

È importante sottolineare, quindi, come le differenze esistenti tra thesauri e ontologie, alle quali si è fatto accenno, non debbano indurre a considerare il thesaurus come uno strumento semanticamente meno ricco e per questo meno valido di un'ontologia: le due tipologie di KOS, infatti, sono nate in risposta ad esigenze diverse e la predisposizione dell'uno o dell'altro dipende dalle caratteristiche e dagli obiettivi dei contesti di applicazione.

Proprio in virtù di tali differenze, quindi, esistono dei casi in

cui un'ontologia[114] assolverebbe meglio di un thesaurus a determinate funzionalità[115]: le metodologie di passaggio da un thesaurus ad un'ontologia sono oggetto di interesse scientifico e in letteratura esistono proposte di approcci sperimentati su thesauri esistenti. Si tratta essenzialmente di processi di arricchimento semantico orientati all'identificazione della natura delle relazioni esistenti tra i concetti. L'approccio proposto da (Chrisment et alii, 2006) per la trasformazione di un thesaurus in un'ontologia leggera di dominio può essere applicata ai thesauri monolingue con struttura gerarchica realizzati conformemente alle norme ISO 2788:1986 e ANSI Z39.19:2005. La metodologia è stata concretamente sperimentata sul thesaurus di astronomia *dell'International Astronomical Union* (IAU) e la sua originalità rispetto ad altri studi aventi i medesimi obiettivi risiede nel fatto che le operazioni di trasformazione sono state condotte non solo a partire dal thesaurus, ma anche da un corpus documentale appositamente costituito, al fine di estrarre le informazioni implicite o non modellizzate nel thesaurus stesso. In particolare, la relazione gerarchica di tipo genere-specie nell'ontologia è definita a partire dalle relazioni di tipo BTG-NTG presenti nel thesaurus. Questa fase prevede, inoltre, l'inserimento di classi più generi-

[114] Per ontologia si intende la rappresentazione formale ed esplicita di una concettualizzazione condivisa, interpretabile tanto da un operatore umano che da una macchina.
Cfr. NICOLA CAPUANO, *Ontologie OWL: Teoria e Pratica*, in «Computer Programming», n 148, luglio-agosto 2005, pp. 59-64.

[115] Si pensi ad esempio all'utilizzo congiunto di ontologie e di sistemi capaci di effettuare delle inferenze e di fare dei ragionamenti a partire dalle informazioni modellizzate, estraendo da queste conoscenza nuova ed implicita e sfruttando le restrizioni espresse sulle relazioni e le proprietà attribuite alle relazioni stesse (transitività, simmetria, ecc.). Esse inoltre rappresentano lo strumento per eccellenza del Semantic Web e consentono di sviluppare sistemi di IR con funzionalità avanzate e che consentono una maggiore interazione con gli utenti.

che, poiché spesso i thesauri al livello gerarchico più elevato contengono un gran numero di termini, al fine di agevolare le operazioni di navigazione. L'identificazione di tali classi, che rappresentano o concetti di dominio o concetti generici per strutturare il dominio stesso, avviene attraverso due processi automatici: il raggruppamento dei concetti le cui etichette presentano la stessa testa lessicale per creare delle classi ad un primo livello della gerarchia e la definizione di categorie astratte (proprietà, fenomeni, eventi, strumenti, oggetti, ecc.) alle quali associare i concetti di dominio a partire da un'ontologia generica esistente, ovvero WordNet[116]. La specificazione delle relazioni associative avviene a partire dall'analisi sintattica del corpus documentale che permette di estratte il contesto linguistico delle etichette relative a ciascun concetto. Tali contesti sono poi raggruppati sulla base delle categorie astratte alle quali i concetti appartengono, le relazioni sono inizialmente definite a livello di categorie astratte, ma vengono poi inserite tra i concetti a queste appartenenti a partire dai termini interessati da una relazione RT all'interno del thesaurus. Ulteriori relazioni associative non presenti nel thesaurus sono identificate a partire dall'analisi dei documenti del corpus, basandosi sulla frequenza dei termini che cooccorrono con le etichette dei concetti e sui risultati di un'analisi distribuzionale che tiene conto della similitudine dei contesti in cui occorrono i sintagmi.

(Soergel et alii, 2004)[117] descrive invece l'approccio definito per la trasformazione in ontologia del thesaurus multilingue AGROVOC. Il modello proposto prevede una netta distinzione tra il livello concettuale, che si riferisce al significato, il livello terminologico, relativo ai termini utilizzati per rappresentare i

[116] <http://wordnet.princeton.edu/>.

[117] Cfr. DAGOBERT SOERGEL, BORIS LAUSER, ANITA LIANG, FREHIWOT FISSEHA, JOHANNES KEIZER, STEPHEN KATZ, *Reengineering Thesauri for New Applications: the AGROVOC Example*, in «Digit. Inf.», vol. 4, n. 4, 2004.

concetti, e il livello di stringa, ovvero le varianti lessicali possibili per ciascun termine. I concetti devono essere assegnati a categorie generiche (processi, sostanze, ecc.) che vincolano le tipologie di relazioni dalle quali gli stessi possono essere interessati. L'inserimento di nuove informazioni e la precisazione di quelle esistenti avviene con l'ausilio di un *ontology editor* e attraverso il riconoscimento di pattern ricorrenti sui quali l'editor formula delle regole applicabili a casi identificati come simili.

Di notevole interesse anche le più attuali ricerche nell'ambito del Web Semantico o Web di dati, che riguardano lo sviluppo di *linked data*[118] e l'integrazione/allineamento dei sistemi di organizzazione della conoscenza. Esemplificativa in tal senso la Figura 7 che riporta i *dataset open* disponibili sulla rete e collega-

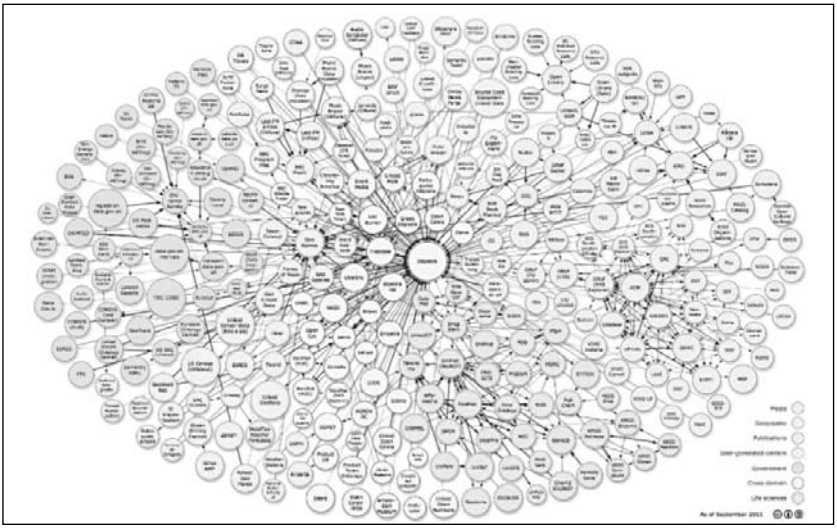


Figura 7. Linked Open Data Cloud.

[118] *«Si costruisce così un reticolo di dati collegati (linked data, appunto) appartenenti a un dominio (che costituisce il contesto di partenza), collega-*

ti tra di loro. La sfida che l'integrazione/condivisione dei dati porta con sé è rappresentata dalla possibilità di accedere a tutte le risorse informative indicizzate tramite i concetti appartenenti a vocabolari controllati interconnessi e allineati tra di loro, per cui tali studi risultano di notevole importanza per la definizione di nuovi strumenti di organizzazione della conoscenza e per l'integrazione di quelli esistenti al fine di garantire un accesso federato e al tempo stesso controllato all'informazione e ai dati presenti nel Web.

Bibliografia

ANSI/NISO Z39-19:2005, *Guidelines for the construction, format, and management of monolingual controlled vocabularies*

AITCHISON, J., BAWDEN, D., GILCHRIST, A., *Thesaurus Construction and use: a practical manual*, ed. 4, Londra, ASLIB, 2000

BIBLIOTECA NAZIONALE CENTRALE DI FIRENZE, *Nuovo Soggettario*, Milano, Editrice Bibliografica, 2006

BROUGHTON, V., (a), *Costruire Thesauri: strumenti per indicizzazione e metadati semantici*, Cavaleri, P. (a cura di), Ballestra L., Venuti L. (traduzione di), Milano, Editrice Bibliografica, 2008

BROUGHTON, V., (b), *A faceted classification as the basis of a faceted terminology: conversion of a classified structure to thesaurus format in the Bliss Bibliographic Classification*, ed. 2, in «Axiomathes», vol. 18, Springer, 2008, pp. 193-210

BS 8723:2004-2008, *Structured vocabularies for information retrieval – Guide*

CABRÉ, M.T., *Terminology: theory, methods and applications*, Sager J.C. (ed.), DeCesaris J.A. (traduzione di), Philadelphia PA, John Benjamins, 1998

CALVITTI, T., VITI, E., *Da ISO 2788 ai nuovi standard per la costruzione e*

to a sua volta ad altri set di dati esterni, ovvero fuori dal dominio, in un contesto di relazioni sempre più estese».

MAURO GUERRINI, TIZIANA POSSEMATO, *Linked data: un nuovo alfabeto del Web Semantico*, in «Biblioteche oggi», aprile 2012, p. 7.

l'interoperabilità dei vocabolari controllati: un'analisi comparativa, in «Bollettino AIB», vol. 49, n. 3, settembre 2009, pp. 307-322

CAPUANO, N., *Ontologie OWL: Teoria e Pratica*, in «Computer Programming», n 148, luglio-agosto 2005, pp. 59-64

CASSON, E., *Dai thesauri ai vocabolari controllati: alcune novità introdotte nell'ultima edizione dello standard ANSI/NISO Z39.19-2005*, in «AIDAinformazioni», a. 24, n. 1-2, gennaio-giugno 2006, pp. 69-77

CHETI, A., *Manuale ipertestuale di analisi concettuale*, 1996 <http://biblioteche.unibo.it/manuals/html_1/HOME.HTML>

DELL'ORLETTA, F., LENCI, A., MARCHI, S., MONTEMAGNI, S., PIRRELLI, V., VENTURI, G., *Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio*, in «Aida Informazioni», a. XXVI gennaio-giugno, n. 1-2, 2008, pp. 185-206

DEXTRE CLARKE, S.G., LEI ZENG, M., *From ISO 2788 to ISO 25964: The Evolution of Thesaurus Standards towards Interoperability and Data Modeling*, in «Information Standards Quarterly», vol. 24, n. 1, 2012, pp. 20-26

ERTZSCHEID, O., GALLEZOT, G., *Etude exploratoire des pratiques d'indexation sociale comme une renégociation des espaces documentaires. Vers un nouveau big bang documentaire?*, in Document numérique et société, Charton G., Broudoux E. (a cura di), ADBS Éditions, 2006

FLORIO, I., *La normativa standardizzata per la gestione delle documentazione tra Italia e Francia*, Rubbettino Editore, 2011

FOLINO, A., IOZZI, F., TAVERNITI, M., *Gestione documentale in ambiente digitale*, in «Archivistica e Documentazione», Guarasci R. (a cura di), vol. 7, Marzi, Cosenza, Comet Editor Press, 2012

FRANCIS, É., QUESNEL, O., *Indéxation collaborative et folksonomies*, in «Documentaliste – Sciences de l'Information», vol. 44, n. 1, 2007, pp. 58-63

GNOLI, C., *Coordinazione, ordine di citazione e livelli integrativi in ambiente digitale*, in «Bibliotime», a. 6, n. 1, marzo 2003 <http://www.spbo.unibo.it/bibliotime/num-vi-1/gnoli.htm>

GNOLI, C., MARINO, V., ROSATI, L., *Organizzare la conoscenza: dalle biblioteche all'architettura dell'informazione per il web*, Milano, Tecniche Nuove, 2006

GROUPE LANGAGES DOCUMENTAIRES DE L'ADBS, *Les normes de conception, gestion et maintenance de thésaurus: évolution récentes et perspectives*, in «Documentaliste-Sciences de l'Information», vol. 44, n. 1, 2007, pp. 66-74

GUERRINI, M., POSSEMATO, T., *Linked data: un nuovo alfabeto del Web Semantico*, in «Biblioteche oggi», aprile 2012, pp. 7-15

HODGE, G., *Systems of Knowledge Organization for Digital libraries. Beyond traditional authority files*, 2000 <http://www.clir.org/pubs/reports/pub91/contents.html>

IFLA, WORKING GROUP ON GUIDELINES FOR MULTILINGUAL THESAURI, *Guidelines for multilingual thesauri*, IFLA, 2005

ISO 25964-1:2011, Information and documentation – *Thesauri and interoperability with other vocabularies*, Part 1: *Thesauri for information retrieval*

ISO 25964-2:2013, Information and Documentation – *Thesauri and interoperability with other vocabularies*, Part 2: *Interoperability with other vocabularies*

ISO 2788:1986, Documentation – *Guidelines for the establishment and development of monolingual thesauri*

ISO 5964:1985, Documentation – *Guidelines for the establishment and development of multilingual thesauri*

IYER, H., *Classificatory Structures: Concepts, Relations and Representation*, Würzburg, Ergon Verlag, 2012

LEI ZENG, M., SALABA, A., *Toward an International Sharing and Use of Subject Authority Data*, FRBR Workshop, OCLC, 2005

LENCI, A., MONTEMAGNI, S., PIRRELLI, V., *Testo e computer: Elementi di Linguistica Computazionale*, Roma, Carocci Editore, 2005

MARINO, V., *Classificazioni per il Web. I vantaggi dell'adozione di schemi a faccette*, Associazione Italiana Biblioteche (AIB) - WEB, 2004 <http://www.aib.it/aib/contr/marino1.htm>

MEDELYAN, O., WITTEN, I. H., *Thesaurus Based Authomatic Keyphase Indexing*, in Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, Chapel Hill, NC, USA, ACM, 2006, pp. 296-297

RANGANATHAN, S.R., *Colon Classification, I: Schedules for Classification*, ed. 7, Gopinath M.A. (a cura di), Sarada Ranganathan Endowment for Library Science, 1989, (ed. 1, 1933)

RAYBURN, G., *Taxonomies and Thesauri*, 2011 <http://www.llrx.com/system/files?file=taxonomiesthesauri.pdf>

RIEDIGER, H., *Cos'è la terminologia e come si fa un glossario*, 2012 <http://www.term-minator.it/corso/doc/mod3_termino_glossa.pdf>

ROSATI, L., *La classificazione a faccette fra Knowledge Management et Information Architecture (parte I)*, It Consult, 2003 <http://www.itconsult.it/knowledge/articoli/pdf/itc_rosati_faccette_e_KM.pdf>

ROSENFELD, L., MORVILLE, P., *Information Architecture for the World Wide Web*, ed. 2, O'Reilly, 2002

SERRAI, A., *Le classificazioni: idee e materiali per una teoria e per una storia*, Firenze, Leo S. Olschki Editore, 1970

SINCLAIR, J., *Trust the text: language, corpus and discourse*, Londra, Routledge, 2004

SOERGEL, D., LAUSER, B., LIANG, A., FISSEHA, F., KEIZER, J., KATZ, S., *Reen-*

gineering Thesauri for New Applications: the AGROVOC Example, in «Digit. Inf.», vol. 4, n. 4, 2004

SPINELLI, S., *Introduzione ai thesauri*, 2005 <http://biocfarm.unibo.it/~spinelli/indicizzazione/thesauri.htm>

SPINELLI, S., *Introduzione all'indicizzazione*, 2006 <http://biocfarm.unibo.it/~spinelli/indicizzazione/>

SPITERI, L.F., *The Essential Element of Faceted Thesauri*, in «Cataloging & Classification Quarterly», vol. 28, n. 4, The Haworth Press, Inc, 1999, pp. 31-52

TAVERNITI, M., *Fra terminologia e documentazione: estrazione automatica di voci indice da corpora documentali della Pubblica Amministrazione*, in «AIDAinformazioni», a. XXVI gennaio-giugno, n. 1-2, 2008, pp. 239-250

TUDHOPE, D., BINDING, C., *Faceted Thesauri*, in «Axiomathes», vol. 18, n. 2, giugno 2008, pp. 211-222

UNI ISO 5963:1985, Documentazione - *Metodi per l'analisi dei documenti, la determinazione del loro soggetto e la selezione dei termini di indicizzazione*, 1985

ZACKLAD, M., *Classification, thésaurus, ontologies, folksonomies : comparaison du point de vue de la recherche ouverte d'information (ROI)*, in CAIS/ACSI 2007, 35e Congrès annuel de l'Association Canadienne des Sciences de l'Information.Partage de l'information dans un monde fragmenté: Franchir les frontières, Montréal, 10-12 maggio 2007, Arsenault C., Dalkir, K. (a cura di), 2007

Sitografia

<http://adlib.alzheimers.org/adear/alzdb/thesaurus.aspx>
<http://aims.fao.org/agrotagger>
<http://aims.fao.org/standards/agrovoc/about>
<http://aims.fao.org/standards/agrovoc/about>
<http://data.nasa.gov/nasa-thesaurus/>
<http://databases.unesco.org/thesaurus/>
<http://dublincore.org/>
<http://eurovoc.europa.eu/drupal/?q=it>
<http://thesaurus.com/Roget-Alpha-Index.html>
<http://wordnet.princeton.edu/>
<http://www.csa.com/factsheets/supplements/LISAguide.pdf>
<http://www.freethesaurus.info/redined/en/index.php>
<http://www.getty.edu/research/tools/vocabularies/aat/index.html>
<http://www.iso.org/iso/home.html>

<http://www.nlm.nih.gov/mesh/>
<http://www.w3.org/>
<http://www.w3.org/2004/02/skos/>
<http://www.w3.org/RDF/>
<http://www.w3.org/TR/owl-features/>
<http://www.w3.org/XML/>
<http://zthes.z3950.org/>

Le ontologie

VINCENZO LOIA*

1 Le ontologie

Sebbene la natura dinamica, decentralizzata ed eterogenea dei servizi e dell'informazione favorisca la visione del Web quale ambiente consono allo sviluppo ed al consolidamento di attività dinamiche, distribuite e cooperanti, la disponibilità di informazioni orientate alla sola comprensione umana, la mancanza di interoperabilità semantica sia a livello di risorse che di applicazioni diffuse sul Web inibiscono l'effettiva fruibilità di tali vantaggi compromettendo la stessa accessibilità delle risorse.

In accordo alla visione del World Wide Web Consortium (W3C)[1], una maggiore fruibilità del Web si raggiunge solo garantendo la disponibilità di informazioni *machine-understandable* (i.e., accessibili agli elaboratori). Dunque, le iniziative in merito al Semantic Web muovono dalla constatazione che le principali limitazioni in termini di fruibilità del Web risiedono principalmente nell'assenza di un adeguato supporto semantico alla rappresentazione delle informazioni disponibili. Ciò spinge ver-

* Università degli Studi di Salerno, Dipartimento di matematica e informatica.

1 Il W3C (<www.w3.org/>) sviluppa tecnologie che garantiscono l'interoperabilità (specifiche, guidelines, software e applicazioni) per portare il World Wide Web al massimo del suo potenziale agendo da forum di informazioni, comunicazioni e attività comuni.

so la diffusione di standard specifici, con regole ben definite, basate su una maggiore strutturazione dei contenuti.

Nel corso degli anni il Web si è arricchito di nuove tecnologie, cosicché la visione originaria quale mero deposito di informazioni lascia il posto ad una concezione del tutto diversa che lo colloca al centro di un'enorme rete di sistemi distribuiti che cooperano scambiandosi reciprocamente non solo informazioni ma funzionalità, servizi. Assistiamo, dunque, alla transizione dal web statico, caratterizzato principalmente dalla disponibilità di ipertesti HTML[2], al web dinamico, in cui i browser si arricchiscono delle potenzialità offerte dai linguaggi di scripting[3] e l'elaborazione lato server è ampiamente supportata dagli *application server*[4] sino all'odierno web computing in cui la visione del Web come piattaforma applicativa trova la sua massima espressione nei Web Services, che permettono la connessione di applicazioni ad altre disponibili e distribuite sulla rete, cui si rivolge l'interesse dell'industria mondiale del software. Lo stesso ruolo degli utenti è cambiato e con il Web 2.0 essi sono divenuti anche autori dei contenuti amplificando la quantità d'informazioni disponibili e ostacolando maggiormente la fruibilità di informazioni di reale interesse.

---

2 HTML (<www.w3.org/html/>), acronimo di *HyperText MarkupLanguage*, è il linguaggio standard per descrivere e definire il contenuto e l'aspetto delle pagine sul World Wide Web.

3 I linguaggi di scripting sono linguaggi di programmazione interpretati, che prevedono, cioè, l'esecuzione riga per riga delle istruzioni contenute in file di codice, detti script appunto, da parte di un programma che funge da interprete. Sono molto utilizzati in ambiente web per permettere lo scambio di informazione tra client e server e si distingue tra scripting lato client e scripting lato server a seconda che l'interprete sia il browser o il server stesso. Lo script può essere un file indipendente o essere incorporato nel codice HTML di una pagina web.
<http://ennebi.solira.org/linprog/pag 13.html>.

4 Piattaforme che forniscono servizi di base richiesti in ambiente web.

Il Semantic Web rappresenta *«un'estensione del web in cui le informazioni sono strutturate con senso compiuto, migliorando la cooperazione tra persone e macchine»* (Berners-Lee et alii, 2001)[5]. L'informazione quindi assume in tal senso una forma ben-definita, *machine-processable* governata da un insieme di linee guida che consentano al Web di acquisire maggiore robustezza, come strumento completo, efficace, utile. Lo sviluppo tecnologico, la diffusione di applicazioni e di sistemi di supporto all'integrazione e all'interoperabilità semantica, costituiscono gli elementi cardine, che negli anni hanno portato alla definizione di uno *stack* di formalismi che trova una corretta raffigurazione nel *Semantic Web Wedding Cake* mostrato in Figura 1. Nello specifico, negli ultimi anni l'affermarsi delle iniziative relative al Semantic Web trova riscontro nell'utilizzo di alcuni standard per la descrizione delle informazioni (i.e., XML[6], RDF[7], etc.), fino all'introduzione delle *ontologie* e di linguaggi ontologici.

Il concetto di ontologia aveva avuto, sino ad allora, quasi esclusivamente un'accezione filosofica che, tuttavia, ben si adattava all'esigenza di strutturare le informazioni scambiate in Rete. Un'ontologia è, per Aristotele, *«lo studio dell'essere in quan-*

---

5 TIM BERNERS-LEE, JAMES HENDLER, ORA LASSILA, *The semantic web*, in «Scientific American», vol. 284, n. 5, 2001, p. 36.

6 *eXtensible Markup Language* (<www.w3.org/XML/>), sviluppato dal W3C, è un linguaggio che consente la rappresentazione di documenti e dati strutturati su supporto digitale.

7 *Resource Description Framework* (<www.w3.org/RDF/>), è uno standard proposto dal W3C come set di linguaggi dichiarativi basato su sintassi XML ed adatto a descrivere la struttura di una parte della realtà. Per realtà intendiamo qualsiasi risorsa sia possibile identificare sulla rete con un indirizzo univoco, mentre per descrizione indichiamo l'insieme delle proprietà, degli attributi e delle relazioni con altre realtà.

to tale»[8], dunque, lo strumento ideale per dare all'uomo la possibilità di stabilire quale sia la *sostanza* che intende modellare e poi interrogare, costruendone la descrizione semantica.

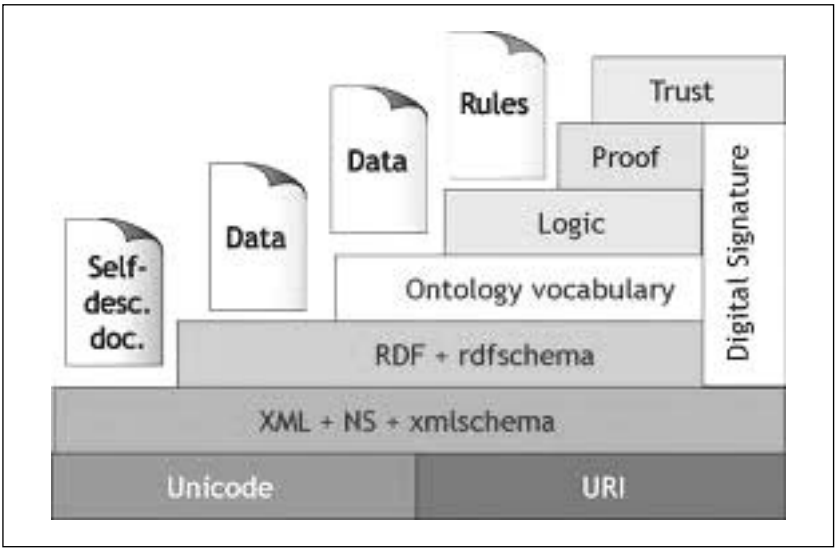


Figura 1. Semantic Web Wedding Cake.

Le ontologie consentono una dichiarazione formale ed esplicita della conoscenza in un particolare dominio applicativo, in modo da poter essere *compresa* sia dalle persone che dai sistemi o agenti software che la condividono. Tuttavia, esiste un'ulteriore ragione che giustifica il forte interesse per questo strumento di modellazione: un guadagno a lungo termine in risorse investite. Infatti, sebbene le tecnologie per la realizzazione dei sistemi

[8] NICOLA ABBAGNANO, *Dizionario di filosofia*, Torino, UTET, 1968, p. 561.

software siano in continua evoluzione, non succede lo stesso con ciò che essi vogliono rappresentare: il dominio. Poter contare su una organizzazione sotto forma di ontologia della realtà che si intende simulare, consente di definire un *linguaggio* comune per uno specifico dominio, per i membri di uno stesso team, etc. Inoltre, un modello concettuale di così lunga durata può essere utilizzato in diverse applicazioni fornendo opportunità di riuso ed interoperabilità. Anche la possibilità di condividere l'ontologia con il proprio cliente diventa un'opportunità allettante in quanto gli consente di capire quanto essa sia rappresentativa del proprio dominio, migliorando l'*appeal* dei prodotti o servizi che si vanno a realizzare, o favorendo eventuali miglioramenti, laddove necessari.

Questa capacità di rappresentazione da parte delle ontologie è diventata, negli anni, il presupposto per la realizzazione di sistemi software anche di natura e obiettivi diversi. Esistono sistemi che sfruttano la modellazione ontologica per migliorare le funzionalità di ricerca (De Maio et alii, 2011)[9] o per utilizzare e valorizzare il patrimonio di conoscenza interno a un'azienda favorendone il ciclo di vita, mentre altri si basano sull'esperienza pregressa nel campo della medicina per formulare nuove diagnosi (De Maio et alii, 2011)[10]. In (Happel, Sedorf, 2006)[11] gli autori

9 Cfr. CARMEN DE MAIO, GIUSEPPE FENZA, VINCENZO LOIA, SABRINA SENATORE, *Hierarchical web resources retrieval by exploiting Fuzzy Formal Concept Analysis*, in «Information Processing & Management», vol. 48, n. 3, Elsevier, 2012, pp. 399-418.

10 Cfr. CARMEN DE MAIO, VINCENZO LOIA, GIUSEPPE FENZA, MARIACRISTINA GALLO, ROBERTO LINCIANO, ALDO MORRONE, *Fuzzy knowledge approach to automatic disease diagnosis*, in 2011 IEEE International Conference on Fuzzy Systems Proceedings, Taipei, 27-30 giugno 2011, pp. 2088-2095.

11 Cfr. HANS-JÖRG HAPPEL, STEFAN SEEDORF, *Applications of Ontologies in Software Engineering*, in Atti del «2nd International Workshop on Semantic Web Enabled Software Engineering (SWESE 2006) - 5th Interna-

mostrano come la modellazione ontologica possa migliorare anche l'ingegneria del software nelle sue diverse fasi di progettazione e sviluppo. Altri scenari d'uso riguardano il discovery di servizi Web.

In generale, le ontologie risultano particolarmente utili per il design di sistemi orientati al problem solving basato sulla conoscenza, come i sistemi per la diagnosi, la pianificazione delle attività etc., grazie al loro potere espressivo ma soprattutto grazie alla possibilità di strutturare un sistema di ragionamento inferenziale.

1.1 *Modellazione Ontologica*

L'obiettivo del Semantic Web è quello di strutturare le risorse per renderle comprensibili alle macchine così come agli umani. A tal fine si ricorre alla modellazione semantica dei suoi contenuti mediante l'utilizzo di particolari linguaggi. Uno dei primi strumenti utilizzati è stato il linguaggio XML.

Un documento XML si struttura in base ad un DTD (*Document Type Definition*)[12] associato che genera una sorta di gerarchia di marcatori simile ad un *albero di Porfirio*: uno schema di coordinazione e subordinazione dei generi e delle specie, prodotto partendoda un genere *sommo* e scendendo fino alle specie più specifiche secondo il processo della dicotomia (e.g., la sostanza può essere corporea e incorporea, quella corporea si divide in animata e inanimata, quella animata in sensibile e insensibile ecc.) (Blum, 1999)[13].

Gli alberi di Porfirio, e quindi anche l'XML, non danno alcu-

tional Semantic Web Conference (ISWC 2006)», Athens, GA, USA, 05-09 novembre 2006.

12 Una DTD definisce la struttura di un documento XML.

13 Cfr. PAUL RICHARD BLUM, *Dio e gli individui: L'«Arbor Porphyriana» nei secoli XVII e XVIII*, in «Rivista di filosofia neo-scolastica», vol. 91, 1999, pp. 18-49.

na informazione riguardo il significato, la semantica dei nodi che definiscono. Inoltre, è impossibile stabilire una sovrapposizione fra i marcatori, ovvero è necessario che vi sia univocità tra i nodi dell'albero. Supponiamo di dover descrivere una realtà composta da sostanza *materiale* e *immateriale* in cui, quella materiale è fatta di *corpi animati* (gli esseri viventi) o *inanimati* (i minerali). Gli *esseri viventi*, a loro volta, possono essere *animali*, *piante* o *uomini*. La *sostanza immateriale*, invece, contempla la *scienza* e l'*arte*; la scienza, a sua volta, si compone di *scoperte scientifiche*. In Figura 2 è mostrata una possibile rappresentazione ad albero di questa realtà.

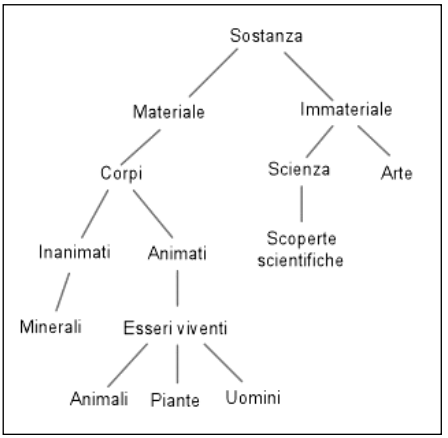


Figura 2. Esempio di albero di classificazione.

Dalla rappresentazione grafica è evidente che risulta impossibile, con il solo strumento ad albero, definire un'eventuale relazione tra gli *uomini* e le *scoperte scientifiche* senza inficiarne la correttezza sintattica (vedi Figura 3 e Figura 4). Nella prima soluzione, mostrata in Figura 3, infatti, sembrerebbe esistere una relazione di subordinazione tra uomini e scoperte scientifiche; la soluzione in Figura 4, invece, non è sintatticamente corretta per la mancanza di univocità tra i nomi dei nodi.

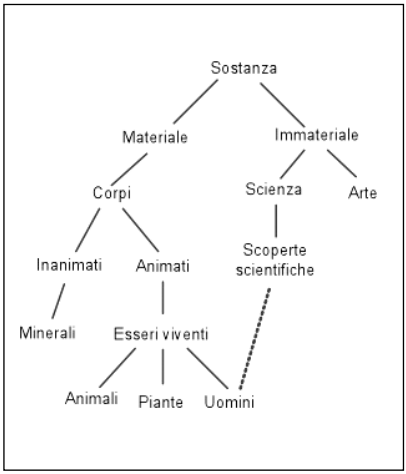


Figura 3. Inserimento di relazioni (a).

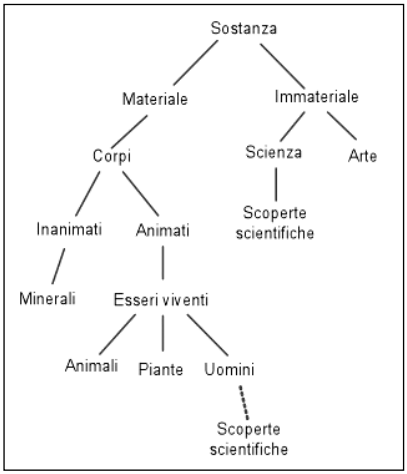


Figura 4. Inserimento di relazioni (b).

Da questa analisi emerge che è assolutamente necessario uno strumento che consenta sia la definizione della semantica dei nodi, sia la definizione di relazioni che vadano al di là dei semplici concetti di appartenenza o subordinazione: il modello logico, del quale fanno parte anche le ontologie.

In una visione generale, le ontologie sono una forma di rappresentazione della conoscenza del mondo o di parte di esso. Nello specifico dei sistemi informatici, esse sono utilizzate per la rappresentazione della conoscenza in una forma *machine understandable* e sono applicabili ad una grande varietà di obiettivi funzionali, tra cui il ragionamento induttivo, la classificazione e l'utilizzo di tecniche di problem solving, oltre che per facilitare la comunicazione e lo scambio di dati fra sistemi diversi. Le ontologie, pertanto, sono legate a vocabolari *controllati*, in base ai quali i concetti devono essere descritti.

I componenti più comuni definitiall'interno di un'ontologia sono: individui, classi, attributi, relazioni, termini funzione, restrizioni, regole, assiomi e eventi.

Gli *Individui* sono le istanze o gli oggetti (detti anche *livello base*); le *Classi* sono gli insiemi, le collezioni, i concetti, i tipi di oggetto, i generi di cose; gli *Attributi* sono gli aspetti, le proprietà, le caratteristiche, i caratteri o i parametri che l'oggetto o la classe può avere; le *Relazioni* sono i modi in cui le classi e gli individui sono collegati tra loro; i *Termini funzione* sono strutture complesse formate da certe relazioni che possono essere usate al posto di un termine individuale in una dichiarazione; le *Restrizioni* sono descrizioni formali che le asserzioni devono rispettare; le *Regole* sono espressioni, scritte nella forma IF-THEN, che descrivono l'inferenza logica che deve essere ottenuta da una espressione in una forma particolare; gli *Assiomi* sono le affermazioni e le regole espresse in una forma logica tale che, se prese nel loro insieme, rappresentano tutta la realtà che l'ontologia descrive nel suo dominio di applicazione. Gli *Eventi* rappresentano il cambiamento di un attributo o di una relazione.

Le ontologie sono comunemente codificate usando linguaggi ontologici che possono essere sia proprietari che basati su standard emergenti nell'ambito del Semantic Web. Tali linguaggi si differenziano per potenzialità espressive offerte a supporto della definizione concettuale di un'ontologia. Il più affermato è sicuramente OWL[14] (Capuano, 2005)[15] acronimo di *Web Ontology Language*. Si tratta di un linguaggio basato sul modello assertivo di RDF (*Resource Description Framework*) (i.e., Soggetto, Predicato, Oggetto) che introduce la terminologia standard per poter definire i componenti di un'ontologia. OWL è stato ideato per un utilizzo nel World Wide Web; infatti, tutti i

[14] *Web Ontology Language* (<http://www.w3.org/TR/owl-features/>) è un linguaggio per definire e istanziare Ontologie Web, attraverso il quale si aggiunge semantica ai documenti reperibili in rete.

[15] Cfr. NICOLA CAPUANO, *Ontologie OWL: Teoria e Pratica*, in «Computer Programming» nn. 148, 149, 150, luglio/agosto, settembre, ottobre 2005.

suoi elementi sono definiti in RDF come risorse ed identificati tramite URI.

L'OWL è diviso in tre sottolinguaggi a potenza espressiva crescente, ognuno dei quali è designato per un uso specifico: OWL Lite, OWL DL e OWL Full[16]. *OWL Lite* serve come supporto per quegli utenti che hanno bisogno di rappresentare classificazioni gerarchiche e vincoli semplici. Esso fornisce anche una migrazione veloce di thesauri e di altre tassonomie; *OWL DL* offre massima espressività mantenendo completezza computazionale (tutte le conclusioni sono garantite essere computabili) e decidibilità (tutte le elaborazioni terminano in un tempo finito). L'aggettivo DL è dovuto alla corrispondenza del linguaggio con la *Description Logic* (logica descrittiva), un campo della ricerca che ha studiato le logiche che formano la base formale dell'OWL. Infine, *OWL Full* è stato sviluppato per quegli usi che necessitano della massima espressività e della massima libertà sintattica dell'RDF, ciò però non dà garanzie computazionali. L'OWL Full permette alle ontologie di incrementare il significato di vocabolari predefiniti, appartenenti sia all'RDF che all'OWL.

Inoltre, un linguaggio ontologico per il Web Semantico derivante da OWL è OWL2[17]. Il linguaggio fornisce classi, proprietà e individui, immagazzinando il tutto come documenti per il Semantic Web. Essenzialmente, rappresenta una sorta di estensione di RDFS (RDF-Schema), aggiungendone costrutti ed estendendone l'espressività.

---

16 Sono state definite 3 versioni di OWL, con potere espressivo e complessità crescenti: (i) OWL Lite – sintatticamente più semplice; (ii) OWL DL (description logics) – versione intermedia con potere espressivo più elevato (mantiene la completezza computazionale e la decidibilità); (iii) OWL Full – massima espressività con nessuna garanzia di completezza e decidibilità.

17 <http://www.w3.org/TR/owl2-overview/>.

Essendo le ontologie una rappresentazione del dominio di interesse, esse vanno costruite seguendo un processo di modellazione come quello illustrato in Figura 5, che tenga conto dei termini e dei concetti significativi per il dominio stesso.

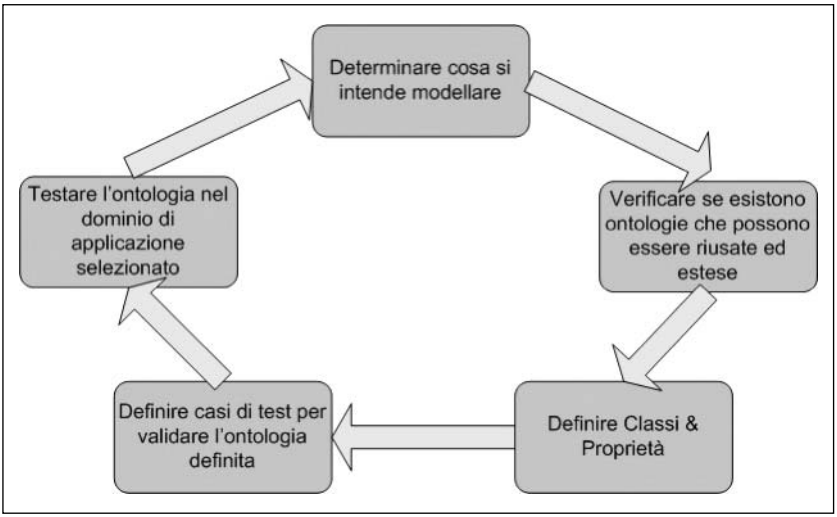


Figura 5. Processo di modellazione di schemi ontologici.

Il processo si propone di selezionare l'insieme delle tecnologie di riferimento a supporto della rappresentazione della conoscenza, ma anche di individuare eventuali modelli di conoscenza esistenti ed accettati dalle comunità che operano nell'area delle tecnologie semantiche, per incentivare la politica del riuso.

I passi fondamentali per il processo di modellazione ontologica sono i seguenti:

Acquisire la conoscenza del dominio: raccogliere quante più informazioni possibili sul dominio di interesse, comprendere i termini usati formalmente per descriverne le entità in maniera consistente, in collaborazione con gli esperti del dominio. Le domande da porsi sono pressoché le seguenti: (i) quale dominio co-

prirà l'ontologia? (ii) qual è lo scopo dell'ontologia? (iii) a quali tipi di domande l'informazione espressa dall'ontologia può fornire risposte? (iv) chi userà e chi sarà il responsabile della manutenzione dell'ontologia?

Riconsiderando l'esempio precedente, abbiamo un dominio formato dai seguenti concetti fondamentali: la sostanza materiale e immateriale; la sostanza materiale è fatta di *corpi animati* (gli esseri viventi) o *inanimati* (i minerali). Gli *esseri viventi*, a loro volta, possono essere *animali, piante* o *uomini*. La *sostanza immateriale*, invece, si differenzia in *scienza* o *arte*; la scienza si compone di *scoperte scientifiche*.

Organizzare l'ontologia: progettare la struttura concettuale complessiva del dominio. Questa operazione semplifica l'identificazione dei principali concetti del dominio e delle loro proprietà, stabilendo le relazioni tra i concetti, creando concetti astratti, individuando quali di questi hanno delle istanze etc. Le domande da porsi in questa fase sono: (i) quali sono i termini importanti? (ii) quali sono le proprietà? Vi sono tre passi fondamentali: (1) *flat glossary*, che consiste nel documentare ciascun termine con una definizione in linguaggio naturale (e fornire esempi dove appropriato) in cui i nomi diventano oggetti o attori, e i verbi si trasformano in relazioni o processi, (2) *structured glossary*, che consiste nella decomposizione e/o specializzazione dei termini e nell'individuazione degli attributi di un concetto (predicazione), (3) identificare tutte le relazioni concettuali fra gli oggetti.

Per l'esempio iniziale abbiamo i seguenti risultati:

(1) Definizione in linguaggio naturale dei seguenti termini: Sostanza materiale, Sostanza immateriale, Corpo animato, Corpo inanimato, Minerale, Essere Vivente, Animale, Pianta, Uomo, Arte, Scienza,Scoperta scientifica;
(2) Definizione delle specializzazioni *animali*, *piante* e *uomini* per l'entità *esseri viventi*. Inoltre, si vanno ad identificare gli attributi utili alla descrizione delle varie entità

(e.g., nome e specie per animale; forma e colore per minerali ecc.);

(3) Le relazioni concettuali identificate sono:
   – La scienza si compone di scoperte scientifiche;
   – Gli uomini fanno scoperte scientifiche.

Popolare l'ontologia: aggiungere concetti, relazioni ed entità fino a raggiungere il livello di dettaglio necessario a soddisfare gli obiettivi dell'ontologia. Per individuare nuovi concetti è possibile adottare tre tipi di approcci: (i) *top-down*, che prevede l'identificazione dei concetti generali e attraverso un raffinamento successivo si procede verso i concetti particolari (e.g., da computer a workstation), (ii) *bottom-up*, che procede per livelli di astrazione, partendo dalle entità particolari del dominio per astrarre i concetti generali che racchiudono o fanno uso di quelli particolari (da workstation a computer) e (iii) *middle-out* (o combinato) che prevede di individuare prima i concetti salienti e poi generalizzarli e specializzarli. I concetti da soli non forniscono informazioni sufficienti? è importante definire anche le relazioni tra gli oggetti del dominio. Inoltre è consigliabile imporre dei vincoli sulle relazioni, quali la cardinalità (ad esempio un computer ha [1,n] processori) sul dominio e sul codominio della relazione[18], sul tipo di valore (e.g., un indirizzo ha un CAP il cui codominio è di tipo stringa, una bottiglia ha un'etichetta con un codominio di tipo istanza, un'idea coinvolge concetti con un codominio di tipo concetto).

Applicazioni di questo step sono, ad esempio, l'aggiunta di ulteriori concetti per la distinzione tra piante sempreverdi e non; in più, l'aggiunta delle cardinalità per le relazioni già definite:

– La scienza si compone di scoperte scientifiche: [1, n];
– Gli uomini fanno scoperte scientifiche: [n,m].

[18] Per dominio si intende la classe alla quale si applica la relazione, mentre per codominio la classe alla quale la relazione stessa si riferisce.

Controllare il proprio lavoro: risolvere inconsistenze sintattiche, logiche e semantiche tra gli elementi dell'ontologia. I controlli di consistenza possono anche favorire una classificazione automatica che definisce nuovi concetti sulla base delle proprietà delle entità e delle relazioni tra le classi. Per i controlli di consistenza possono essere utilizzati in maniera efficiente i tool di verifica automatici messi a disposizione da alcuni strumenti di disegno come Protégé[19].

Considerare il riuso di risorse esistenti: è sempre utile pensare di ridefinire ed estendere risorse esistenti, quali glossari, dizionari dei termini e dei sinonimi, documenti, standard e altre ontologie.

Consegnare l'ontologia: al termine dello sviluppo di un'ontologia, così come per qualunque altro sviluppo software, è necessaria una verifica da parte degli esperti del dominio.

Unitamente al processo di modellazione, nella composizione delle ontologie è bene sottostare anche alle seguenti regole: evitare sovrapposizioni, cioè non assegnare uno stesso nome ad un concetto e ad una relazione; aggiungere capitalizzazioni e delimitatori (*dot-notation*, trattino); utilizzare prefissi e suffissi per le relazioni (ha-produttore, produttore-di); distinguere gli oggetti dai processi; non usare abbreviazioni; l'albero dell'ontologia deve essere bilanciato nella granularità; evitare classi con un unico figlio.

Pertanto, un'ontologia per essere definita *buona* deve possedere le seguenti caratteristiche:prevedere tutte le distinzioni chiave; non fare assunzioni implicite; essere chiara e concisa: ciascun concetto deve essere rilevante e non vi devono essere duplicati; essere coerente: avere tutte e sole le inferenze consistenti con le definizioni dei concetti; richiedere scelte di progetto motivate (*design options*); aderire ad un *ontological commitment*

19 <http://protege.stanford.edu/>.

(accordo sull'uso di un *vocabolario* consistente con il dominio di interesse) essere modificabile; essere riusabile.

Il risultato di questo processo è un'ontologia espressa in linguaggio RDF come quella che segue:

```
<rdf:RDF>
<owl:ObjectPropertyrdf:about=”#fanno”>
<rdfs:rangerdf:resource=”#Scoperte_scientifiche”/>
<rdfs:domainrdf:resource=”#Uomini”/>
</owl:ObjectProperty>

<owl:ObjectPropertyrdf:about=”#si-compone-di”>
<rdfs:domainrdf:resource=”#Scienza”/>
<rdfs:rangerdf:resource=”#Scoperte_scientifiche”/>
</owl:ObjectProperty>

<owl:Classrdf:about=”#Animali”>
<rdfs:subClassOfrdf:resource=”#Esseri_viventi”/>
</owl:Class>

<owl:Classrdf:about=”#Animati”>
<rdfs:subClassOfrdf:resource=”#Corpi”/>
</owl:Class>

<owl:Classrdf:about=”#Arte”>
<rdfs:subClassOfrdf:resource=”#Immateriale”/>
</owl:Class>

<owl:Classrdf:about=”#Corpi”>
<rdfs:subClassOfrdf:resource=”#Materiale”/>
</owl:Class>

<owl:Classrdf:about=”#Esseri_viventi”>
<rdfs:subClassOfrdf:resource=”#Animati”/>
</owl:Class>
```

```
<owl:Classrdf:about=”#Immateriale”>
<rdfs:subClassOfrdf:resource=”#Sostanza”/>
</owl:Class>

<owl:Classrdf:about=”#Inanimati”>
<rdfs:subClassOfrdf:resource=”#Corpi”/>
</owl:Class>

<owl:Classrdf:about=”#Materiale”>
<rdfs:subClassOfrdf:resource=”#Sostanza”/>
</owl:Class>

<owl:Classrdf:about=”#Minerali”>
<rdfs:subClassOfrdf:resource=”#Inanimati”/>
</owl:Class>

<owl:Classrdf:about=”#Piante”>
<rdfs:subClassOfrdf:resource=”#Esseri_viventi”/>
</owl:Class>

<owl:Classrdf:about=”#Scienza”>
<rdfs:subClassOfrdf:resource=”#Immateriale”/>
</owl:Class>

<owl:Classrdf:about=”#Scoperte_scientifiche”/>

<owl:Classrdf:about=”#Sostanza”/>

<owl:Classrdf:about=”#Uomini”>
<rdfs:subClassOfrdf:resource=”#Esseri_viventi”/>
</owl:Class>
</rdf:RDF>
```

1.2 *Tipi di Ontologie*

I modelli ontologici sono generalmente strutture gerarchiche complesse in cui si crea una fitta rete di relazioni tra i concetti, solitamente divise in tipologiedifferenti in base al loro livello di astrazione ed in base agli obiettivi applicativi: ontologie di *dominio*, ontologie di *riferimento* e ontologie *fondamentali*.

Un'ontologia di dominio (detta anche *Domain-specific Ontology*) modella una parte del mondo: essa rappresenta il significato specifico dei termini e come questi sono utilizzati in quel dominio. Esempi di tali ontologie sono quelle per la georeferenziazione (e.g., Geonames[20]), e le ontologie per la classificazione dei prodotti (e.g., UNSPC[21]).

Un'ontologia di riferimento (detta *Reference Ontology*) modella le entità in accordo a specifici schemi e obiettivi. Si dicono *di riferimento* in quanto sono referenziate dallo stato dell'arte.

Un'ontologia fondamentale (detta anche fondazionale o superiore o *UpperOntology*) è un modello degli oggetti comuni che sono, di solito, utilizzati trasversalmente in un ampio campo di domini ontologici. Essa contiene un glossario dei termini nel quale gli oggetti in un insieme di domini possono essere descritti. Per esempio, tali ontologie sono utili per rappresentare le persone e i loro interessi (e.g., FOAF[22]), per la descrizione delle caratteristiche delle folksonomie (e.g., SKOT[23]), dei social network e delle attività (e.g., SIOC[24]) ecc.

Al fine di chiarire la differenza fra i vari modelli ontologici, di seguito saranno forniti dettagli ed esempi di alcune *upper ontology* e *domain ontology* esistenti in letteratura.

---

20 <www.geonames.org/>.
21 <www.unspsc.org/>.
22 *Friend-of-A-Friend.* <http://www.foaf-project.org/>.
23 *Social Semantic Cloud of Tags.* <http://www.scot-project.org/>.
24 *Semantically-Interlinked Online Communities.* <http://sioc-project.org/>.

1.2.1 *'Upper Ontology'*

Le *upper ontology* (LOA, 2005)[25] rappresentano concetti generali indipendenti dai singoli domini scientifici e applicativi.

Fondamentalmente, sono ontologie indipendenti dal dominio con particolari caratteristiche di flessibilità, estendibilità e riusabilità. Un'upper ontology è limitata alla definizione di meta-concetti generici, astratti, e quindi abbastanza generali da indirizzare un ampio range di aree di dominio. I concetti di dominio non sono definiti all'interno di tale tipologia di ontologie, ma tali artefatti forniscono una struttura su cui costruire e integrare ontologie di dominio specifiche.

Esempi di upper ontology sono descritti di seguito.

OWL-S[26] (*Web Ontology Language for Services*) è lo standard più noto per la definizione di servizi web semantici: si pone l'obiettivo di fornire i meccanismi base per definire una codifica più ricca di semantica ed espressività, al fine di consentire l'automatizzazione del discovery, dell'invocazione, della composizione e dell'interoperazione tra i servizi web.

L'OWL-S è una particolare upper ontology basata su OWL che fornisce i concetti e le relazioni necessari per descrivere le *service capability* generali. In particolare, come evidenziato in Figura 6, si basa su tre livelli descrittivi: *Service Profile, Service Model e Service Grounding*. *ServiceProfile* fornisce una descrizione concisa del servizio (e.g., cosa fa, come lavora ecc.); *ServiceModel* descrive le funzionalità del servizio; *Service Grounding* decrive come accedere al servizio e stabilisce una corrispondenza tra frammenti di OWL-S e frammenti di WSDL[27].

---

[25] Cfr. LABORATORY FOR APPLIED ONTOLOGY (LOA), Institute of Cognitive Science and Technology, *Foundational Ontologies & Their Library*, 2005, pp. 32-34.

[26] *OWL-Services.* <http://www.w3.org/Submission/OWL-S/>.

[27] *Web Services Description Language.* <http://www.w3.org/TR/wsdl>.

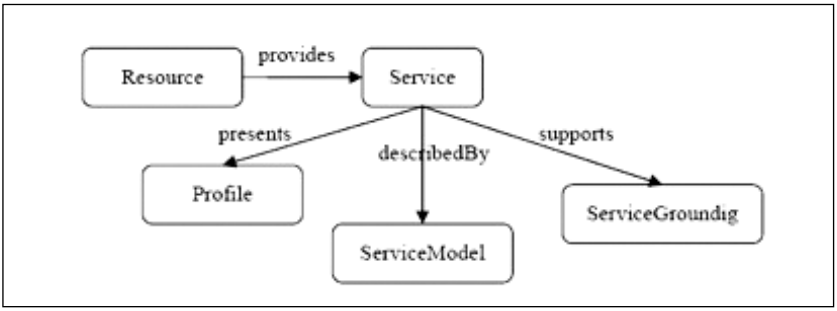


Figura 6. Ontologia OWL-S.

Il servizio è definito in termini dei suoi processi costituenti (*Atomic Process, Simple Process, Composite Process*), ognuno dei quali viene poi definito mediante una serie di IOPE (*Input Output Preconditions Effects*), e gli stessi rapporti tra processi sono specificati in maniera rigorosa. OWL-S assicura una descrizione dei servizi disposta su tre livelli. Il livello top dell'ontologia descrive come una risorsa è correlata ad un servizio, successivamente come il servizio è relato al *profile*, al *Service Model* e al *Service Grounding*. I tre livelli concettuali di OWL-S sono progettati per fornire insieme un quadro complessivo e generale delle *capabilities* di un servizio.

Il progetto *FOAF* (*Friend Of A Friend*)è un'applicazione del Semantic Web che può essere utilizzata per descrivere le persone, le loro attività e le relazioni con altre persone. In particolare, FOAF è stato progettato per permettere di strutturare le informazioni relative alle persone di una rete sociale in modo *machine-readable*.

Fondato da Dan Brickley e Libby Miller, il progetto FOAF mira a definire un vocabolario standard, denominato *FOAF Vocabulary Specification*[28], basato su RDF per esprimere i metada-

[28] <http://xmlns.com/foaf/spec/>.

ti riguardo le persone, i loro interessi, i rapporti e le attività. Le definizioni del vocabolario FOAF sono state scritte utilizzando linguaggi di programmazione RDF e OWL; attraverso i termini presenti nel vocabolario è possibile creare facilmente documenti FOAF.

Un documento FOAF, che tuttavia è diverso da una pagina Web tradizionale, può essere combinato con altri documenti della medesima natura per generare un'unica base di dati di informazioni. Ciò consente ad un software di elaborare queste descrizioni, magari all'interno di un motore di ricerca, allo scopo di trovare informazioni su una persona e sulle comunità delle quali fa parte.

Le principali classi e proprietà dell'ontologia FOAF, sono: *Core*, che contiene le classi e le proprietà che costituiscono il nucleo di FOAF e che descrivono le caratteristiche di persone e gruppi sociali che sono indipendenti dal tempo e dalla tecnologia, *Social Web*, che contiene una serie di classi utilizzate per descrivere account Internet, indirizzi Web e altre attività sul Web. Inoltre, FOAF definisce le classi per descrivere altre entità quali: Progetto, Organizzazione e Gruppo.

SIOC (*Semantically-Interlinked Online Communities*), è un'iniziativa per la rappresentazione semantica delle informazioni prodotte dalle community on-line: descrive nel linguaggio RDF i contenuti generati dagli utenti sui siti come blog, forum, wiki e social network, e le interazioni che hanno suscitato (Bojars et alii, 2005)[29].

Il cuore del framework SIOC è l'omonima ontologia, definita per la rappresentazione di dati provenienti dal social Web in RDF. L'ontologia è costituita da un insieme di classi e proprietà

[29] Cfr. ULDIS BOJARS, ALEXANDRE PASSANT, JOHN G. BRESLIN, STEFAN DECKER, *The SIOC Project: Semantically-Interlinked Online Communities*, in COIN@AAMAS&IJCAI&MALLOW, vol. 6069, Springer, 2009, pp. 179-194.

quali: *Site* definisce il luogo di una community online o di un insieme di community; *Forum* è uno spazio di discussione, ospitato su un sito; *Post* può essere costituito da un articolo, un messaggio, un audio o un video. Un post è scritto da un *autore*, ha un *topic* specifico, un *contenuto*, *link esterni*, ecc.; *User* rappresenta l'account di un membro della community on-line; *Usergroup* è un insieme di account di utenti interessati ad un argomento comune.

SIOC consente agli sviluppatori di collegare gli elementi creati dagli utenti ad altri oggetti relazionati ad essi, alle persone (tramite il loro account utente associato) e ad argomenti (utilizzando tag specifici o categorie gerarchiche), come rappresentato in Figura 7 e può rappresentare vari tipi di contenitori (ossia Wiki, blog, MessageBoard) e vari *content item* (cioè WikiArticle, blogpost, BoardPost).

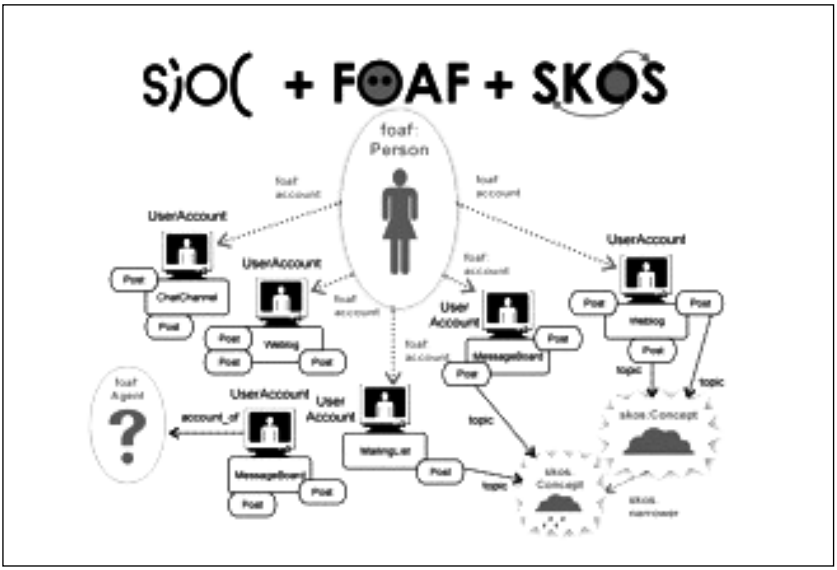


Figura 7. InterlinkSIOC,FOAFeSKOS.

SCOT (*Social Semantic Cloud of Tags*) è un'ontologia che mira alla descrizione delle caratteristiche delle folksonomie. In particolare, ha lo scopo di descrivere la struttura e la semantica dei dati taggati e le loro relazioni in modo esplicito attraverso RDF/OWL e consentire l'interoperabilità sociale per la condivisione e il riutilizzo di metadati dei tag semantici attraverso diverse fonti. Infatti, il modello descrive le proprietà dei tag, comprese le loro frequenze di occorrenza, quali altri tag sono utilizzati in combinazione con essi e quali tag sono relati ad essi.

SCOT incorpora e riutilizza i vocabolari esistenti (e.g., SIOC, FOAF, SKOS ecc.) al fine di evitare ridondanze e per consentire l'utilizzo di una descrizione dei metadati più ricca per specifici domini (e.g., per descrivere un insieme di utenti, risorse ecc.). Durante il processo di progettazione dell'ontologia SCOT, alcune classi e proprietà esterne sono state identificate come adatte ad essere riutilizzate. Tali concetti non sono inclusi all'interno di SCOT ma sono utilizzati unitamente ai suoi termini per descrivere le attività di tagging (vedi Tabella1). Questi sono rappresentati da un insieme di classi e proprietà RDF che possono essere utilizzate per esprimere il contenuto e la struttura dell'attività di tagging come un grafo RDF.

Prefisso	***Specifica***
tag	Newman's Tag Ontology
dcterms	Dublin Core MetadataTerms
foaf	FriendofaFriend(FOAF)
sioc	SemanticallyInterlinkedOnlineCommunity(SIOC)
skos	SimpleKnowledgeOrganizationSystem(SKOS)
xsd	XMLSchema(Datatypes)

Tabella 1. Classi e proprietà esterne appartenenti a vocabolari esistenti.

1.2.2 *Domain Ontology*

Le ontologie di dominio (Corcho et alii, 2003)[30] modellano i concetti e le relazioni rilevanti per i singoli campi disciplinari, affrontando la conoscenza dettagliata sulla quale esiste già sostanzialmente una condivisione dei significati e ampia convergenza tra gli studiosi. Ad un livello di astrazione inferiore rispetto alle *domain ontology* troviamo le *task-ontology*, che formalizzano i concetti relativi ad una specifica attività; e le *application ontology*, che coprono le attività relative ad un particolare campo scientifico e richiedono un livello molto elevato di condivisione della conoscenza all'interno del campo applicativo.

Esempi di *domain ontology* sono descritti di seguito.

Geonames è un progetto gratuito per la creazione di un database geografico mondiale. Ha l'obiettivo di fornire gli strumenti per tradurre il nome di un luogo (e.g.,una montagna, una città, una regione) in dati che lo rappresentino: latitudine, longitudine, altezza, popolazione, CAP ecc.

Altri progetti di *geocoding* sono regolati da norme e specifiche restrizioni, Geonames, invece, è distribuito con licenza *Creative Commons*: chiunque è libero di utilizzare come preferisce l'enorme mole di informazioni disponibile, purché ne venga citata la fonte. Integrare la geografia nel Web semantico diventa così un'operazione alla portata di qualunque webmaster.

Oltre alla disponibilità dei nomi di luoghi in lingue diverse, sono presenti dati come latitudine, longitudine, altitudine, popolazione, suddivisione amministrativa e codici postali. Tutte le coordinate utilizzano il sistema WGS84 (*World Geodetic System*,

---

30 Cfr. OSCAR CORCHO, MARIANO FERNANDEZ-LOPEZ, ASUNCION GOMEZ-PEREZ, *Methodologies, Tools and Languages for Building Ontologies: Where is the Meeting Point?*, in «Data & Knowledge Engineering», vol. 46, 2003, pp. 41-64.
<http://www.dia.fi.upm.es/~ocorcho/documents/DKE2003_CorchoEtAl.pdf>.

standard per l'utilizzo in cartografia, geodesia e navigazione). I nomi degli Stati vengono codificati in adesione allo standard ISO 3166[31].

Le *feature* di Geonames prevedono una categorizzazione basata sulla selezione di una classe scelta da una tassonomia disponibile nella stessa ontologia. Ad esempio, una *feature* geografica relativa ad un hotel sarà etichettata nella base dati con la classe tassonomica che identifica gli edifici (i.e., classe S) e nello specifico dal codice che individua un hotel (i.e., S.HTL), in modo da differenziare l'annotazione geografica per l'hotel da tutte le altre.

UNSPSC è il più noto standard di classificazione che, con la sua gerarchia di codifica, consente di definire circa 15.000 categorie. È di pubblico dominio e non è soggetto a copyright.

Si tratta di uno schema di classificazione di *commodity* e non di prodotti e include solo gli attributi primari usati per differenziare una *commodity* da un'altra (attributi come marchi o caratteristiche fisiche sono normalmente esclusi).

Il primo obiettivo della classificazione gerarchica UNSPSC è di evitare le duplicazioni e fornire uno strumento che consenta, a chi è responsabile della classificazione dei prodotti o servizi, di identificare facilmente il codice appropriato (mediante l'utilizzo di parole chiave).

La caratteristica più importante dello schema UNSPSC è che una *commodity* deve essere individuata da un titolo e da una definizione univoci e facilmente distinguibili da tutti gli altri.

1.3 Vantaggi legati alla Modellazione Ontologica

Come già ampiamente argomentato, le tecnologie semantiche e in particolar modo le ontologie forniscono uno strumento ca-

[31] ISO 3166, *Codici per la rappresentazione di nomi di paesi e delle loro suddivisioni.*

pace di associare a qualunque tipo di risorsa una semantica in termini di concetti descritti, di gerarchie, di relazioni ecc.

Sebbene i promotori dell'artefatto ontologico erano fermamente convinti del fatto che i vantaggi offerti fossero di gran lunga superiori alle difficoltà dovute alla gestione delle ontologie, era necessario convicere anche tutta la comunità scientifica. In questo senso (Gruninger, Lee, 2002)[32] hanno proposto una classificazione ad alto livello dei benefici offerti dalle ontologie. Le classi individuate sono: (i) *Comunicazione* - le ontologie assicurano interoperabilità tra sistemi (e uomini) a livello di dati e di processi, identificando in maniera univoca il significato dei concetti in un dominio; (ii) *Inferenza computazionale* - le ontologie consentono di derivare in maniera automatica informazioni implicite non direttamente presenti nella base di conoscenza per migliorare le tradizionali tecnologie di *browsing* e *retrieval* delle informazioni. È inoltre possibile modellare la conoscenza di dominio indipendentemente dall'implementazione del sistema sottostante; (iii) *Riuso ed organizzazione della conoscenza* - sviluppando descrizioni di dominio sistematiche e ampiamente accettate se ne abilita il riuso evitando i costi per eventuali nuovi sviluppi e si migliora la qualità dei modelli forniti.

In letteratura sono numerose le applicazioni che sfruttano questi benefici. Le tecnologie semantiche, infatti, hanno permesso la progettazione di funzionalità che consentono un'ottimizzazione dell'accesso all'informazione attraverso una descrizione formale in ontologie delle risorse da recuperare, indipendentemente dalla loro tipologia (e.g., documenti testuali, immagini, video ecc.). L'applicazione della semantica consente in questo modo la progettazione di funzionalità innovative che permettono

[32] Cfr. MICHAEL GRUNINGER, JINTAE LEE, *Introduction – ontology: different ways of representing the same concept*, Communication of the ACM, vol. 45, n. 2, 2002, pp. 39-41.

una classificazione multidimensionale delle risorse. In tal senso risulta interessante lo schema noto come a *faccette* (o *faceted*) proposto da Ranganathan negli anni Trenta (Revelli, 2010)[33], che permette la classificazione di documenti rispetto a soggetti multipli e/o complessi. Tale classificazione consente una ricerca degli elementi d'interesse a partire da più punti di vista, ma avendo come risultato sempre lo stesso oggetto/valore.

Grazie ai formalismi del Semantic Web, la disponibilità di informazioni *machine understandable* e *machine processable* permetterebbero, tra le altre cose, di supportare ricerche sul Web più intelligenti e di ridurre lo sforzo, a carico dell'utente, di selezionare le risorse pertinenti e di interesse per la sua richiesta. Infatti, cominciano ad affermarsi nello scenario di riferimento motori di ricerca semantici che, da un lato, non considerano le parole della query come semplici keyword, ma vi assegnano un significato più ampio e, dall'altro, utilizzano la propria base di conoscenza per costruire a *runtime* una risposta contenente dati strutturati che soddisfano appieno le esigenze conoscitive dell'utente. In questa direzione si stanno muovendo grandi compagnie come Wolfram[34], Google (nello specifico la recente funzionalità *Knowledge Graph*[35]), Bing (Snapshot e Sidebar)[36], ecc.

1.3.1 *Classificazione a Faccette e Ontologie*

Le faccette possono essere viste come assi di uno spazio cartesiano n-dimensionale, in cui il valore di ciascuna di esse corri-

33 Cfr. CARLO REVELLI, *Le Cinque leggi in italiano*, in «Biblioteche Oggi», vol. 28, n. 8, ottobre 2010, pp. 7-9.

34 <www.wolframalpha.com/>.

35 <http://googleblog.blogspot.it/2012/05/introducing-knowledge-graph-things-not.html>.

36 Bing Community. <http://www.bing.com/community/site_blogs/b/search/archive/2012/05/10/spend-less-time-searching-more-time-doing-introducing-the-new-bing.aspx>.

sponde alla posizione (coordinata) dell'oggetto da classificare nella dimensione corrispondente. Ogni oggetto viene quindi considerato sotto vari aspetti, che sono ortogonali fra loro.

Secondo Ranganathan (Ranganathan, 1985)[37], le faccette di qualsiasi classe si possono ricondurre a cinque categorie fondamentali: personalità (l'oggetto centrale di un discorso), materia (i componenti e le proprietà dell'oggetto), energia (le caratteristiche dinamiche dei processi che lo interessano), spazio (i suoi elementi geografici o in genere spaziali), e tempo (le sue fasi cronologiche).

Con una classificazione a faccette, il contenuto di un documento può essere descritto analiticamente nei suoi diversi aspetti; questi sono poi espressi tutti insieme, secondo una sequenza determinata da regole di funzionalità. Questo approccio (analitico-sintetico) consente, in generale, di ottenere una poligerarchia costituita da gerarchie disgiunte (semanticamente ortogonali). Ogni facceta può essere specializzata in categorie via via più specifiche, fra le quali ci può essere overlap.

La classificazione a faccette differisce dalle classificazioni tradizionali in alcuni punti (Vickery, 2008)[38]. Le faccette non sono rigide liste, ma sono libere di essere combinate le une con le altre con la massima libertà, così che ogni tipo di relazione fra termini e fra soggetti possa essere espressa. Essa, inoltre, non consiste esclusivamente in relazioni genus-species: combinando termini in soggetti composti, introduce nuove correlazioni logiche fra essi, riflettendo quindi meglio la complessità della conoscenza. Un oggetto classificato può presentare, se necessario, più

[37] Cfr. SHIYALI RAMAMRITA RANGANATHAN, *Facet Analysis. Fundamental Categories*, in «Theory of Subject Analysis», Littleton, Colorado, 1985, pp. 86-93.

[38] Cfr. BRIAN C. VICKERY, *Faceted Classification for the Web*, in «Axiomathes», vol. 18, n. 2, giugno 2008, pp. 145-160.

foci per la stessa faccetta, risolvendo così l'imbarazzo, tipico nelle classificazioni gerarchico-enumerative, di non sapere dove collocare, per esempio, un'automobile che è allo stesso tempo, ad esempio, *sport car*, *convertible* e *compact car*.

In generale, quindi, tale approccio prevede la creazione di tante faccette, appunto, quante sono le proprietà di una risorsa e di accedere o ricercare tale risorsa a partire da una qualunque combinazione di queste proprietà (mutuamente esclusive). Non una singola grande tassonomia, ma tante piccole tassonomie con diversi focus di partenza.

In Figura 8 è riportato un esempio di una classificazione a faccette di automobili: come categorie di classificazione sono state scelte la tipologia (i.e., nuovo, usato), il segmento (i.e., berlina, station wagon, monovolume) e il tipo di alimentazione (i.e., benzina, diesel, GPL, metano). Selezionando una o più di queste caratteristiche, il sistema va a raffinare i risultati proponendo ulteriori criteri di classificazione fino a mostrare i risultati che soddisfano le richieste.

È evidente che l'approccio multidimensionale permette la definizione di funzionalità di ricerca delle informazioni di tipo user-friendly, favorendo una maggiore interazione con l'utente: esso è guidato verso la definizione di query di raffinamento, riducendo i risultati, senza però ritrovarsi a fare ricerche che non ne restituiscono alcuno. In tal senso, è interessante notare come nell'esempio in Figura 8, andando avanti nella ricerca, all'utente siano presentati solo i criteri di ricerca che portano a dei risultati concreti, ad esempio, una volta selezionata la caratteristica *usato*, il sistema si accorge della mancanza di automobili usate alimentate a GPL e metano e visualizza solo *benzina* e *diesel* come ulteriori criteri di raffinamento.

La classificazione o navigazione a faccette è realizzata mediante la definizione di istanze ontologiche ciascuna appartenente a classi stabilite: i criteri di navigazione saranno le proprietà definite per le classi ontologiche a cui le istanze appartengono.

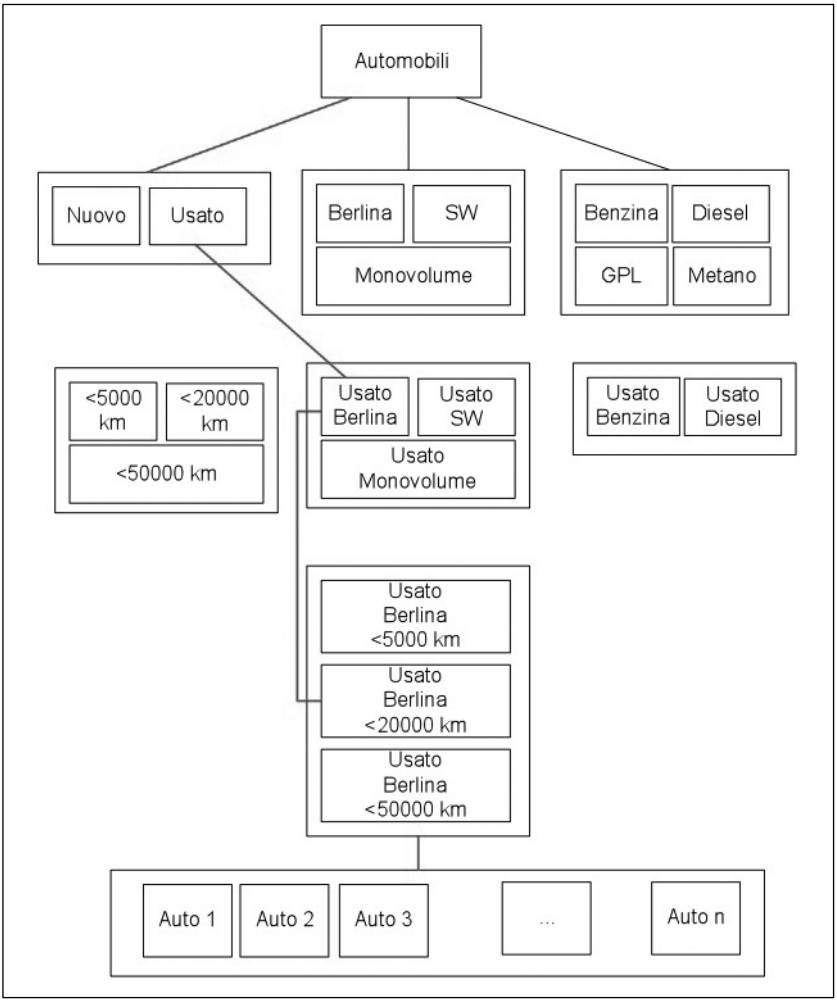


Figura 8. Esempio classificazione a faccette.

L'uso di ontologie consente una descrizione uniforme e controllata dei contenuti, indipendente dal formato e, nel caso di documenti testuali, dai termini posti nel documento. Il vantaggio

fornito dell'approccio a faccette è dovuto essenzialmente dall'uso di strutture dati dinamiche. In pratica l'uso di classificazioni multidimensionali (i.e., organizzate in faccette) fornisce un singolo e coerente framework in cui gli utenti possono concentrarsi su uno o più concetti dell'ontologia ed immediatamente vedere una sintesi concettuale dei loro obiettivi, in forma ridotta derivata dalla classificazione originale attraverso il filtraggio dei concetti non correlati. A sua volta, i concetti presenti nella classificazione ridotta possono essere usati per settare focus addizionali e dipendenti in modo da iterare la ricerca in un percorso non predeterminato, fintanto che l'utente non raggiunge un insieme di risultati sufficientemente piccolo. Un esempio di ricerca basata su faccette è mostrata in Figura 9. Tale approccio è sempre pronto ad accogliere nuove faccette ed è quindi intrinsecamente scalabile.

Inoltre, sebbene l'aggiunta di una faccetta renda necessaria la riclassificazione di tutte le entità secondo la nuova caratteristica, i vantaggi sono di gran lunga superiori all'inserimento o alla modifica di una o più categorie in uno schema gerarchico. Una faccetta in più determina, infatti, un aumento esponenziale del numero delle combinazioni potenziali e, quindi, un livello di specificazione maggiore del carattere descrittivo delle classi.

Un esempio di ricerca a faccette è il progetto *Semantic Web Environmental Director* (SWED)[39] che sfrutta la navigazione a faccette e utilizza tesauri, ontologie e liste per categorizzare, pubblicare e rappresentare le informazioni contenute nel proprio sito.

1.3.2 *Supporto all'interoperabilità*

Il problema dell'interoperabilità tra sistemi eterogenei può essere supportato mediante approcci sistematici rivolti, da un lato,

[39] <www.swed.org.uk/>.

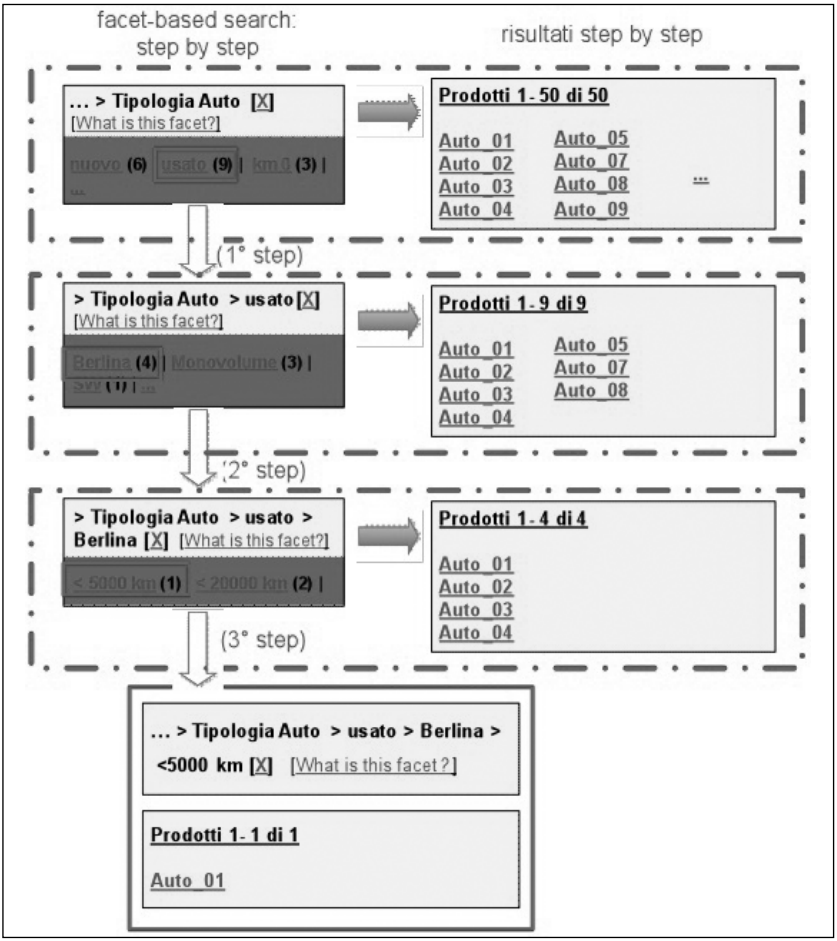


Figura 9. Un esempio di ricerca basate su faccettestep by step.

alla rappresentazione ed alla condivisione del dato e, dall'altro, alla riconciliazione, ossia alla possibilità di correlare tra loro concetti *simili* ma formalizzati in maniera differente. In tal senso le ontologie, ed in generale i formalismi del Semantic Web, sup-

portano il primo dei due aspetti evidenziati, cioè la rappresentazione *machine understandable* delle informazioni e della loro semantica. Il secondo aspetto, ossia la riconciliazione di schemi concettuali eterogenei, è una tematica ancora aperta e può essere supportata in modi diversi a seconda del contesto e degli obiettivi di interoperabilità che si intende ottenere. L'utilizzo delle ontologie, sebbene supporti adeguatamente la rappresentazione e la condivisione del dato, non risolve in modo nativo il problema della riconciliazione.

In letteratura scientifica, esistono diverse metodologie (e.g., (Noy, Musen, 2001)[40], (Stumme, Mädche, 2001)[41], (Kalfoglou, Schorlemmer, 2003)[42], ecc.) a supporto della riconciliazione semi-automatica di schemi concettuali eterogenei. D'altro canto, ci sono approcci che ben si adattano al dominio applicativo di riferimento, più concreti e sistematici rivolti a supportare l'interoperabilità tra modelli ontologici eterogenei. Questi, per lo più manuali, hanno un taglio più industriale e introducono benefici in termini di flessibilità e scalabilità dei modelli ontologici stessi e aumentano la longevità delle concettualizzazioni di dominio prodotte. In particolare, tali approcci riguardano principalmente la predisposizione di ontologie indipendenti dal dominio (i.e., *upper ontology*) in grado di abilitare la specificazione di costrutti

40 Cfr. NATALYA F. NOY, MARK A. MUSEN, *Anchor-PROMPT: Using non-local context for semantic matching*, in Atti del «17th IJCAI 2001 workshop on ontology and information sharing», 4-5 agosto 2001, Seattle, WA US, pp. 63–70. <http://sunsite.informatik.rwthaachen.de/Publications/CEUR-WS/Vol-47/>.

41 Cfr. GERD STUMME, ALEXANDER MÄDCHE, *FCA-merge: bottom-up merging of ontologies*, in Atti del «17th IJCAI2001 workshop on ontology and information sharing», 4-5 agosto 2001, Seattle, WA US, pp. 225–230.

42 Cfr. YANNIS KALFOGLOU, MARCO SCHORLEMMER, *If-map: an ontology mapping method based on information flow theory*, in «Journal of data semantics», vol. 1, 2003, pp. 98–127.

concettuali che supportano le correlazioni ontologiche in modo esplicitamente formalizzato e in sostanza, introducono delle ontologie in grado di supportare la specificazione formale del mapping (o *bridging*) tra rappresentazioni concettuali eterogenee. Pertanto, mediante tecnologie semantiche standard, questi approcci abilitano la definizione di meccanismi che supportano a *run time* un sistema software nelle attività di conversione e traduzione dei concetti tra le diverse rappresentazioni.

In molti sistemi di gestione della conoscenza esiste l'esigenza di modellare e correlare domini concettuali eterogenei a supporto della condivisione di informazioni. Pertanto, l'esigenza è più concretamente quella di correlare, piuttosto che convertire o tradurre, informazioni e rappresentazioni aderenti a schemi concettuali eterogenei.

Tale esigenza riguarda principalmente la definizione e l'utilizzo di ontologie indipendenti dal dominio per la rappresentazione di entità che vedono il coinvolgimento di concetti diversi. Dunque, s'intende abilitare attraverso le *upper ontology* la definizione delle entità oggetto di elaborazione e supportare attraverso le ontologie di dominio la correlazione e la mediazione tra risorse eterogenee che si collocano in domini applicativi affini.

Dunque, seguire tale tipo di approccio richiede essenzialmente le seguenti fasi: (i) identificazione delle onotologie necessarie a modellare le risorse che si intende gestire (e.g. profili utenti, documenti, task, servizi ecc.); (ii) identificazione delle *lightweight domain ontology,* ovvero ontologie spesso in forma di schemi tassonomici utili a contestualizzare la definizione delle istanze delle *upper ontology.*

A valle dell'identificazione la modellazione potrà considerare ciò che collega le *upper ontology* alle *domain ontology* per il conseguimento di una modellazione che, facendo perno sulla conoscenza di dominio, consente di correlare risorse eterogenee (Figura 10).

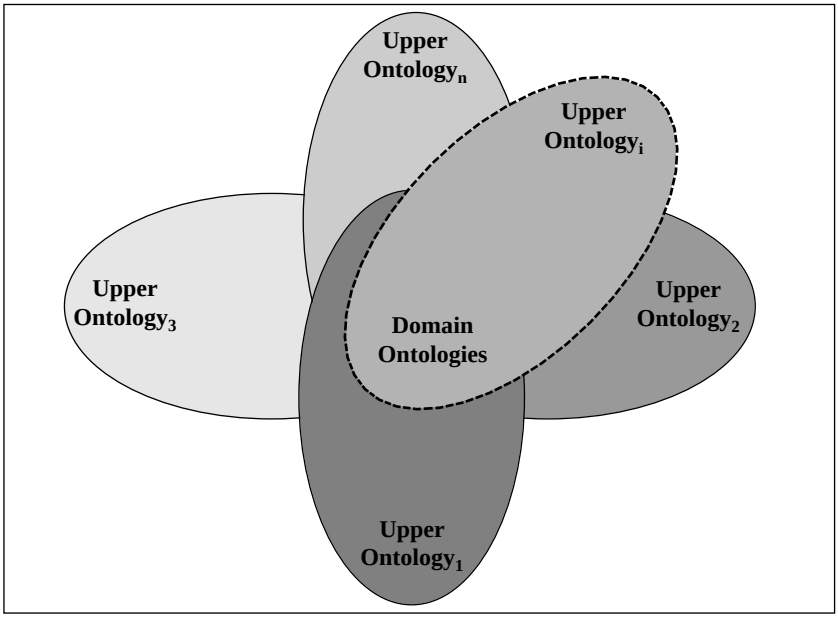


Figura 10. Valore aggiunto in termini di correlazione di risorse eterogenee.

1.4 Conclusioni

Come si è detto, il problema dell'interoperabilità tra sistemi eterogenei può essere supportato mediante approcci sistematici rivolti, da un lato, alla rappresentazione *machine understandable* del dato con ontologie e, dall'altro, alla riconciliazione di tali rappresentazioni concettuali inevitabilmente eterogenee.

Dunque, le ontologie sono un valido strumento formale per supportare la rappresentazione e la condivisione della semantica e dell'informazione. Tuttavia, l'uso delle ontologie abilita la predisposizione di modelli eterogenei che possono essere prima di tutto compresi e poi integrati attraverso opportune metodologie di riconciliazione. Pertanto, la riconciliazione, la mediazione e la correlazione di ontologie eterogenee sono attività che assumono

un ruolo cruciale per supportare l'interoperabilità nel suo complesso.

L'adozione di un approccio alla modellazione ontologica incentrato sulla correlazione di risorse eterogenee per mezzo di *lightweight domain ontology* consente di rilevare taluni valori aggiunti dall'approccio messo in atto.

Infatti, da un lato, si evidenziano valori aggiunti in termini di modellazione dello strato di conoscenza in un'ottica orientata al riuso, alla modularità ed alla scalabilità dei domini modellati, dall'altro, valori aggiunti in termini di supporto al filtraggio ed alla selezione dello specifico oggetto (e.g., profilo), scelto tra quelli disponibili, basti pensare alla *ricerca a faccette* che l'approccio supporta in modo nativo (Giunchiglia et alii, 2009)[44].

Specificatamente, la creazione di uno strato logico intermedio di correlazione, permette di mantenere la compatibilità verso ontologie note alla comunità che opera sulle tecnologie semantiche e di supportare evoluzioni future della modellazione di dominio. Infatti, l'eventuale definizione di ulteriori ontologie di dominio o evoluzioni di ontologie di dominio disponibili non impatta sulla validità della correlazione e della classificazione delle risorse stesse che potranno essere organizzate rispetto ai nuovi modelli ontologici.

Bibliografia

ABBAGNANO, N., *Dizionario di filosofia*, Torino, UTET, 1968

BERNERS-LEE, T., HENDLER, J., LASSILA, O., *The semantic web*, in «Scientific American», vol. 284, n. 5, 2001, pp. 35-43

BLUM, P.R., *Dio e gli individui: L'«ArborPorphyriana» nei secoli XVII e XVIII*, in «Rivista di filosofia neo-scolastica», vol. 91, 1999, pp. 18-49

[44] Cfr. FAUSTO GIUNCHIGLIA, BISWANATH DUTTA, VINCENZA MALTESE, *Faceted Lightweight Ontologies*, in Conceptual Modeling: Foundations and Applications, 2009, pp. 36-51.

BOJARS, U., PASSANT, A., BRESLIN, J.G., DECKER, S., *The SIOC Project: Semantically-Interlinked Online Communities*, in COIN@AAMAS&IJCAI&MALLOW, vol. 6069, Springer, 2009, pp. 179-194

CAPUANO, N., *Ontologie OWL: Teoria e Pratica*, in «Computer Programming» nn. 148, 149, 150, luglio/agosto, settembre, ottobre 2005

CORCHO, O., FERNANDEZ, M., GOMEZ-PEREZ, A., *Methodologies, Tools and Languages for Building Ontologies: Where is the Meeting Point?*, in «Data & Knowledge Engineering», vol. 46, 2003, pp. 41-64 <http://www.dia.fi.upm.es/~ocorcho/documents/DKE2003_CorchoEtAl.pdf>

DE MAIO, C., LOIA, V., FENZA, G., GALLO, M., LINCIANO, R., MORRONE, A., *Fuzzy knowledge approach to automatic disease diagnosis*, in 2011 IEEE International Conference on Fuzzy Systems Proceedings, Taipei, 27-30 giugno 2011, pp. 2088-2095

DE MAIO, C.,FENZA, G., LOIA, V.,SENATORE, S.,*Hierarchical web resources retrieval by exploiting Fuzzy Formal Concept Analysis*, in «Information Processing & Management», vol. 48, n. 3, Elsevier, 2012, pp. 399-418

GIUNCHIGLIA, F., DUTTA, B., MALTESE, V., *Faceted Lightweight Ontologies*, in Conceptual Modeling: Foundations and Applications, 2009, pp. 36-51

GRUNINGER, M., LEE, J., *Introduction - ontology: different ways of representing the same concept*, Communication of the ACM, vol. 45, n. 2, 2002, pp. 39-41

HAPPEL, H.J., SEEDORF, S., *Applications of Ontologies in Software Engineering*, in Atti del «2nd International Workshop on Semantic Web Enabled Software Engineering (SWESE 2006) - 5th International Semantic Web Conference (ISWC 2006)», Athens, GA, USA, 05-09 novembre 2006

ISO 3166-*Codici per la rappresentazione di nomi di paesi e delle loro suddivisioni*

KALFOGLOU, Y., SCHORLEMMER, M., *If-map: an ontology mapping method based on information flow theory*, in «Journal of data semantics», vol. 1, 2003, pp. 98-127

LABORATORY FOR APPLIED ONTOLOGY (LOA), Institute of Cognitive Science and Technology, *Foundational Ontologies & Their Library*, 2005. pp. 32-34

NOY, N., MUSEN, M., *Anchor-PROMPT: Using non-local context for semantic matching*, in Atti del «17th IJCAI 2001 workshop on ontology and information sharing», 4-5 agosto 2001, Seattle, WA US, pp. 63-70 <http://sunsite.informatik.rwthaachen.de/Publications/CEUR-WS/Vol-47/>

RANGANATHAN, S. R., *Facet Analysis. Fundamental Categories*, in «Theory of Subject Analysis», Littleton, Colorado, 1985, pp. 86-93

REVELLI, C., *Le Cinque leggi in italiano*, in «Biblioteche Oggi», vol. 28, n. 8, ottobre 2010, pp. 7-9

STUMME, G., MÄDCHE, A., *FCA-merge: bottom-up merging of ontologies*, in Atti del «17th IJCAI2001 workshop on ontology and information sharing», 4-5 agosto 2001, Seattle, WA US, pp. 225-230
VICKERY, B. C., *Faceted Classification for the Web*, in «Axiomathes», vol. 18, n. 2, giugno 2008, pp. 145-160

Sitografia

<http://googleblog.blogspot.it/2012/05/introducing-knowledge-graph-things-not.html>
<http://protege.stanford.edu/>
<http://sioc-project.org/>
<http://www.bing.com/community/site_blogs/b/search/archive/2012/05/10/spend-less-time-searching-more-time-doing-introducing-the-new-bing.aspx>
<http://www.foaf-project.org/>
<http://www.scot-project.org/>
<http://www.w3.org/Submission/OWL-S/>
<http://www.w3.org/TR/owl2-overview/>
<http://www.w3.org/TR/owl-features/>
<http://www.w3.org/TR/wsdl>
<http://xmlns.com/foaf/spec/>
<www.geonames.org/>
<www.swed.org.uk/>
<www.unspsc.org/>
<www.w3.org/>
<www.w3.org/html/>
<www.w3.org/RDF/>
<www.w3.org/XML/>
<www.wolframalpha.com/>