MATERIALI

# Biblioteche digitali e Web semantico

di Maria Teresa Biagetti

Si manifesta in modo sempre più evidente una convergenza di interessi e uno scambio di conoscenze tra gli specialisti che si occupano della realizzazione e della gestione delle Biblioteche digitali e gli esperti che si dedicano allo sviluppo del Web semantico. Le tecniche adottate per individuare e recuperare i documenti gestiti dalle Biblioteche digitali possono essere affinate sfruttando in particolare le possibilità offerte dalle ontologie create per il Web semantico, e le raccolte di documenti messi a disposizione dalle Biblioteche digitali possono costituire un insieme di informazioni e di concetti per l'incremento delle ontologie stesse.

È importante inoltre ricordare che il complesso di esperienze maturate nell'ambito della *Library and information science* nella realizzazione dei sistemi di organizzazione della conoscenza, in particolare classificazioni biblioteconomiche, tesauri e liste di soggetti, è stato fatto proprio dal Semantic Web Deployment Working Group, gruppo di lavoro del World Wide Web Consortium (W3C) che si occupa dello sviluppo di RDF (*Resource Description Framework*)[1] e dell'applicazione e dell'uso di OWL (*Web Ontology Language*)[2], e trasferito nell'ambito delle tecnologie e degli strumenti sviluppati per l'organizzazione semantica del Web, con l'utilizzo di linguaggi di rappresentazione della conoscenza definiti formalmente, come RDF e OWL. Il risultato di questa integrazione di competenze è SKOS (*Simple Knowledge Organization System*), un ambito di ricerche e applicazioni dedicato alla realizzazione di un modello adatto ad esprimere la struttura di base e i contenuti dei tradizionali sistemi semiformali per l'organizzazione della conoscenza. Utilizzando lo standard *SKOS Simple Knowledge Organization System Reference*[3] (Recommendation del W3C del 18 agosto 2009), tesauri e sistemi di classificazioni possono essere re-ingegnerizzati come ontologie formali, trasformando la struttura di concetti esplicitata in un tesauro, ad esempio, in un insieme più sfaccettato ed elaborato di classi e proprietà, identificate da URI.

MARIA TERESA BIAGETTI, Sapienza Università di Roma, Scuola speciale per archivisti e bibliotecari, viale Regina Elena 295, 00161 Roma, e-mail mariateresa.biagetti@uniroma1.it.

**1** <http://www.w3.org/TR/rdf-primer/> (W3C Recommendation 10 February 2004).

**2** OWL Overview (W3C Recommendation 10 February 2004) <http://www.w3.org/TR/owl-features/>. Recentemente è stata realizzata la nuova versione di OWL, OWL 2: <http://www.w3.org/TR/owl2-overview/> (W3C Recommendation 27 October 2009).

**3** <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>.

Bollettino **AIB**, ISSN 1121-1490, vol. 49 n. 4 (dicembre 2009), p. 577-582.

La cooperazione, l'influenza reciproca e l'interazione sempre più stretta, che si esplicano attraverso un ventaglio di attività e di applicazioni, tra le Biblioteche digitali e la tecnologia dedicata allo sviluppo del Web semantico, è stato l'argomento centrale della seconda *International conference for digital libraries and the semantic Web* (ICSD), che si è svolta il 10 e l'11 settembre 2009 all'Università di Trento, organizzata dal Dipartimento di ingegneria e scienza dell'informazione, a cura di Paolo Bouquet, e ospitata dalla Facoltà di sociologia. La prima conferenza era stata organizzata a Bangalore dall'Indian Statistical Institute nel 2007.

Preceduta da due workshop svoltisi nei giorni 8 e 9 settembre e dedicati a *Advanced technologies for digital libraries e multilinguality in information access to digital libraries. User needs and evaluation of multilingual resources use*, la conferenza di Trento ha visto la partecipazione di centodue esperti provenienti da diciannove nazioni, rappresentative di quattro continenti. Le diciotto comunicazioni presentate, articolate in quattro sessioni di lavoro, ciascuna introdotta da un ampio intervento dei relatori invitati (Dagobert Soergel, Sankar Kumar Pal, Julien Masanes, Fausto Giunchiglia), hanno proposto l'analisi di temi e progetti che possono essere raggruppati nei seguenti ambiti:

### 1 Generazione automatica di ontologie utilizzando *corpora* di documenti gestiti dalle biblioteche digitali

Sulla cooperazione tra biblioteche digitali e Web semantico ha incentrato il suo intervento di apertura della prima giornata del convegno, *Digital libraries and the semantic Web. A conceptual framework and an agenda for research and practice*, Dagobert Soergel (University of Maryland), affrontando l'argomento da una prospettiva eminentemente biblioteconomico-documentaria, con particolare attenzione all'arricchimento semantico dei documenti. Soergel ha sottolineato che la tecnologia e gli strumenti elaborati per il Web semantico possono essere impiegati per migliorare le possibilità di ritrovamento dei documenti nelle Biblioteche digitali, e dal momento che le biblioteche digitali gestiscono collezioni organizzate di documenti, a loro volta esse possono fornire materiali per generare, incrementare e popolare le ontologie, strumento tra i più efficaci per la realizzazione del Web semantico. L'estrazione automatica o semiautomatica, con il supporto del controllo umano, di liste strutturate di concetti, risulta più agevole se la fonte è costituita da tesauri, dizionari e sistemi di classificazione, dando luogo alla realizzazione di archivi di proposizioni espresse in un linguaggio formalizzato, con il riferimento alla fonte.

Un nuovo strumento per la generazione automatica delle ontologie, e per la modifica di ontologie esistenti - SPRAT (*Semantic Pattern Recognition and Annotation Tool*) - è stato ad esempio presentato da Diana Maynard (Sheffield University). Basandosi su modelli di analisi lessicale e sintattica già sperimentati, SPRAT permette di aggiungere ad esempio sottoclassi in ontologie già esistenti, sfruttando la semantica fornita dal contesto. I concetti che risultano già presenti in una ontologia possono essere così arricchiti di elementi che non erano presenti. Oppure, si possono creare nuove classi sulla base dei concetti rilevabili tramite l'analisi contestuale: ad esempio, la presenza in uno stesso documento di termini come *hornsharks* (squali), *leopard sharks* (squali tigrati) e c*atsharks* (altra famiglia di squali piccoli), i primi due già presenti nell'ontologia che si vuole arricchire, permette, attraverso l'analisi del contesto e l'analisi lessicale, di aggiungere automaticamente una classe (c*atsharks*), assente nell'ontologia.

Tuttavia, come ha riconosciuto la stessa Maynard, per raggiungere un livello accettabile di correttezza nella generazione automatica di sottoclassi sono necessari ulte-

riori studi e la messa a punto di tecniche più raffinate. I risultati della valutazione presentati da Diana Maynard mostrano infatti una percentuale di correttezza inferiore al 50% nella individuazione di nuove sottoclassi, e un livello basso (22%) di correttezza nella individuazione di proprietà, che in una ontologia definiscono le relazioni che collegano gli individui di una classe.

**2 Arricchimento semantico e indicizzazione automatica dei documenti**

L'annotazione semantica dei documenti aumenta le possibilità di ritrovamento dei documenti. Le singole entità precisamente definibili – i nomi di persone, di luoghi, di enti, le date e gli avvenimenti – e i concetti cui nel testo si fa riferimento, possono essere identificati facendo ricorso ad ontologie (ma anche a tesauri e a nomenclature ufficiali). Nel suo intervento, Dagobert Soergel, riferendosi alle modalità di mappatura dei concetti e delle entità[4], ha sottolineato l'importanza di corredare i documenti digitali anche con annotazioni che identifichino concetti che non sono esplicitamente citati nel documento, ma ai quali il documento rinvia, concetti cioè che possono essere semanticamente implicati dal documento nella sua interezza o da una sua parte. Questa tipologia di indicizzazione può avvenire in modo automatico, applicando metodologie di tipo statistico, oppure può essere realizzata dall'operatore umano con il supporto delle ontologie, o infine prodotta automaticamente sulla base di regole d'inferenza previste da ontologie provviste di un alto grado di espressività logica. I concetti potranno essere rappresentati attraverso RDF oppure attraverso OWL, il linguaggio per le ontologie nel Web, e la ricerca potrà poi avvenire tramite motori di ricerca o agenti intelligenti.

È ancora Soergel a sottolineare l'importanza del fatto che i documenti *digital born* che entrano a far parte di collezioni digitali presentino la mappatura dei concetti fin dalla loro origine, preferibilmente realizzata dagli autori stessi all'atto della scrittura del testo, e la necessità, che ne consegue, di disporre di strumenti di supporto che convertano in un linguaggio formale le proposizioni create dagli autori usando il linguaggio naturale. Anche questa attività potrebbe contribuire ad un arricchimento delle ontologie per il Web semantico: alcune proposizioni, convertite in un linguaggio formale, potrebbero infatti essere accolte nelle ontologie.

Nel suo intervento - *Faceted lightweight ontologies* - Fausto Giunchiglia (Università di Trento) ha illustrato i fondamenti su cui si basa la realizzazione di ontologie utilizzabili nella indicizzazione automatica dei documenti, mostrando apprezzamento per la *Library science* nel suo complesso e per la realizzazione dei sistemi di classificazione biblioteconomica, in particolare di quelli a faccette, e dei linguaggi d'indicizzazione. Le ontologie *lightweight*, ontologie con struttura ad albero, espresse in linguaggio formale e rappresentabili utilizzando OWL, usate allo scopo di realizzare la condivisione dei significati e favorire l'interoperabilità semantica, risultano più vicine alle rappresentazioni semiformalizzate di concetti, come i sistemi di classificazione e i tesauri, che alle *heavyweight ontologies*, strutture complesse di concetti, relazioni e assiomi, dotate di capacità inferenziale. Nella realizzazione di questo tipo di ontologie Giunchiglia ha seguito il modello delle classificazioni a faccette, in particolare applicando le faccette stabilite da G. Bhattacharyya, basate sulle

**4** Per le entità, il riferimento è al Progetto di ricerca OKKAM <http://www.okkam.org/>, finanziato dalla Comunità Europea nell'ambito del 7 Programma quadro, che provvederà entro il 2010 a realizzare l'identificazione univoca sul Web di persone, enti, eventi e prodotti. Il Consorzio di istituzioni dei paesi che stanno curando questa iniziativa è coordinato dall'Università di Trento (Project Coordinator: Paolo Bouquet).

teorie di Ranganathan: *Domain* (o *Discipline*), *Entity*, *Property*, *Action*, *Modifier*, ed esplicitate nelle categorie del linguaggio d'indicizzazione POPSI (Postulate based Permuted Subject Indexing). Le entità, gli oggetti, le proprietà e le azioni relative a ciascun dominio scientifico costituiscono il *background* di conoscenza formalizzata, organizzata per gruppi omogenei di concetti (faccette), da cui estrarre i termini che costituiscono l'ontologia. L'uso delle faccette rende esplicite le relazioni logiche tra i concetti, superando i limiti delle strutture gerarchiche, e permette di considerare entità complesse secondo prospettive diverse, a seconda della disciplina.

La necessità di strutturare i documenti fin dalla loro origine appositamente per il Web è stata sottolineata da Henrik Eriksson (Linköping University), che nel suo intervento, *Towards semantic documents for Digital libraries and document repositories*, ha chiarito che l'arricchimento semantico dei documenti gestiti da Biblioteche digitali o presenti nei *repositories*, può essere ora realizzabile non solo tramite *word processors* dotati di apposite funzionalità (come KWrite, OntOffice o WickOffice), ma anche adottando per la creazione di documenti il formato PDF, estremamente diffuso e di facile impiego, e utilizzando ad esempio l'*editor* per ontologie p*rotégé*. Gli sviluppi futuri sono legati alle possibilità di realizzare la modularizzazione delle ontologie, moduli o subontologie, ai quali collegare concetti e sezioni dei documenti.

### 3 Miglioramento delle strategie per il ritrovamento e l'accesso ai documenti nelle biblioteche digitali

L'utilizzazione delle ontologie permette ad esempio di migliorare soprattutto il livello di *precision* nella ricerca dei documenti. Matias Frosterus e Eero Hyvönen (Helsinki University), nel loro importante intervento - *Bridging the search gap between the Web of pages and Web of data by combining ontological document expansion with text search* - hanno sottolineato la necessità di adottare strategie che consentano di combinare la ricerca attraverso le parole del testo, che permette un basso livello sia di *recall* che di *precision*, con la ricerca dei documenti condotta a livello semantico puro, in uno spazio concettuale creato attraverso la *document expansion*, resa possibile dall'uso di ontologie. Se l'uso delle ontologie limitato all'espansione della *query*, durante la fase della ricerca, poteva permettere di condurre ricerche impiegando tutti i sinonimi dei termini adottati o tutti i termini che denominano i concetti sussunti sotto una classe, l'uso delle ontologie durante la fase dell'indicizzazione consente invece una mappatura automatica dei termini e dei concetti presenti nel documento con quelli presenti nell'ontologia, ma anche con i concetti ad essi semanticamente correlati, ai quali si propone di attribuire un peso semantico inferiore.

Le difficoltà, che offrono quindi l'opportunità di ulteriori indagini per l'immediato futuro, nascono dal fatto che nell'espansione semantica dei documenti si possono seguire diversi modelli per definire le relazioni tra concetti. Ciascun modello adottabile prevede un certo numero di percorsi e relazioni di livello diverso, e può condurre a risultati diversi.

Le prospettive offerte, nell'organizzazione dei documenti e dei servizi nel Web, dall'insieme di metodologie e orientamenti conosciuto come *Soft computing*, che utilizza sia la logica *Fuzzy*, che prevede la parziale sovrapposizione dei contenuti nella definizione delle classi e la determinazione di confini imprecisi tra di esse, sia le potenzialità offerte dalla realizzazione dei *Rough sets*, insiemi approssimativi, sono state illustrate nell'ampio intervento che ha aperto la seconda sessione dei lavori, particolarmente tecnico e caratterizzato da un taglio logico, di Sankar Kumar Pal (Indian Statistical Institute, Calcutta): *Soft computing, rough sets, information granules and applications in Web intelligence.* Pal ha delineato, in prospettiva, la realizza-

zione di *Granular information retrieval*, applicato in particolare a insiemi di documenti multimediali in rete, testi, immagini, ipertesti, che permetterà di ritrovare più efficacemente i documenti, sfruttando l'applicazione di *Information granules*, casi, prototipi o modelli informativi rappresentativi di classi non rigorosamente determinate, ma dai confini imprecisi e indistinti.

Inserendosi nell'ambito delle ricerche sul *Cross-language information access*, la tecnologia elaborata per permettere il ritrovamento dei documenti in lingue diverse partendo da una *query* espressa in una lingua, ad esempio in italiano, Alessio Bosca (Politecnico di Torino) e Luca Dini hanno presentato una metodologia innovativa per aumentare le possibilità di traduzione automatica delle *queries* rivolte dagli utenti ai sistemi di ricerca nelle Biblioteche digitali in lingue diverse. Propongono infatti di sfruttare i *log-files*, che registrano le transazioni degli utenti. Gli utenti di Biblioteche digitali in ambiente di ricerca spesso traducono la loro *query* anche in inglese, per avere maggiori possibilità di recupero. Usare i *log-files* per permettere l'accesso a documenti in lingue diverse, assicura alta rilevanza per l'utente e l'adozione di termini specializzati nei diversi domini scientifici, mentre non sempre una traduzione letterale può rendere ragione di ciò.

**4 Conversione dei tradizionali strumenti di organizzazione della conoscenza**

Uno spazio marginale è stato riservato a SKOS (*Simple Knowledge Organization System*), che avrebbe potuto essere invece più valorizzato durante il convegno. Sotto l'egida di W3C, SKOS fornisce un insieme di tecnologie, strumenti e modelli adatti a rappresentare la struttura di base dei tradizionali sistemi semiformali per l'organizzazione della conoscenza (tesauri, sistemi di classificazione, liste di intestazioni per soggetto e glossari), e dei sistemi di concetti in essi presenti, e utilizzando RDF permette una facile migrazione nel Web dei sistemi di organizzazione della conoscenza già esistenti, ne consente la condivisione e la riutilizzazione.

Un riferimento al sistema SKOS è presente nell'intervento di Benjamin Zapilko (GESIS –Leibniz - Institut für Sozialwissenschaften, Bonn) e York Sure (Universität Koblenz-Landau), *Transferring the Shell model to the semantic Web and the impact on text-fact-integration*. I due autori propongono di riformulare, trasferendolo nell'ambito del Web semantico, *Shell model*, un modello che è già stato utilizzato nelle Biblioteche digitali, che permette il collegamento tra documenti dal contenuto eterogeneo, indicizzati secondo modalità diverse in quanto provenienti da diverse fonti, e adottato nei portali tedeschi di informazione interdisciplinare <http://vascoda.de> o dedicati alle Scienze sociali (ad esempio: <http://www.sowiport.de/>). La ricerca integrata di dati testuali e fattuali (articoli in full-text, ma anche riferimenti bibliografici, dati sui progetti di ricerca e profili dei ricercatori), che presentano quindi un alto livello di eterogeneità semantica e possono essere indicizzati seguendo diversi livelli di profondità, è una necessità avvertita in particolare dai ricercatori nell'ambito delle Scienze sociali. La trasformazione di *Shell model* comporterebbe il rimodellamento in RDF dei dati testuali presenti nei database e il trasferimento del tesauro tedesco delle Scienze sociali *TheSoz* all'interno del sistema SKOS, che permetterebbe di utilizzare gli strumenti del Web semantico, come il linguaggio OWL.

Nel complesso, la conferenza tenutasi a Trento ha offerto la possibilità di entrare in contatto con un ricco ventaglio di esperienze e ha rappresentato un contributo significativo, che apre prospettive alla ricerca, contribuendo a collegare la *Library and information science* alle tecnologie innovative e agli strumenti per la realizzazione del governo semanticamente avanzato del Web. Alcuni interventi, in particolare

quelli nei quali è stata trattata l'indicizzazione automatica dei documenti, suggeriscono tuttavia alcune riflessioni.

L'indicizzazione automatica dei documenti e la mappatura dei concetti, realizzate col supporto delle ontologie e che permettono di recuperare oggetti digitali nei quali è stato trattato il singolo concetto, costituiscono indubbiamente uno strumento potente di ritrovamento, una "forza bruta", come la definiscono gli esperti di intelligenza artificiale, della quale sarebbe sciocco proporre di fare a meno. Se è possibile però indicizzare efficacemente in modo automatico i documenti prodotti nell'ambito delle scienze esatte o in campo medico, il livello di efficacia e di rispondenza alle necessità della ricerca si abbassa decisamente quando a essere indicizzata è la letteratura prodotta in particolare nell'ambito delle scienze umane. In quest'ultimo caso, infatti, emergono alcuni aspetti peculiari, determinati dal fatto che i concetti elaborati sono più facilmente suscettibili di molteplici interpretazioni, possono acquistare significatività diverse in relazione a interpreti diversi, e gli argomenti trattati possono essere indagati secondo prospettive differenti e nell'ambito di vari scenari di ricerca. Considerando questi aspetti, e tenendo presente la tipologia di documenti che la gran parte delle Biblioteche digitali mette a disposizione, permangono alcune perplessità sull'efficacia dell'indicizzazione automatica dei documenti nelle Scienze umane, nonostante le stimolanti prospettive di ricerca tracciate da alcuni interventi, come quelli di Sankar K. Pal e di Matias Frosterus e Eero Hyvönen.

Infine, bisogna rilevare che si è sentita la mancanza di un esame delle opportunità e delle criticità offerte dalle ontologie di livello logico superiore, le ontologie dotate di motori per le inferenze logiche, che possono lavorare applicando contemporaneamente diversi modelli di abduzione.