Sviluppi e trasformazioni delle biblioteche digitali: dai *repositories* di testi alle *semantic digital libraries*

di Maria Teresa Biagetti

1. Introduzione

Le trasformazioni che si sono prodotte sul versante delle biblioteche digitali nell'ultimo decennio hanno determinato il manifestarsi di una pluralità di tipologie di sistemi, di un universo di biblioteche digitali, e perciò risulta praticamente impossibile formulare una definizione sintetica di biblioteca digitale che possa essere rappresentativa di tutte le caratteristiche che sono attualmente riscontrabili.

Nell'ambito di questa molteplicità di modelli si possono comprendere sia le realizzazioni di *repositories* di documenti, nati digitali o digitalizzati, per lo più istituiti dai governi nazionali e dedicati alla preservazione e alla fruizione dei patrimoni culturali delle singole nazioni, basti citare Gallica e la Biblioteca digitale italiana, sia quelle strutture, da Europeana alla World digital library, che organizzano la fruizione di oggetti digitali, libri, documenti d'archivio, musica, film, oggetti museali, appartenenti a diversi partner, sia quelle realizzazioni avanzate che offrono, in aggiunta, strumenti per l'analisi dei testi e il loro studio[1].

Su di un altro piano, ma comunque, secondo l'opinione di molti, riconducibili all'universo delle biblioteche digitali, i depositi ad accesso aperto di documenti prodotti da istituzioni scientifiche e università, messi liberamente a disposizione degli studiosi per l'incremento della conoscenza e la prosecuzione degli studi e, infine,

MARIA TERESA BIAGETTI, Sapienza Università di Roma, Dipartimento di Scienze documentarie, linguistico-filologiche e geografiche, viale Regina Elena n. 295, 00161 Roma, e-mail mariateresa.biagetti@uniroma1.it

Ultima consultazione dei siti web: 12 marzo 2014

1 Gallica, <http://gallica.bnf.fr/?&lang=FR>, attiva dal 1997, attualmente offre 2 milioni e mezzo di documenti digitalizzati, tra libri a stampa e manoscritti, registrazioni sonore, partiture musicali. La Biblioteca digitale italiana, <http://www.internetculturale.it/opencms/opencms/it/>, aperta nel 2001, oggi permette di consultare liberamente libri, spartiti, fotografie, carte geografiche. Europeana, <http://www.europeana.eu/>, a partire dal 2008, offre circa 5 milioni di oggetti digitali (libri, documenti d'archivio, trasmissioni televisive, musica, film, oggetti museali) di cui i partners forniscono l'accesso. World digital library, <http://www.wdl.org/en/>, curata dal 2005 da Library of Congress e Unesco, per il momento fornisce l'accesso a 1.300 documenti rappresentativi delle culture di tutto il mondo, dando visibilità a documenti posseduti da nazioni che non avrebbero la possibilità di gestire autonomamente una biblioteca digitale.

AIB studi, vol. 54 n. 1 (gennaio/aprile 2014), p. 11-34. DOI 10.2426/aibstudi-9955.

anche i sistemi per lo sviluppo di biblioteche digitali da utilizzare come supporto dell'*e-learning*, i quali possono fornire documenti scientifici, monografie e materiale di studio validato dai docenti per agevolare l'apprendimento a distanza degli studenti[2].

La definizione fornita nel 1998 da Digital Library Federation, e ancora oggi una tra le più citate e condivise, poneva l'accento sulla biblioteca digitale considerata come una organizzazione di risorse digitali selezionate e preservate nel tempo per la fruizione da parte di un'utenza specifica:

> Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities[3].

Nel dicembre 2007, in pieno *Sixth Framework Programme* della Commissione europea, dedicato, tra l'altro, allo sviluppo delle tecnologie per la società dell'informazione, nel *Manifesto for digital libraries*, IFLA ha definito invece la biblioteca digitale come una parte fondamentale dei servizi di una biblioteca, in quanto rende possibile l'accesso a vaste comunità di utenti di collezioni di documenti digitali dalla qualità certificata, gestiti secondo principi accettati e condivisi a livello internazionale, assicurando loro i servizi necessari alla consultazione e alla fruizione. L'accento si sposta sui bisogni espressi dalla società dell'informazione e sulla necessità di assicurare l'interoperabilità dei sistemi.

> A digital library is an online collection of digital objects, of assured quality, that is created or collected and managed according to internationally accepted principles for collection development and made accessible in a coherent and sustainable manner, supported by services necessary to allow users to retrieve and exploit the resources.
> A digital library forms an integral part of the services of a library, applying new technology to provide access to digital collections. Within a digital library collections are created, managed and made accessible in such a way that they are readily and economically available for use by a defined community or set of communities.
> A digital library provides a mechanism for collaboration between public and research libraries to form a network of digital information in response to the needs of the Information Society. The systems of all partners in a digital library must be able to interoperate.

2 L'apporto che le biblioteche digitali possono fornire al sistema dell'apprendimento a distanza, sia nel caso dei corsi universitari, sia nel caso dei corsi di apprendimento permanente e di formazione professionale, coinvolge problematiche complesse che spaziano dalla possibilità di digitalizzare fonti e monografie suggerite dai docenti, con il rispetto del copyright, all'opportunità di fornire supporto in caso di videoconferenze da parte della biblioteca dell'Ateneo. La tematica è stata dibattuta nell'ambito del Progetto Digital Libraries Applications. *Biblioteche digitali in Italia. Scenari, utenti, staff e sistemi informativi. Rapporto di sintesi del Progetto Digital Libraries Applications.* Coordinato e curato da Anna Maria Tammaro in collaborazione con Stefano Casati e Damiana Luzzi. Fondazione Rinascimento digitale, 2006, p. 31-37, <http://www.rinascimento-digitale.it/documenti/dla/BibliotecheDigitaliItalia2006.pdf>.

3 Digital Library Federation fu fondata nel 1995 da Library of Congress insieme ad alcune importanti Università, tra cui Yale, Harvard, Stanford, Princeton, e da New York Public Library, con lo scopo di gestire e preservare gli oggetti digitali che documentano il patrimonio culturale americano, perché fossero disponibili per i cittadini e gli studiosi. Oggi accoglie 35 istituti, tra cui British Library, MIT e Bibliotheca Alexandrina, <www.diglib.org/about/dldefinition.htm> [1998].

> A digital library complements initiatives to develop digital archives for the preservation of digital content. In this way, a digital library helps to improve preservation of and access to cultural and scientific heritage[4].

Il termine Digital Library può essere utilizzato per indicare sia semplici *repositories*, sia sistemi più complessi che prevedono caratteristiche avanzate. Una biblioteca digitale si considera un semplice *digital repository* di oggetti digitali o digitalizzati, forniti dei loro metadati, quando si limita ad acquisire, descrivere, preservare e rendere accessibili e consultabili i documenti da essa gestiti. In sostanza, si tratta di servizi di accesso a documenti digitali selezionati, descritti e mantenuti; ciascuna biblioteca digitale opera isolatamente, concentrando gli sforzi sulla gestione dei documenti digitali.

Una biblioteca digitale che offre invece funzionalità avanzate, può permettere agli utenti di ricercare testi, oggetti, immagini e oggetti compositi, attraverso strumenti semanticamente raffinati, utilizzando ad esempio anche le annotazioni realizzate dagli altri fruitori. Funzionalità più evolute possono consentire inoltre l'uso di strumenti per l'elaborazione e l'*editing* di nuovi documenti e per la creazione di annotazioni e links, di dispositivi per estrarre le citazioni presenti nei documenti e convertirle in records bibliografici, di strumenti per analizzare ed esportare i dati, per creare simulazioni utilizzando i dati esistenti, e infine di *tools* utilizzabili dall'utente per creare grafi o pagine web riutilizzando i testi che sono a disposizione nel *repository*[5].

Un esempio significativo di biblioteca digitale che rientra in quest'ultima tipologia, arricchita in particolare con strumenti adatti all'analisi testuale, è costituito da Perseus Digital Library[6], progettata nel 1987, sito web dal 1995, e rifondata su un *Digital Library System* da Tufts University (Boston e Talloires) con caratteristiche di modularità e interoperabilità, nei primi anni del nuovo millennio. Perseus mette a disposizione collezioni di testi in greco, latino e arabo, per lo più indirizzando alle copie digitali di Google Books, di Open Content Alliance e di Internet Archive, ma anche immagini digitali di artefatti fisici, di spazi storicamente rilevanti, unitamente a servizi all'utenza, prevalentemente nei settori della storia, della letteratura e della cultura del mondo greco-romano.

La caratteristica principale è l'offerta ai fruitori di strumenti per utilizzare i documenti nel modo più confacente alle necessità di studio. Oltre a predisporre link a dizionari online per le lingue latina, greca e araba, che forniscono i significati delle parole incontrate nei testi durante la lettura, Perseus offre un potente supporto linguistico all'utente che si esplica attraverso la realizzazione della lemmatizzazione automatica dei testi gestiti dalla biblioteca digitale, allestendo link dalle forme flesse alla forma da cui derivano, cui si aggiunge l'analisi morfologica delle parole.

I documenti del passato sono posti a disposizione degli utenti attraverso la fruizione diretta delle immagini digitalizzate di oggetti, documenti, luoghi e iscrizioni. Sono offerti al lettore ulteriori strumenti di informazione e conoscenza, come la possibilità di visualizzare i record catalografici relativi alle diverse edizioni di ciascuna opera e di accedere alla copia digitale, la possibilità di consultare le relative voci di

4 *IFLA Manifesto for digital libraries,* <http://www.ifla.org/publications/iflaunesco-manifesto-for-digital-libraries>.

5 Dagobert Soergel, *Digital libraries and knowledge organization*. In: *Semantic digital libraries*, editors Sebastian R. Kruk; Bill McDaniel, Berlin: Springer, 2009, p. 9-39.

6 <http://www.perseus.tufts.edu/hopper/>.

enciclopedia e altre fonti informative strutturate, ma anche di esaminare le edizioni moderne in lingua originale o in traduzione delle opere consultate nella biblioteca digitale. Infine, è sempre possibile usufruire dell'analisi sintattica e dell'identificazione dei personaggi e dei luoghi citati in un documento[7].

Per quanto riguarda le funzionalità della ricerca, gli utenti di Perseus hanno a disposizione un insieme di strumenti in tutte le lingue offerte dal sistema, che permettono la ricerca per parole e per frasi nei singoli testi e nelle collezioni di testi. La ricerca per parole prevede un'opzione che consente di ritrovare tutti i documenti nei quali la stessa parola si presenta anche nelle forme flesse, rafforzando le possibilità di ritrovamento di testi in lingue ricche dal punto di vista morfologico, come il Greco, il Latino e l'Arabo[8].

Come si può constatare, questo modello di biblioteca digitale è fondamentalmente basato sui contenuti, tuttavia, a corredo dei documenti sono stati introdotti servizi e strumenti avanzati, in particolare a supporto dell'analisi testuale, che contribuiscono ad identificarlo come un modello che costituisce un ponte tra le biblioteche di prima e di seconda generazione.

2. Le biblioteche digitali della seconda generazione

Le trasformazioni in atto, dalle biblioteche digitali realizzate durante il primo periodo alle biblioteche digitali della seconda generazione, riguardano sostanzialmente il passaggio da una impostazione incentrata sui contenuti (*content-centric*) per lo più orientata all'agevole fruizione del patrimonio culturale e più o meno arricchita con strumenti dedicati all'elaborazione e all'analisi dei contenuti stessi, ad un'impostazione decisamente incentrata sui servizi agli utenti (*person-centric*).

In particolare, i requisiti richiesti alle singole biblioteche digitali per essere considerate appartenenti alla seconda generazione, riguardano la possibilità di disporre di servizi specializzati per la ricerca a seconda dei diversi media presenti, la possibilità di ricerca attraverso query che presentano molteplici caratteristiche e l'uso del *relevance feedback*[9], oltre ai servizi di indicizzazione, ai servizi di annotazione, di gestione dei metadati, di gestione dei contenuti e di gestione delle *resources*, cioè tutte le entità e le relazioni gestite dalla biblioteca digitale.

In aggiunta alla realizzazione dei servizi localmente offerti da un *repository*, è ritenuto requisito imprescindibile la possibilità di disporre di servizi specializzati tra diversi *content providers*. Si possono comprendere tra queste caratteristiche l'indicizzazione realizzata senza alcun controllo centrale, la gestione distribuita dei metadati, la gestione di servizi eterogenei distribuiti e autonomi tra loro (per il calcolo intensivo, con possibilità di incremento *on demand*), la possibilità di realizzare personalizzazioni e un alto grado di scalabilità, cioè la capacità di mantenere lo stesso

7 <http://www.perseus.tufts.edu/hopper/research>.

8 <http://www.perseus.tufts.edu/hopper/opensource>.

9 Tecnica definita nell'ambito dell'*Information retrieval*, volta a permettere agli utenti di riformulare richieste di documenti sulla base del riconoscimento della rilevanza delle risposte ottenute. Il *Relevance feedback* di tipo esplicito è in sostanza una tecnica di *query expansion*, cioè l'espansione automatica dei termini usati nella ricerca, basata su algoritmi che rilevano l'equivalenza rispetto a un documento ritrovato e definito esplicitamente rilevante dall'utente, mentre il *Relevance feedback* implicito si basa sul monitoraggio del comportamento degli utenti nella ricerca, senza considerare il giudizio di rilevanza degli utenti.

livello di qualità nel caso in cui le richieste dell'utenza o la mole dei contenuti possano crescere o decrescere in scala[10].

Le funzionalità della ricerca predisposte da una tipologia avanzata di biblioteca digitale, permettono di ricercare documenti sia attraverso i metadati, le annotazioni degli utenti e i social tags, per ritrovare testi, immagini e files sonori, sia attraverso l'analisi del full-text. Tra queste funzionalità di livello superiore esiste la possibilità di espandere i termini della ricerca usando i thesauri e le ontologie, ma anche i sistemi di classificazione biblioteconomica esistenti, insomma l'intero complesso dei Knowledge Organization Systems (KOS), in modo da assistere gli utenti nelle ricerche all'interno dei documenti e sui documenti.

L'elemento innovativo, e la sfida più ambiziosa, tuttavia, sono costituiti principalmente dalla possibilità di integrare in un unico spazio informativo i documenti gestiti da biblioteche, archivi, musei e data bases esterni, quindi di permettere l'effettuazione di ricerche in sistemi diversi, sfruttando l'interoperabilità sintattica e semantica, e di consentire anche la ricerca multilingue.[11]

Gli obiettivi proposti vengono raggiunti oggi in gran parte grazie all'uso dell'architettura *Service-oriented* (SOA) e delle reti *peer-to-peer* (P2P) e Grid.

Service-oriented Architecture è un modello organizzativo dei software basato sull'uso di singoli componenti, o moduli, che permette il riuso delle funzionalità di un programma come 'servizi' per altri programmi. La tecnica di progettazione di programmi modulari adotta la strategia della separazione delle funzionalità di un programma in moduli indipendenti, ciascuno dei quali contiene un certo numero di processi atti ad eseguire solo un aspetto della funzionalità complessiva desiderata. In un sistema modulare i diversi componenti separati costituiscono un programma eseguibile che su richiesta attiva le specifiche funzionalità dei moduli disponibili e permette anche l'utilizzazione di più moduli contemporaneamente.

Le diverse funzionalità di un software, denominate 'servizi', sono le unità di lavoro realizzato da un *service provider* per raggiungere i risultati desiderati dal *service consumer.* Grazie all'architettura modulare, queste possono essere riutilizzate come 'servizi' per altri programmi. I 'servizi' sono unità funzionali definite, ciascuna indipendente dallo stato delle altre. Interagiscono tra loro attraverso interfacce semplificate che utilizzano la modalità richiesta/risposta.

SOA è uno stile architetturale che permette ai componenti dei software che interagiscono di mantenere un basso livello di dipendenza, *loose coupling*[12]. Nell'ambito dell'architettura dei sistemi, che si occupa di definire l'organizzazione dei moduli e delle interfacce, i sistemi in cui ciascun componente ha scarsa o nessuna conoscenza della definizione degli altri componenti sono detti *loosely coupled.* La necessità di adottare sistemi con basso livello di dipendenza reciproca dei componenti si è messa in evidenza in ambito aziendale per il bisogno di adattarsi più agilmente ai cambia-

10 Maristella Agosti [*et al.*], *Evaluation and comparison of the Service Architecture, P2P, and Grid approaches for DLs* (Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006), DELOS A Network of Excellence on Digital Libraries, 2006, p. 5, <http://dbis.cs.unibas.ch/static/delos_website/D1.1.1%20-%20Evaluation%20and%20Comparison%20of%20the%20Service%20Architecture,%20P2P,%20and%20Grid%20Approaches%20for%20DLs.pdf >.

11 Dagobert Soergel, *Digital libraries and knowledge organization*. In: *Semantic* d*igital libraries*, cit., p. 9-39.

12 *Coupling* esprime il grado con cui ciascun modulo di un programma dipende da ciascuno degli altri moduli. Per la definizione: <http://en.wikipedia.org/wiki/Loose_coupling>.

menti determinati dall'esistenza di partner diversi e a situazioni diverse di mercato. Grazie all'architettura *service-oriented*, in una rete cooperativa di computer, i servizi possono essere più facilmente combinati tra loro, e questo agevola la capacità delle applicazioni di essere disponibili ad una grande massa di utenti, attraverso servizi e interfacce specializzati e ben definiti.

In uno scenario in cui una biblioteca digitale non si configura più solo come un luogo di archiviazione di dati, ma come uno strumento giornaliero di lavoro inserito nel processo della produzione intellettuale, l'architettura SOA viene utilizzata per supportare, ad esempio, le annotazioni realizzate dagli utenti, utili per chiarire i significati delle risorse informative. Le annotazioni sono il mezzo attraverso cui gli utenti interagiscono con la biblioteca digitale e permettono loro di creare naturalmente un ipertesto che collega i contenuti della biblioteca digitale con i contenuti personali. Necessitano di un'architettura flessibile e beneficiano dell'organizzazione modulare che permette infatti di aggiungere facilmente nuove funzionalità senza dover ridisegnare l'architettura, e consente l'utilizzazione diffusa delle annotazioni da parte di vaste comunità di utenti[13].

Prendiamo in considerazione nei dettagli alcuni di questi sistemi. Fedora (Flexible Extensible Digital Object Repository Architecture), realizzato dalla Cornell University tra 1997 e 2002 col sostegno di DARPA[14], poi divenuto Open source Project dopo il 2002, con il contributo di Andrew W. Mellon Foundation, è un sistema per la gestione di documenti complessi e multimediali, immagini digitali e metadati. Assunto da Cornell University e da Virginia University come sistema per la gestione di archivi e biblioteche digitali, Fedora ha continuato ad essere sviluppato fino a presentare oggi un'architettura *service-based* e risultare, quindi, particolarmente adatto per la gestione di biblioteche, archivi, *repository*, pubblicazioni on line e piattaforme per l'apprendimento. Fedora presenta un'architettura estensibile per la gestione e la disseminazione di oggetti complessi e per la rappresentazione delle relazioni esistenti tra di essi[15].

Flessibilità massima è anche la caratteristica dell'architettura di OpenDLib, Digital Library Management System curato da ISTI-CNR di Pisa[16]. Basato su *Service-oriented Architecture*, consiste di una *open federation of services*, che possono essere distribuiti e replicati su diversi server. La flessibilità si evidenzia nella facilità di aggiungere nuove classi di servizi. L'interoperabilità è assicurata a livello dei metadati con le altre Digital Libraries compatibili all'interno di OAI-PMH. OpenDLib è un *Digital Library System* per gestire biblioteche digitali, capace di venire incontro alle necessità di comunità eterogenee: ad esempio, può gestire oggetti informativi di diversa tipologia, multimediali e multilingue, organizzati in insiemi di collezioni, ciascuna con le sue *policies* per l'accesso, a seconda della tipologia[17]. La nuova versione, OpenDLibG, può gestire oggetti

13 Maristella Agosti [*et al.*], *Evaluation and comparison of the Service Architecture, P2P, and Grid approaches for DLs*, cit., p. 6-10 e 20-23.

14 Defense Advanced Research Projects Agency (DARPA) è l'Agenzia del Dipartimento della difesa degli Stati Uniti che si occupa di ricerca e sviluppo tecnologico.

15 <http://www.fedora.info/>.

16 <http://opendlib.iei.pi.cnr.it/home.html>.

17 Leonardo Candela [*et al.*], *OpenDLib: a digital library service system,* <http://www.researchgate.net/profile/Leonardo_Candela/publication/200462040_OpenDLib_A_Digital_Library_Service_System/file/>.

complessi, particolari tipi di immagini, video, oggetti 3D, sfruttando la possibilità di integrazione con infrastrutture Grid e un ambiente di computo distribuito[18].

I sistemi *peer-to-peer* (P2P) consentono l'interazione tra singoli *providers* di servizi indipendenti tra loro. Sono sistemi distribuiti, organizzati in nodi (*peers*) che possono comunicare direttamente l'uno con l'altro. L'organizzazione può seguire una architettura completamente decentralizzata che prevede che ciascun *peer* nella rete funga sia da client che da server, senza alcun server centrale, oppure un'architettura ibrida, in cui si fondono le caratteristiche della organizzazione decentralizzata e quelle dell'architettura client/server, e che prevede alcuni servizi centralizzati e altri basati sul modello di comunicazione *peer to peer*.

Il modello ibrido può anche presentare un'architettura gerarchica, o architettura *super-peer*, in genere usata nel caso di cooperazione inter- e intra-istituzionale. Un *super-peer* è un *peer* che agisce come un server per un gruppo di *peer* ordinari, in genere organizzati in raggruppamenti (*clusters*) che possono coprire anche istituzioni diverse, e interagisce con gli altri *super-peer*. Le diverse biblioteche digitali, con le loro collezioni e i loro utenti, costituiscono i singoli *peer*[19].

Un sistema dinamico e flessibile, adatto a permettere la condivisione della conoscenza nel campo del patrimonio culturale, è particolarmente utile nel caso della gestione dei patrimoni culturali appartenenti ad istituzioni diverse. L'architettura SOA, insieme all'uso della rete *peer-to-peer*, è al fondamento anche dell'infrastruttura realizzata dal progetto europeo lanciato nel 2004 Building Resources for Integrated Cultural Knowledge Services - BRICKS[20], nell'ambito del *Sixth Framework Programme*, per la realizzazione dell'accesso integrato alle risorse distribuite nel settore del *Cultural Heritage*. Sono state sviluppate soluzioni software open source per la condivisione degli oggetti culturali e la gestione di biblioteche digitali. Obiettivo è facilitare l'integrazione delle risorse digitali esistenti in una digital library comune e condivisa, che può riguardare quindi sia archivi digitali, sia musei digitali, sia altre tipologie di realizzazioni di memorie digitali. La rete è costituita da *BNodes*, ciascuno corrispondente ad una istituzione, che comunicano in Internet. Ciascun nodo sviluppa le funzionalità necessarie alla propria istituzione, ad esempio le annotazioni, e assicura l'interoperabilità tra metadati diversi utilizzando la mappatura con l'ontologia CIDOC[21]. Per le funzioni di ricerca sono previste interrogazioni integra-

18 Leonardo Candela [*et al.*], *OpenDLibG: Extending OpenDLib by exploiting a gLite Grid infrastructure*. In: *Research and advanced technology for digital libraries,* Proceeding of the 10TH European Conference, ECDL 2006. Alicante, Spain, September 2006. Editors Julio Ponzalo [*et al*.]. Berlin: Springer-Verlag, 2006, p. 1-13.

19 Ludger Bischofs [*et al.*], *Adaptive replication strategies and software architectures for peer-to-peer systems*. In: *Future digital library management systems: system architecture and information access.* 8th DELOS thematic workshop Schloss Dagstuhl, Germany March 29 – April 1, 2005. Editors: Yannis Ioannidis, Hans-Jörg Schek, Gerhard Weikum. DELOS: a network of excellence on digital libraries, <http://www.delos.info/files/pdf/Proceedings/Dagsthul_2903_010405/delos-dagstuhl-handout-all.pdf>.

20 <http://www.brickscommunity.org/>.

21 <http://www.cidoc-crm.org/index.html.>, CIDOC CRM *Conceptual Reference Model* è una ontologia formale, elaborata da CIDOC (*International Committee for Documentation*) e ICOM (*International Council of Museums*) allo scopo di facilitare l'integrazione e lo scambio delle informazioni in ambienti eterogenei del patrimonio culturale. E' standard ISO 21127 dal settembre 2006.

te su tutti i tipi di risorse[22]. Al progetto hanno aderito, tra gli altri, la Galleria degli Uffizi, l'Archivio segreto apostolico vaticano, il Consorzio Forma, l'Università di Firenze e, nel 2007, l'ICCU.

L'architettura *peer-to-peer* permette a ciascuna biblioteca digitale di agire come un *peer* in una rete federata di istituti, attuando anche la ricerca federata tra biblioteche digitali e altre fonti informative. Le reti P2P permettono la ricerca collaborativa tra biblioteche digitali, che possono gestire ampie quantità di dati in modo autonomo e consentono ai motori di ricerca di beneficiare degli input di tipo intellettuale degli utenti, *bookmarks*, *logfile* delle *query* e *clickstream*.

L'architettura progettata per il sistema Minerva (Max-Plank Institut für Informatik), ad esempio, facilita la ricerca federata tra un gran numero di biblioteche digitali, attraverso la selezione dei *peer* sulla base di misure della stima della rilevanza di una particolare biblioteca digitale rispetto ad una query. Le misure di rilevanza, calcolate in base a computi statistici, sono usate per stabilire quali biblioteche digitali siano più promettenti e adatte a rispondere ad una determinata query. Le query possono essere eseguite prima localmente in una biblioteca iniziale, usando tecniche di *feedback* implicito e di *query expansion* automatica (tecniche di Information retrieval per *relevance feedback*, o tecniche basate su *query logfile* e *clickstream*). Nel caso in cui la risposta di una determinata biblioteca sia ritenuta insoddisfacente dall'utente, il sistema fornisce una lista di potenziali *peer* utili, la query viene reindirizzata alle biblioteche digitali federate, eseguita sugli indici locali di ciascuna, e i risultati vengono infine combinati in una unica lista. Inoltre, attraverso l'impiego di *user recommendations*, attraverso *bookmarks* locali, che rappresentano raccomandazioni forti per singoli documenti, le query possono essere indirizzate a biblioteche tematicamente rilevanti, che presentano un'utenza dagli interessi simili, per migliorare le possibilità di ritrovare pagine 'soggettivamente' rilevanti[23].

Le potenzialità di computo e archiviazione di un'infrastruttura GRID, infrastruttura di hardware e software capace di coordinare la condivisione delle risorse e la risoluzione dei problemi, possono essere utilizzate per strutturare funzioni tipiche delle biblioteche digitali come la ricerca, la visualizzazione dei documenti, la realizzazione delle annotazioni, la personalizzazione dei servizi, e per creare un'infrastruttura di conoscenza che permetta ai membri di comunità di utenti di produrre *on-demand transient digital libraries*. Queste si costituiscono nel momento in cui un utente autorizzato ne ha bisogno, sfruttando risorse condivise, capaci di soddisfare i loro bisogni, come *repositories* di contenuti, software, servizi, possibilità di computo. Il progetto Diligent (Digital Library Infrastructure on Grid Enabled Technology)[24], coordinato da ISTI-CNR all'interno del *Sixth Framework Programme*, e che coinvolge 14 partner europei, ad esempio, ha inteso sviluppare un'infrastruttura capace di far collaborare differenti organizzazioni che, basandosi sulla tecnologia Grid, potessero accedere alla condivisione controllata delle risorse, ai servizi e alle

22 Bernhard Haslhofer; Predrag Kne evi , *The BRICKS digital library infrastructure*. In: *Semantic digital libraries*, cit., p. 151-161.

23 Matthias Bender [*et al.*], *Challenges of distributed search across digital libraries*. In: F*uture digital library management systems: system architecture and iInformation access.* 8th DELOS thematic workshop (Schloss Dagstuhl, Germany March 29 – April 1, 2005). Pisa, ISTI-CNR, 2005.

24 I servizi offerti dal sistema DILIGENT si basano sulle possibilità computazionali dell'infrastruttura Grid messa a punto dal progetto EGEE (Enabling Grids for E-sciencE) tra 2004 e 2010. (EGEE, <http://www.eu-egee.org>, <http://eu-egee-org.web.cern.ch/eu-egee-org/index.pl?id=134>).

possibilità di computo avanzato, in particolare per la gestione facile e sicura dei documenti multimediali dinamici, con immagini e video, con grafici e tabelle che riportano dati automaticamente aggiornabili. L'infrastruttura Grid costituisce il supporto per la creazione di biblioteche digitali virtuali, *on demand*, basate su risorse e possibilità di computo condivisi, dinamici e modificabili in base a diversi requisiti. L'utilizzazione di questa biblioteca virtuale potrà essere rivolta a comunità diverse di utenti autorizzati[25].

3. Biblioteche digitali con funzionalità 'semantiche'

Un ulteriore avanzamento nel processo di trasformazione delle biblioteche digitali è determinato dall'apporto di funzionalità 'semantiche', che le caratterizzano e le differenziano dalle biblioteche digitali della seconda generazione. L'uso degli strumenti impiegati nei social network e l'adozione degli elementi costitutivi del Semantic Web, sono infatti le caratteristiche peculiari delle biblioteche digitali dell'ultima generazione.

In questo caso, ad esempio, l'accesso ai contenuti delle biblioteche digitali attraverso tecnologie che sfruttano le possibilità offerte dalle annotazioni semantiche realizzate dagli utenti, già adottato dalle biblioteche digitali di seconda generazione, potrà essere potenziato dall'adozione di canali di ritrovamento basati su sistemi che permettono la ricerca in base a criteri di similarità e usano il *collaborative filtering*. Si tratta di una tecnica basata sull'uso di algoritmi per calcolare il grado di similarità degli interessi di vaste comunità di utenti. Si attribuiscono pesi alle diverse scelte adottate nelle ricerche e si elaborano modelli in base ai quali fare previsioni sulla rilevanza dei documenti per una categoria di utenti.

Il modello della *Semantic Digital Libray* prevede alcuni elementi caratterizzanti e imprescindibili:

– uso di RDF (Resource Description Framework)[26] come comune denominatore per rappresentare i metadati, allo scopo di rendere possibile l'interoperabilità con altri sistemi, non necessariamente solo biblioteche digitali, al livello dei metadati e delle potenzialità comunicative,

– possibilità di integrare informazioni basate su metadati che provengono da fonti diverse, come descrizioni bibliografiche, vocabolari controllati, ma anche profili degli utenti e bookmarks,

– predisposizione di interfacce per l'utenza che non siano soltanto *user friendly*, ma siano dotate di sistemi a supporto della ricerca e del browsing che, oltre ad utilizzare i dati frutto dell'indicizzazione semantica proveniente dalle biblioteche (*legacy data*), siano rafforzati dall'uso di annotazioni semantiche realizzate dagli utenti e, quindi, sfruttino le potenzialità dei social networks,

– offerta di interfacce per la ricerca che siano basate sui profili dell'utenza, siano fornite di motori di ricerca dotati di funzionalità di 'ragionamento' e siano dotate di *recommendation systems*[27].

Quest'ultimo elemento, cioè l'impiego di motori arricchiti con funzionalità avanzate, costituisce una delle sfide più impegnative che attendono le biblioteche digi-

25 Maristella Agosti [*et al.*], *Evaluation and comparison of the service architecture, P2P, and Grid approaches for DLs*, cit., p. 28.

26 <http://www.w3.org/RDF/>.

27 Sebastian R. Kruk; Bill McDaniel, *Goals of semantic digital libraries*. In: *Semantic digital libraries*, cit., p.71-76.

tali dell'ultima generazione. *Reasoning engines* sono componenti di software capaci di realizzare inferenze, cioè derivare conseguenze logiche da un insieme di assiomi, asserzioni dichiarate in una ontologia costruita per il Semantic Web.[28] Rientrano nella categoria dei software agenti, programmi con un certo grado di autonomia, realizzati nell'ambito dell'Intelligenza artificiale, che sono in grado di apprendere dal comportamento dell'utente e di ricavare nuova conoscenza dalle relazioni assiomatiche istituite in una ontologia. Interrogano le ontologie in rete per cercare le informazioni, disambiguare i significati e 'comprendere' logicamente i messaggi.[29]

Tra le realizzazioni che possono rientrare in questa categoria, possiamo citare come esempio il motore di ricerca OntoLook, progettato dalla cinese Huazhong University of Science and Technology[30] e recentemente ripreso e sviluppato da ricercatori indiani con il nome di Semantic Look[31]. L'elemento innovativo è costituito dalla possibilità di individuare e utilizzare le relazioni tra i concetti istituite dalle ontologie. Viene applicato nella ricerca delle pagine Web partendo dalle relazioni tra i concetti presenti nella query dell'utente. Il motore analizza i termini della query e le relazioni che esistono tra essi e raggruppa i termini-concetti in coppie. Queste vengono successivamente utilizzate per interrogare l'ontologia collegata al sistema e scoprire tutte le relazioni che l'ontologia dichiara per quei concetti negli assiomi istituiti. Vengono infine scelte automaticamente le relazioni tra i termini da utilizzare nella ricerca. I risultati saranno più precisi e meno 'rumorosi'. Se applicato a biblioteche digitali, può costituire la base per ricerche mirate all'interno dei contenuti dei documenti col supporto delle ontologie.

Le biblioteche digitali di ultima generazione devono essere in grado di gestire oggetti informativi complessi, risorse *streaming*, oggetti statici e dinamici e contenuti eterogenei. Gli utenti dovrebbero poter usufruire di informazioni interconnesse che riguardano le diverse risorse gestite dalla biblioteca digitale, attuando il browsing, o usando i filtri predisposti per la ricerca, o compiendo la ricerca di oggetti che presentano elementi in comune[32]. Le biblioteche digitali devono inoltre poter sod-

28 La letteratura sulle ontologie per il Semantic Web è sterminata e non è possibile citarla qui. Rinvio al sito dell'Istituto di scienze e tecnologie della cognizione del CNR di Trento, in particolare al Laboratory for Applied Ontology e ai lavori di Nicola Guarino, uno dei massimi esperti italiani di ontologie, <http://www.loa.istc.cnr.it/>. Utili per una prima informazione i lavori di Maria Teresa Pazienza, della quale in italiano si può leggere *Ontologie e Web semantico: proprietà e problematiche connesse al loro uso diffuso*. In: Numero speciale monografico su Le ontologie, a cura di Maria Teresa Biagetti, «AIDA Informazioni», 28 (2010), n. 1-2, p. 33-61.

29 Sulla rappresentazione della conoscenza, sulla Logica del primo ordine, e per un approfondito esame delle diverse tipologie di agenti intelligenti e delle possibilità di 'ragionamento' attraverso le inferenze, è molto utile leggere Stuart J. Russell; Peter Norvig, *Intelligenza artificiale: un approccio moderno,* 2. ed., Milano: Pearson, 2005. (Terza edizione aggiornata, a cura di Francesco Amigoni, Milano: Pearson, 2010).

30 Yufei Li; Yuan Wang; Xiaotao Huang, *A relation-based search engine in Semantic Web*, «IEEE Transactions on knowledge and data engineering», 19, (2007), n. 2, p. 273-282.

31 Leena Giri G. [*et al.*], *Mathematical model of semantic look. An efficient context driven search engine*, «arXiv»:1402.7200v1 [cs.IR], submitted 28 Feb 2014, < http://arxiv.org/abs/1402.7200v1>.

32 Per una prima informazione è utile anche il *Tutorial on semantic digital libraries* di Sebastian R. Kruk [*et al.*], <http://www.slideshare.net/skruk/tutorial-on-semantic-digital-libraries-eswc2007#>.

disfare le necessità dei lettori formatisi nell'era di Internet, abituati a interagire con una molteplicità di altri utenti, a non lavorare da soli, ma a sentirsi parte di una comunità. Soprattutto per questo, esse devono essere corredate di servizi basati sulle potenzialità del Semantic Web e sulle soluzioni offerte dal Social Networking.

Se consideriamo i metadati, le biblioteche digitali 'semantiche' si distinguono dalle altre biblioteche digitali perché usano metadati aperti, non strutturati e altamente interconnessi. L'uso di metadati aperti e destrutturati dovrebbe offrire alle biblioteche digitali dell'ultima generazione la possibilità di adottare soluzioni nuove a favore della ricerca da parte degli utenti.

Le soluzioni proposte per aiutare gli utenti nella ricerca, per il momento, sono: estensione della navigazione a faccette, interazione basata sulla lingua naturale, *collaborative filtering* basato sulla comunità degli utenti, ad esempio utilizzando Friend-of-a-friend (FOAF). Gli utenti, inoltre, possono utilizzare interfacce flessibili, modificabili a seconda della tipologia di fruizione. Il sistema riconosce gli utenti come appartenenti ad una categoria, e alcune funzioni possono essere disabilitate automaticamente, mentre altre lo saranno a discrezione dell'utente stesso. I profili degli utenti possono essere catturati con una speciale ontologia che ne descrive le caratteristiche, i gruppi cui appartengono e le relazioni intrattenute.

L'architettura *service-oriented* stabilisce l'organizzazione dei diversi servizi, tra cui l'accesso ai contenuti da parte degli utenti finali, sia esseri umani che macchine, le modalità di presentazione dei dati attraverso RDF e il supporto ontologico per definire i significati e le relazioni tra i concetti, il controllo delle politiche di accesso e la validazione dei dati che entrano nella biblioteca digitale. Nell'ambito di questa architettura, il settore dedicato specificamente alla gestione del trattamento dei dati comprende quattro sottosezioni:

– Servizi di ricerca delle informazioni (access, search, browsing), con i quali interagiscono gli utenti: interfacce che usano il linguaggio naturale, possibilità di navigazione attraverso faccette, utilizzo delle annotazioni di altri utenti, uso del *ranking* e del *collaborative filtering*, insieme all'impiego dei profili degli utenti.
– Servizi di gestione avanzata, per i bibliotecari, per la gestione degli utenti e delle collezioni, ma anche dell'insieme di ontologie utilizzate.
– Servizi di gestione dei contenuti e dei metadati, che assicurano la gestione dell'accesso e il controllo delle politiche d'accesso. I bibliotecari possono fruire di servizi che li assistono nell'indicizzazione e nella classificazione.
– Servizi per l'interoperabilità, che non coinvolgono l'utenza, e offrono diverse soluzioni: dall'esposizione dei metadati all'uso del metadata *harvesting*, ai servizi di mediazione di metadati[33].

Le infrastrutture di rete *peer-to-peer* (P2P) e Grid costituiscono il supporto tecnologico alla biblioteca digitale di ultima generazione. L'infrastruttura P2P, senza controllo centralizzato, è impiegata per la condivisione delle risorse che risiedono su macchine diverse, i singoli *peer*, e permette la ricerca delle risorse, l'accesso e la disseminazione delle risorse stesse.

Le annotazioni semantiche delle risorse mantenute dalle biblioteche digitali semantiche possono contribuire a rendere più potente l'infrastruttura P2P. Una delle strategie adottate per migliorare l'efficienza dell'infrastruttura P2P è infatti la creazione di reti logiche che permettono di raggruppare le reti fisiche, modellate attraverso l'uso di classifi-

33 Sebastian R. Kruk; Adam Westerki; Ewelina Kruk, *Architecture of semantic digital libraries*. In: *Semantic digital libraries*, cit., p. 77-85.

cazioni e gerarchie di concetti generati automaticamente, impiegando algoritmi di *clustering*, o in modo semi-automatico, sfruttando i tag semantici degli utenti. Obiettivo principale delle reti logiche è il miglioramento della ricerca delle risorse nella rete di *peer*.

Semantic Overlay Networks (SON) si offrono come reti logiche che permettono di strutturare un sistema di *peer*, o nodi, che conservano contenuti simili dal punto di vista semantico, e nel quale le query vengono trattate in modo da permettere il massimo livello di ritrovamento. La rete logica permette di gestire dati eterogenei, offrendo un supporto per la correlazione di schemi diversi di dati che siano simili dal punto di vista semantico, ad esempio, modi diversi di rappresentare un riferimento cronologico (*semantic mediation*). Alla base del sistema di *Semantic Overlay Networks* vi sono tecniche di classificazione dei 'nodi', cioè dei contenitori di documenti, fondate su gerarchie di concetti, che permettono di prendere in considerazione solo i documenti che rientrano nella categoria concettuale richiesta o che presentano un legame gerarchico di tipo ascendente o discendente con l'oggetto della ricerca[34].

In realtà, queste organizzazioni gerarchiche, a mio avviso, soffrono di tutti i limiti che le classificazioni biblioteconomiche tradizionali manifestano, dal momento che i documenti spesso trattano una molteplicità di argomenti, difficilmente incasellabili in un'unica categoria. Il loro impiego sembra orientato soprattutto a rendere più efficiente la ricerca nei sistemi P2P in termini di tempo impiegato nel recupero di documenti rilevanti, mentre non sembra che il grado di *precision* dei documenti possa beneficiarne in modo determinante.

La rete Grid è usata invece per trattare dati pesanti, realizzare ad esempio i servizi di indicizzazione, che coinvolgono il trattamento del linguaggio naturale, *reasoning* ed estrazione automatica delle caratteristiche dei contenuti multimediali[35].

Particolare rilievo viene dato non solo all'uso delle annotazioni realizzate dagli utenti, ma anche alle funzionalità che devono permettere il raffinamento delle *queries*, basato sui profili degli utenti, e all'ampliamento dei risultati della ricerca attraverso l'impiego di *recommendation systems* che sfruttano le annotazioni delle risorse o sul *collaborative filtering* degli utenti, che in tal modo assumono il ruolo centrale di collaboratori.

Sviluppati alla metà degli anni '90 dall'interazione tra Information retrieval e Scienze cognitive, i sistemi di *recommendation* di tipo collaborativo o di *collaborative filtering*, invece di basarsi sui profili degli utenti e sulle loro preferenze, esplicitamente dichiarate o implicitamente desunte attraverso il tracciamento delle ricerche, usano le intere collezioni di giudizi di *rating* forniti dagli utenti per apprendere un modello, e applicare stereotipi per elaborare le previsioni di *rating*. Vengono applicati anche modelli statistici, ad esempio i modelli Bayesiani, e complessi modelli probabilistici, come quelli basati sui processi decisionali Markoviani. L'elemento migliorativo è costituito dal fatto che, rispetto ai sistemi di *recommendation* basati sul contenuto (*content-based*), che si limitano a suggerire documenti che hanno un grado di similarità con documenti già conosciuti dall'utente di cui è noto il profilo, i sistemi collaborativi si basano sul giudizio di *rating* di tutti gli utenti. Possono quindi suggerire qualsiasi item, e favorire l'ampliamento delle ricerche, invece di incoraggiare esclusivamente la specializzazione[36].

Tra i *Semantic Digital Library System* che hanno predisposto anche funzionalità

34 Arturo Crespo; Hector Garcia-Molina, *Semantic overlay networks for P2P systems*. Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003, < http://ilpubs.stanford.edu:8090/627/1/2003-75.pdf>.

35 Sebastian R. Kruk; Adam Westerki; Ewelina Kruk, *Architecture of semantic digital libraries*., cit.

36 Khan Munnezzah; Nair Sreeja, *A survey of content-based recommendation systems in a nutshell*. «International journal of advanced research in computer science and electronics engineering (IJARC-

avanzate di tipo sociale, possiamo prendere in considerazione JeromeDL e Greenstone. JeromeDL è stato sviluppato da Digital Enterprise Research Institute (DERI) della National University of Ireland, Galway, e da Gdansk University of Technology (Polonia). In questo caso, le tecnologie del Semantic Web, RDF, FOAF e ontologie, vengono usate insieme alle possibilità offerte dai *social network*. Jerome DL supporta differenti standard descrittivi (Dublin Core, BibTeX e MARC21), tuttavia le descrizioni basate su questi non sono facilmente utilizzabili dalle macchine, per le quali è necessario il supporto di formalismi. Le risorse vengono indicizzate utilizzando tutti i possibili KOS (Knowledge Organization Systems), quindi vocabolari controllati, sistemi di classificazione, tesauri; tuttavia, si dichiara di preferire l'uso di *folksonomies*, dei sistemi di *tagging*, e le annotazioni realizzate dagli utenti. Gli elementi descrittivi messi in evidenza sono quelli adottati negli usuali formati bibliografici, come MARC21 e Dublin Core. Nell'ambito degli studi realizzati per JeromeDL è stata definita una ontologia minima per organizzare le relazioni tra le risorse e i loro autori/creatori, le loro parti, i topici trattati definiti attraverso parole chiave, gli eventi e le persone citate etc.

Come strumento per l'intermediazione tra i diversi formati è stata sviluppata invece MarcOnt Ontology che, partendo dagli elementi presenti nei formati bibliografici comuni (MARC21, BibTex, Dublin Core), crea un nuovo standard descrittivo in forma di ontologia per le biblioteche digitali, per migliorare il livello di interoperabilità, scritta in OWL DL[37]. Ad esempio, le descrizioni in MARC21 possono essere trasformate in descrizioni semantiche MarcOnt creando file MARC-XML e convertendoli poi in file MARC-RDF. Adottare formati ampiamente utilizzati, insieme ad ontologie bibliografiche, è ritenuto sufficiente a creare una descrizione semantica delle risorse.

Uno degli obiettivi di JeromeDL è permettere la ricerca in fonti bibliografiche diverse e su un'ampia varietà di risorse, testi, documenti multimediali, risorse in PDF. L'algoritmo usato per la ricerca si snoda attraverso tre fasi: ricerca nel full text delle risorse e delle annotazioni degli utenti circa le risorse; ricerca nelle descrizioni bibliografiche nei formati MARC21 e BibTeX; ricerca *user-oriented*, che si basa sulle descrizioni 'semantiche', in particolare utilizzando le informazioni che provengono dalle categorie più interessate, e usa l'espansione della *query* fondata su *social collaborative filtering*, definiti secondo gli interessi degli utenti con il supporto di FOAFRealm[38]. FOAFRealm consente l'interconnessione degli utenti registrati, permette

SEE)», 3, (2014), n. 1, p. 24-30. Alcune tecniche per realizzare *Collaborative filtering systems*, basate sui coefficienti di correlazione, sulla similarità misurata con gli spazi vettoriali, sui metodi statistici bayesiani, sono stati presentati in John S. Breese; David Heckerman; Carl Kadie, *Empirical analysis of predictive algorithms for collaborative filtering*, UAI '98. Proceedings of the Fourteenth conference on uncertainty in artificial intelligence, San Francisco: Morgan Kaufmann Publishers Inc., 1998, p. 43-52.

37 Marcin Synak; Sebastian R. Kruk, *MarcOnt Initiative–The ontology for the librarian world,* 2005, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.114.3986&rep=rep1&type=pdf>; Sebastian R. Kruk; Marcin Synak; Kerstin Zimmermann. *MarcOnt: integration ontology for bibliographic description formats,* < http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.155.201&rep=rep1&type=pdf>. OWL DL (*Description Logics*) è uno dei sottolinguaggi di OWL (Ontology Web Language), il linguaggio formale per specificare le ontologie realizzato dal Web Ontology Working Group come parte della W3C Semantic Web Activity, <http://www.w3.org/TR/owl-features/>. Nel 2009 è stata rilasciata la seconda edizione: *OWL 2 Web Ontology Language Document Overview* (W3C Recommendation 27 October 2009), <(http://www.w3.org/TR/owl2-overview/>.

loro anche di annotare, dare una valutazione delle risorse, creare bookmarcks e scambiarli. Costituisce uno sviluppo di FOAF in quanto, oltre ad usare RDF per stabilire le relazioni di conoscenza tra le persone (*Friend-of-a-friend*), aggiunge la possibilità di modulare il livello di conoscenza tra le persone, attribuendo un peso al grado del collegamento interpersonale.

Greenstone[39], nella versione 3., è un insieme di software open source sviluppati da New Zeland Digital Library Project insieme all' Unesco per realizzare biblioteche digitali dotate di una struttura semantica. Può essere impiegato per impiantare biblioteche digitali che accolgono contenuti multimediali, o anche per *repositories* scientifici istituzionali. Le funzioni di ricerca sono sostanzialmente limitate a permettere la ricerca delle parole nel full-text, realizzare il browsing attraverso i titoli dei documenti, le organizzazioni e i soggetti. I soggetti offerti, in verità, sono ampie classi, e non permettono l'identificazione precisa degli argomenti: ad es. *History, Cultural heritage and Language* è presentato come soggetto, ma è un insieme di tre classi disciplinari molto vaste.

La struttura semantica impiegata per ampliare le funzionalità delle biblioteche digitali create usando l'insieme di software Greenstone open source, invece, viene identificata con:

a) Il servizio di *alerting* tra diverse biblioteche, attraverso la creazione di profili dell'utenza, basati sui loro bisogni informativi. Un insieme di *federated servers* vengono esplorati per conoscere se un nuovo documento elettronico è stato aggiunto, se un manoscritto è stato digitalizzato, o se una risorsa elettronica è stata sostituita.

b) L'applicazione del modello FRBR, che può essere usato per raggruppare le traduzioni della stessa opera, con titoli differenti, in lingue diverse, o per raggruppare riviste che hanno modificato i loro titoli. Questa caratteristica può contribuire a migliorare il servizio di *alerting*, organizzando pagine dedicate a elencare dati biografici e opere di ciascun autore[40].

La potenzialità semantica è limitata alla creazione dei profili degli utenti, in base ai quali le risorse informative vengono filtrate in modo da realizzare servizi di *alerting* finalizzati a loro, e con l'uso limitato di una ontologia bibliografica, FRBR.

Il sistema Fedora[41], open source, nelle versioni più recenti, dalla 2.0 a quella recentissima 3.7.1 dell'ottobre 2013, prevede un modello di relazioni basato su RDF che permette di rappresentare le relazioni tra le parti componenti degli oggetti, e che lo rende adatto all'uso in biblioteche digitali, archivi e depositi istituzionali. Permette l'aggregazione di materiali provenienti da diverse fonti, l'associazione dei metadati e la gestione delle funzionalità relative alla combinazione di materiali diversi, come

38 Sebastian R. Kruk; Stefan Decker; Lech Zieborak, *JeromeDL – Adding semantic web technologies to digital libraries*. In: *Database and expert systems applications*. Lecture Notes in Computer Science. Vol. 3588, Editors Kim V. Andersen; John Debenham; Roland Wagner. Berlin Heidelberg: Springer (2005), p. 716-725. Sebastian R. Kruk [*et al.*], *JeromeDL: The social semantic digital library.* In: *Semantic digital libraries*, cit., p.139-150. Sebastian R. Kruk; Stefan Decker; Lech Zieborak, *Jerome DL – Reconnecting digital libraries and the semantic web,* <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.3149&rep=rep1&type=pdf>, *WWW2005*, May 10–14, 2005, Chiba, Japan.

39 < http://www.greenstone.org/>. Greenstone Version 3, currently under study, is designed as a distributed network of independent modules.

40 Annika Hinze [*et al.*], *Semantics in Greenstone*. In: *Semantic digital libraries*, cit., p. 163-176.

41 <http://www.fedora.info ; http://fedorarepository.org/software/current>.

testi e immagini. Permette di mettere in evidenza le relazioni, anche di tipo semantico, esistenti tra oggetti digitali in una collezione, e le relazioni strutturali tutto-parte, ad esempio tra un articolo e la rivista sulla quale è pubblicato, o tra un capitolo di una monografia e la monografia stessa.

L'unità di una risorsa può essere suddivisa, creando oggetti digitali separati (ad esempio, capitoli di una vasta monografia) che facilmente potranno essere usati in contesti differenti, ad esempio in una piattaforma per l'e-learning, a seconda delle esigenze; questo permette anche di organizzare più facilmente gli oggetti digitali secondo le prerogative di FRBR. Una delle caratteristiche rilevanti è la possibilità di istituire relazioni tra diverse tipologie di risorse utilizzando una ontologia. L'esempio più significativo è l'impiego di Fedora per la *Encyclopedia of Chicago*[42]. Le voci dell'enciclopedia, firmate e fornite di riferimenti bibliografici, sono corredate, se è necessario, di immagini digitali di documenti d'archivio, che naturalmente si possono ingrandire e sfogliare, far ruotare, e di rappresentazioni di fonti storiche e documenti digitalizzati. Le possibilità di applicazione sono presentate attraverso due esempi di mappe della città relative ad alcuni anni dell'Ottocento (*Rich Maps*) disponibili tra le *Special features* dell'*Encyclopedia* online: spostando il mouse su di una zona di Chicago, l'utente apre link a dati statistici sulla popolazione relativi a quell' anno, articoli pubblicati su riviste in quell'anno, dati storici relativi ad avvenimenti rilevanti, ad esempio lo sciopero dei lavoratori di Chicago dal 25 aprile al 4 maggio 1886.[43]

4. Un modello di riferimento per le biblioteche digitali

Il processo di trasformazione delle biblioteche digitali, comprese quelle arricchite da funzionalità 'semantiche', trova un punto di aggregazione dei diversi orientamenti nell'attività realizzata dal Progetto DELOS. *DELOS Network of Excellence on Digital Libraries* fu approvato da *European Commission* come parte del *Sixth Framework Programme* IST (2002-2006), con l'obiettivo di definire metodologie per l'intero ciclo dell'informazione nelle biblioteche digitali, progettare servizi multimodali, multilingue e interoperativi per la gestione integrata dei contenuti e infine sviluppare tecnologie che potessero sostenere funzionalità avanzate.

L'orientamento assunto da DELOS pone gli utenti al centro delle biblioteche digitali, che sono considerate depositi delle conoscenze universali, aperti alla libera fruizione di chiunque e strumenti atti a facilitare la comunicazione e la collaborazione tra studiosi. Il Report che accoglie le risposte alle interrogazioni poste dalla Commissione europea a Delos (Nizza, 5 e 6 dicembre 2005), oltre ad esporre i principi generali per lo sviluppo del programma di *European Digital Library* (EDL), pone l'accento proprio su questa visione delle biblioteche digitali.

> The DELOS vision is that Digital Libraries will become the universal knowledge repositories and communication conduits for the future, common vehicles by which everyone will access, analyze, evaluate, enhance, and exchange all forms of information. They will be indispensable tools in the daily personal and professional lives of people. They will be accessible at any time and from anywhere, and will offer a friendly, multi-modal, efficient, and effective interaction and exploration environment. Efforts towards this vision require significant changes in the present Digital Library development strategies, with respect to

42 <http://www.encyclopedia.chicagohistory.org/>.

43 Carl Lagoze [*et al.*], *Fedora: an architecture for complex objects and their relationships*, «International journal on digital libraries», 6 (2006) , n. 2, p. 124–138. (DOI 10.1007/s00799-005-0130-3).

> functionality, operational environment, and other aspects. There is the need to overcome the major limitations observed in the development of present-day systems, which are essentially "content-centric" and "one-of-a-kind", i.e. each system has been developed having in mind a specific content, a specific user community, and a specific application[44].

Durante il *Third Brainstorming Meeting* di DELOS, tenutosi nel luglio 2004[45], fu proposto di modificare la denominazione di Digital library in *Dynamic universal knowledge environment* (DUKE), definizione ritenuta maggiormente rappresentativa proprio delle caratteristiche di collaborazione, universalità, cooperazione, della realizzazione dell'interoperabilità e dell'accesso integrato a diversi archivi.

Sembra tuttavia di cogliere in questo orientamento, in particolare laddove si insiste sulla realizzazione di spazi per la collaborazione ed il riuso dei documenti, una sovrapposizione tra strumenti che sono stati progettati in realtà per finalità diverse. *Repository* digitali istituzionali, ad esempio di una università, che offrono la possibilità di condividere report e paper prodotti in campi disciplinari differenti, oppure i risultati parziali di ricerche scientifiche, per permettere ad altri di arricchirli e portarle a compimento, nell'ottica dell'*open access* degli esiti delle ricerche, sono una cosa diversa rispetto ad una biblioteca digitale realizzata sulla base di un progetto culturale specifico, dai precisi contorni e a favore di un'utenza determinata. Proprio le biblioteche digitali pensate, costruite e strutturate secondo un preciso piano culturale e scientifico, manifestano quell'unità di contenuti che spesso le raccolte delle biblioteche fisicamente visitabili non offrono, oltre a presentare un livello di approfondimento degli argomenti per lo più omogeneo. L'intenzionalità del progetto e la definizione di un profilo di utenza costituiscono la differenza tra una biblioteca digitale che 'è una biblioteca' e una biblioteca digitale che è invece un deposito istituzionale aperto.

A partire dal 2005, nell'ambito del Progetto DELOS, è stato sviluppato *The DELOS Digital Library Reference Model*[46], che doveva servire come base e inquadramento concettuale generale attraverso il quale definire le entità e le relazioni collegate all'uni-

44 The DELOS Network of excellence on digital libraries. *Recommendations and observations for a European Digital Library (EDL). In response to the European Commission's call for online consultation*, <http://www.delos.info/files/pdf/events/brainstorming_dec05/DELOSBrainstormingReport_Final.pdf>. Le raccomandazioni presentate per l'attività a lungo e a medio termine, prendono in considerazione molteplici tematiche, dalla necessità di aumentare gli sforzi per evitare duplicazioni di digitalizzazione (già accolti dal progetto MICHAEL), all'opportunità di sviluppare strumenti lessicali e ontologici per collegare tra loro le diverse lingue europee (già presi in considerazione per esempio da MACS, Multilingua Access to Subjects), di sviluppare ontologie per l'integrazione dei metadati, per l'accesso semantico e il ragionamento automatico sui dati; dalla necessità di aumentare le misure di preservazione dei documenti, alla necessità di attuare le opportune misure per unificare l'accesso ai documenti che appartengono a biblioteche, archivi e musei e altre istituzioni che conservano patrimoni culturali e memorie collettive; e infine di assicurare l'interoperabilità a tutti i livelli, dalle infrastrutture tecnologiche ai formati dei dati, fino al livello della interoperabilità dei metadati descrittivi e dei KOS impiegati.

45 DELOS: *Future research directions. 3rd DELOS brainstorming workshop report*. Corvara, Italy, 4-6 July, 2004. Editors Alberto Del Bimbo, Stefan Gradman, Yannis Ioannidis, <http://www.delos.info/files/pdf/events/2004_Jul_8_10/D8.pdf>.

46 Leonardo Candela [*et al.*], *The DELOS Digital Library Reference Model. Foundations for Digital Libraries*. Version 0.98., DELOS Network of Excellence on Digital Libraries: 2007, <http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf>; DL.org. Coordination Action on Digital Library Inter-

verso delle biblioteche digitali, e riuscire a coordinare strategie e sistemi diversi all'interno di un unico modello.

The DELOS Digital Library Reference Model definisce precisamente i tre componenti della Biblioteca digitale, considerata come uno spazio genericamente improntato alla ricerca dell'informazione, senza barriere per quanto riguarda i contenuti, lo spazio fisico o il tempo, e rivolto alla facilitazione e allo scambio delle informazioni.

Digital Libraries (DL) sono le organizzazioni che raccolgono, gestiscono e preservano per lunghi periodi gli oggetti digitali, con le quali, in base a *policies* determinate, entrano in contatto gli utenti finali;

Digital Library Systems (DLS) sono i sistemi software, basati su di una architettura che offre tutte le funzionalità richieste e permette agli utenti di interagire con la DL vera e propria. L'organizzazione del sistema richiede il controllo dei protocolli di comunicazione e di accesso ai dati e della scalabilità.

Digital Library Management Systems (DLMS) sono i sistemi software che consentono la realizzazione di un *Digital Library System* e, inoltre, permettono di integrare altri software per aggiungere funzionalità avanzate. Il DLMS permette di generare e riconfigurare un DLS secondo strategie diverse, consentendo ai gestori del sistema di scegliere le funzionalità più confacenti alle esigenze.

Il Model esplicita le aree concettuali di cui si compone una Digital library:
a) *contents*, b) *users* (esseri umani o macchine), c) *functionalities*, d) *quality levels*, e) *policy rules* (ad esempio la gestione dei diritti, la *privacy*, i DRM), f) *architecture of DLS.* Le sei aree sono definite attraverso 218 Entità e 52 Relazioni. L'interoperabilità coinvolge sia i formati di metadati che le *policies* e i parametri per la qualità, per realizzare la cooperazione tra diverse DL[47].

Le aree concettuali che ricoprono maggiore interesse per noi sono quelle relative al *Content Domain* e al *Functionality Domain*. Il primo copre tutte le forme di informazione che una Biblioteca digitale gestisce per offrire i suoi servizi agli utenti, ed include sia gli oggetti informativi (*Information Objects*), cioè testi, immagini, documenti multimediali, che possono anche essere organizzati in collezioni, sia i metadati e le annotazioni che li descrivono. Se prendiamo in considerazione la tipologia della rappresentazione, gli oggetti informativi possono essere *born digital* oppure prodotti di digitalizzazioni da fonti a stampa, frequenti nel campo delle Scienze umane.

> C6 Content Domain
> Definition: One of the six main concepts characterising the Digital Library universe. It represents the various aspects related to the modelling of information managed in the Digital Library universe to serve the information needs of the *Actors*. [...]
> Rationale: The Content concept represents the information that *Digital Libraries* handle and make available to their *Actors*. It is composed of a set of *Information Objects* organized in *Collections*. *Content Domain* is an umbrella concept that is used to aggregate all forms of information that a *Digital Library* may require in order to offer its services.

operability, Best Practices and Modelling Foundations. Version 1.0. (2009) Candela L (Editor), <http://www.dlorg.eu/uploads/DL%20Reference%20Models/The%20Digital%20Library%20Reference%20Model_v1.0.pdf>. Una sintetica presentazione, in: Leonardo Candela; Donatella Castelli, *Una teoria fondazionale per le biblioteche digitali: il DELOS Digital Library Reference Model*, «Dig*Italia*. Rivista del digitale nei beni culturali», 4 (2009), n. 1, p. 44-82.

47 Leonardo Candela [*et al.*], *The DELOS Digital Library Reference Model. Foundations for digital libraries*., cit., p. 55.

> *Metadata* play an important role in the *Content Domain* because they describe a clearly defined category of *Information Objects* in the domain of discourse[48].

Nel modello i metadati sono definiti in termini piuttosto semplificati e non viene presentata alcuna metodologia relativa alla loro creazione, alcun impianto teoretico e alcuna politica per il loro trattamento.

> C11 Metadata
> Definition: Any *Information Object* that is connected to one or more *Resources* through a <*hasMetadata*> relationship.[...]
> Rationale: Metadata are used for describing different aspects of data, such as the semantics, provenance, constraints, parameters, content, quality, condition and other characteristics. These data can be used in different contexts and for a diversity of purposes; usually, they are associated with an *Information Object* (more generally to a *Resource* through the <*hasMetadata*>) as a means of facilitating the effective discovery, retrieval, use and management of the object[49].

Il *Functionality Domain* riguarda i servizi e gli strumenti realizzati dalla biblioteca digitale per i suoi utenti, sia esseri umani che software, e include le funzioni di ricerca, l'accesso ai documenti, il browsing, la visualizzazione, ecc. *DELOS Digital Library Reference Model* stabilisce tuttavia che ciascuna Biblioteca digitale possa implementare queste funzioni in accordo con le necessità dei propri utenti.

> C30 Functionality Domain
> Definition: One of the six main concepts characterising the Digital Library universe. It represents the various aspects related to the modelling of facilities/services provided in the Digital Library universe to serve *Actor* needs [...]. Rationale: The *Functionality Domain* concept represents the services that *Digital Libraries* offer to their *Actors*. The set of facilities expected from *Digital Libraries* is extremely broad and varies according to the application context. There are a number of Functions that *Actors* expect from each *Digital Library*, e.g. search, browse, information objects visualisation. Beyond that, any *Digital Library* offers additional *Functions* to serve the specific needs of its community of users[50].

Functionality Relations sono limitate a definire l'interazione tra le diverse funzioni, l'influenza di particolari profili di utenti su specifiche funzioni, o la creazione di annotazioni, e anche la relazione di connessione tra la funzione di accesso alle risorse e le risorse che sono state trovate.

Viene riconosciuto ai bibliotecari il compito di coprire diverse tra le funzioni che sono legate alla biblioteca digitale, ad esempio l'importante ruolo di assicurare la qualità dei servizi, che costituisce l'elemento discriminante tra una biblioteca digitale e il Web[51]. Tuttavia, l'attività dei bibliotecari viene considerata tra le attività degli utenti finali della biblioteca digitale, insieme ai creatori e agli utenti dei contenuti

48 *Ivi*, p. 76.

49 *Ivi*, p. 79-80.

50 *Ivi*, p. 90-91.

51 *Ivi*, p. 23 e p. 89.

(*Content Creators*, *Content Consumers* and *Librarians*). Secondo questo modello, i bibliotecari curano i contenuti di una biblioteca digitale e stabiliscono le *policies*, tuttavia non viene affatto preso in considerazione il compito relativo all'indicizzazione e all'arricchimento di metadati, specialmente di quelli semantici, cioè l'organizzazione della conoscenza attraverso i contenuti che una DL offre agli utenti interessati alla ricerca scientifica.

Tra le diverse funzionalità previste in una biblioteca digitale, quella che permette la ricerca, espressa attraverso le *queries* degli utenti, assume particolare rilievo nel modello DELOS. Per quanto concerne le metodologie di ricerca si prende in considerazione la possibilità di compiere ricerche per parole, combinate usando gli operatori booleani, oltre alla possibilità di utilizzare i metadati strutturati. Tuttavia, uno dei punti di maggiore criticità è costituito proprio dal trattamento semplicistico riservato ai metadati:

> Descriptive metadata, i.e. metadata that provide a mechanism for representing attributes describing and identifying the *Resource*. Examples include bibliographical attributes (e.g. creator, title, publisher, date), format, list of keywords characterizing the contents. The term 'descriptive' is used here in a consistent but broader sense than in 'descriptive cataloguing'. [...] Administrative metadata, i.e. metadata for managing a *Resource*, [...] Preservation metadata, i.e. metadata designed to support the long-term accessibility of a Resource by providing information about its content, technical attributes, dependencies, management, designated community(ies) and change history[52].

In particolare, quando vengono presi in considezione i metadati descrittivi, si fa riferimento esclusivamente all'uso delle parole chiave per definire i contenuti degli oggetti informativi, e non si considerano affatto i problemi coinvolti nell'indicizzazione semantica realizzata da esseri umani e in quella semiautomatica.

L'argomento relativo alla ricerca semantica non viene affrontato con l'attenzione e l'approfondimento che sarebbero stati necessari; la scarsa sensibilità alle problematiche della rilevanza e della pertinenza dei risultati ottenuti attraverso una ricerca affiora in modo evidente, dal momento che questi concetti vengono appena menzionati e invece, tra le metodologie utilizzabili, oltre al modello booleano, vengono citate solo quelle incentrate sull'uso degli spazi vettoriali e del *relevance feedback*:

> Moreover, an important characteristic of the *Search* functionality lies in which model is adopted in identifying the pertinence of the objects with respect to a query, e.g. the Boolean model or the vector-space model[53].

Vector Space Model, proposto da Gerard Salton[54] e basato sullo studio degli spazi vettoriali nell'ambito dell'algebra lineare, è un modello che è stato ampiamente utilizzato nell'ambito dell'Information retrieval, sia per l'indicizzazione automatica dei documenti full-text, sia per misurare la distanza semantica tra documenti e richieste

52 *Ivi*, p. 80.

53 *Ivi*, p. 93.

54 Gerard Salton; Andrew Wong; Chung-Shu Yang, *A vector space model for automatic indexing*. «Communications of the ACM», 18 (1975), n. 11, p. 613-620.

degli utenti al fine di determinare il livello di rilevanza. Viene usato per rappresentare come vettori sia gli oggetti che i documenti allo scopo di descriverne il contenuto, ed è basato sostanzialmente sul computo della frequenza dei termini portatori di significato usati in un testo. Tutte le parole significative dei documenti, estratte automaticamente, o le frasi, vengono trasformate in vettori, e lo spazio vettoriale presenterà tante dimensioni quanti saranno i vettori ottenuti. Il computo della frequenza dei termini, sia ad alta frequenza che a bassa frequenza nei documenti, fornisce la possibilità di attribuire pesi, o punteggi, ai termini di indice, utilizzando il prodotto di tre fattori: la frequenza dei termini, la frequenza nella collezione o frequenza inversa (TF/IDF) e la normalizzazione in base alla lunghezza del documento. Il grado di similarità tra i documenti di una collezione viene misurato in base alla distanza esistente tra i documenti e le parole usate nelle *queries.* La similarità tra documenti è sempre basata sul computo della presenza delle parole. Il coefficiente di similarità viene calcolato generalmente usando la misura del coseno, che misura la deviazione dell'angolo formato tra il documento vettore e la *query* vettore. Questa misura non soddisfa tutte le necessità della ricerca, soprattutto nel caso delle Scienze umane e sociali, nonostante le correzioni che sono state introdotte, soprattutto volte a tenere in maggiore considerazione i legami semantici che le parole della lingua naturale esibiscono. Ad esempio, è stato proposto l'uso di un *Context Vector Model*[55], attraverso il quale si può calcolare la frequenza della co-occorrenza dei termini permettendo di identificare la dipendenza delle parole da contesti differenti nei quali esse sono inserite.

Ugualmente, gli algoritmi utilizzati per realizzare il *relevance feedback* esplicito, considerano la misura dell'equivalenza dei documenti soprattutto analizzando i termini presenti nei titoli in relazione ad un documento definito rilevante dall'utente[56].

La filosofia sottesa al DELOS Model trova un fondamento esclusivamente nelle strategie dell'Information retrieval. Per i metadati semantici si fa riferimento esclusivamente all'uso di parole chiave, per determinare la pertinenza delle risposte alle *query* si fa riferimento a criteri quantitativi basati sul computo semantico in base agli spazi vettoriali e al *relevance feedback*, sistemi basati sostanzialmente sul computo dei termini, senza prendere assolutamente in considerazione la possibilità di analizzare i diversi ruoli giocati dai concetti in base ai contesti e alle diverse strategie di ricerca nelle quali i documenti si inseriscono. Sappiamo dalle ricerche di Tefko Sarácevic[57] che la pertinenza di un documento, invece, non è un elemento determinabile una volta per tutte, ma è dinamica e varia in base al soggetto conoscitore, è relativa alla conoscenza pregressa del singolo ricercatore e diversa a seconda del tipo di ricerca che viene realizzata.

5. Ulteriori sviluppi per le biblioteche digitali.

Se prendiamo in considerazione il ventaglio delle necessità di ricerca, in particolare nell'ambito delle Scienze sociali, le caratteristiche che aggiungono una profondità

55 Holger Billhardt; Daniel Borrajo; Victor Maojo, *A context vector model for Information retrieval*, «Journal of the American Society for Information Science and Technology», 53 (2002), n. 3, p. 236-249.

56 Efthimis N. Efthimiadis, *Query expansion*, «Annual review of information science and technology», 31 (1996), p. 121-187.

57 Tefko Sarácevic, *Relevance: a review of and a framework for the thinking on the notion in information science,* «Journal of the American Society for information science», 26, 1975, n. 6, p. 321-343. Tefko Sarácevic, *Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II*, «Advances in librarianship», 30, 2006, p. 3-71.

semantica alle biblioteche digitali appaiono poco soddisfacenti. Le biblioteche digitali semantiche usano semplicemente FRBR come una ontologia bibliografica per raggruppare le diverse espressioni della medesima opera, per realizzare sistemi di *alerting* basati sui profili dell'utenza, e il raffinamento per faccette nella funzionalità della ricerca. Tuttavia, queste nuove funzionalità risultano piuttosto limitate, e tra l'altro non nuove, in quanto già largamente impiegate negli OPAC di ultima generazione.

Sia nel DELOS Model, sia nella organizzazione della struttura e degli obiettivi delle biblioteche digitali semantiche, risulta invece assente il riconoscimento del bisogno di metodologie per l'indicizzazione semantica, che consentano un profondo e dettagliato esame di ciascun documento, e permettano poi quindi di ritrovare i documenti sulla base di relazioni strutturate tra i contenuti. In effetti, anche nei modelli realizzati proprio nell'ambito bibliografico – FRBR, FRSAD e FRBRoo – questo riconoscimento è sfortunatamente assente.

Ritengo che sia necessario perciò spostare l'attenzione verso le metodologie di indicizzazione semantica, analizzare le strategie che potrebbero arricchire le funzioni della ricerca semantica espandendone le potenzialità, e sviluppare quei miglioramenti che renderebbero possibile usare i risultati di una indicizzazione più approfondita e sfaccettata.

La prima sfida consiste nell'investigare la possibilità di espandere le funzionalità della ricerca permessa dalle biblioteche digitali che vengono dichiarate nel DELOS Model, e cioè la ricerca usando i metadati esistenti, gli operatori booleani e il *relevance feedback*. L'obiettivo è quello di armonizzare queste funzioni con quelle che provengono dall'analisi concettuale dei documenti che negli istituti bibliotecari è stata applicata per secoli.

Al fine di migliorare la ricerca nelle biblioteche digitali, Dagobert Soergel[58] prendeva in considerazione l'opportunità di permettere agli utenti di espandere i termini della ricerca usando tesauri e ontologie, col supporto di sistemi di classificazione esistenti e dell'intero insieme dei KOS, che possono assistere gli utenti nella ricerca dei documenti e all'interno dei documenti.

Si tratta di una strategia certamente utile. Tuttavia, l'analisi concettuale dei documenti realizzata da esseri umani, ma impostata su metodologie di indicizzazione semantica più raffinate di quelle ordinariamente adottate nelle biblioteche tradizionali, deve costituire l'obiettivo principale, allo scopo di consentire la realizzazione di metadati semanticamente più ricchi e quindi elevare il livello delle funzionalità della ricerca nelle biblioteche digitali.

Punto focale è la realizzazione di indici semantici per la ricerca che siano adeguati rispetto alle necessità degli utenti. È necessario quindi focalizzare l'attenzione sullo studio delle strategie di indicizzazione che possano avere come risultato un arricchimento delle funzioni di ricerca semantica, espandendone le potenzialità intrinseche, ed elaborare interventi che rendano possibile utilizzare attraverso le interfacce gli esiti di una indicizzazione più approfondita e sfaccettata. In particolare nel vasto settore delle Scienze umane è necessario creare i metadati semantici più ricchi e adeguati ai diversi livelli di utenza. Allestire un ventaglio di indici, ciascuno tarato sulle necessità della tipologia di utenza cui si rivolge, è una strategia che può condurre a buoni risultati. Proprio nelle discipline del settore umanistico le monografie e i saggi scientifici si offrono a interpretazioni molteplici e possono essere utilizzabili in scenari di ricerca differenti, all'interno dei quali acquistano valenze e significatività di peso e

58 Dagobert Soergel, *Digital libraries and knowledge organization*. In: *Semantic digital libraries*, cit., p. 9-39.

spessore differenti. Fruitori e interpreti diversi colgono della letteratura scientifica aspetti diversi, considerando predominanti di volta in volta l'uno o l'altro contenuto, a seconda della pertinenza con le ricerche del momento.

Le diverse proprietà mostrate da un documento, oggettivamente rilevabili, possono avere significati diversi in ambiti disciplinari diversi e all'interno di finalità scientifiche diverse. Come ha messo in evidenza alcuni anni fa Birger Hjørland, i soggetti catalografici si presentano come funzioni delle proprietà del documento, come 'potenziale epistemologico' dei documenti[59]. L'analisi semantica dei documenti dovrebbe quindi considerare un ventaglio oggettivamente sostenibile di possibili soggetti facendo emergere una pluralità di argomenti, ritagliati sulle diverse necessità informazionali delle diverse utenze cui i documenti si rivolgono. Diverse biblioteche e centri d'informazione potrebbero stabilire i soggetti catalografici ciascuna tenendo conto, ad esempio, delle necessità del gruppo di utenti predominante. L'indicizzazione semantica dello stesso libro o del medesimo articolo, da parte di diverse biblioteche, tenendo conto delle differenti prospettive di ricerca della propria utenza, invece di costituire un elemento negativo ai fini della ricerca, costituisce una ricchezza epistemologica e se i diversi indici semantici fossero consultabili attraverso un unico strumento di ricerca lo stesso libro sarebbe rintracciabile attraverso prospettive di analisi diverse.

I sistemi di indicizzazione semantica sviluppati nell'ambito dell'Information retrieval, e cioè i modelli probabilistici basati sulla frequenza dei temini (TF/IDF), il *Vector space model* e la determinazione del grado di similarità tra query e documenti, non perseguono come obiettivo la fruizione dei documenti secondo molteplici modelli interpretativi e in base ad una pluralità di prospettive di ricerca e infatti risultano meno utili nel caso delle Scienze umane.

Nel processo di sviluppo che si è delineato a partire dal 2000 nel mondo delle biblioteche digitali, l'elemento più significativo è stato certamente l'impiego delle infrastrutture *peer-to-peer* e Grid, che hanno consentito la realizzazione di strutture flessibili e indipendenti ed una gestione dei singoli *peer* libera dal controllo centralizzato.

L'uso delle annotazioni realizzate dagli utenti costituisce un aiuto alla ricerca, tuttavia, in una biblioteca digitale, almeno per la porzione di contenuti costituita da monografie digitalizzate e da articoli scientifici, in particolare nell'ambito delle Scienze umane, l'indicizzazione approfondita, realizzata da indicizzatori professionisti, deve continuare a fornire le coordinate per il ritrovamento dei documenti rilevanti.

Per quanto riguarda la metodologia dell'indicizzazione semantica[60], gli elementi su cui è necessario lavorare sono i seguenti:

1. L'analisi concettuale dei libri e degli articoli deve rendere possibile mettere in evidenza qual'è l'argomento principale trattato nella monografia (o gli argomenti principali), e quali argomenti secondari, rispetto a quello principale, sono considerati nel documento. Gli argomenti secondari hanno un trattamento autonomo, ma sono oscurati dal soggetto (o soggetti) di discussione, sono trattati dal punto di vista del-

59 Birger Hjørland, *Information seeking and subject representation. An activity-theoretical approach to information science*. Westport (Conn.)-London, Greenwood Press, 1997.

60 L'impostazione metodologica trova un fondamento nei lavori di Benedetto Aschero, *Teoria e tecnica dell' indicizzazione per soggetto.* Nuova edizione riveduta e ampliata. Milano, Editrice Bibliografica, 1993. Benedetto Aschero, *Manuale pratico di soggettazione. Esercizi graduati per l' apprendimento*. Milano, Editrice Bibliografica, 1982.

l'argomento principale cui la monografia è dedicata. Questa informazione deve essere fornita agli utenti.
2. Le funzioni della ricerca nelle biblioteche digitali devono permettere di ritrovare i documenti tenendo conto dei diversi livelli di approfondimento presentati dalle trattazioni. Attraverso nuove funzionalità, l'utente deve avere la possibilità di conoscere quale è il livello di profondità dell'argomento trattato in un documento, ad esempio, sapere con certezza se un argomento è trattato in modo approfondito e completo in una certa monografia, se il livello è scientifico oppure divulgativo, se è adatto ai ragazzi.
3. Il trattamento degli argomenti 'menzionati' in una monografia è una delle difficoltà più significative. La decisione riguarda la creazione di indici semantici per le entità o i topici che necessariamente appartengono all'argomento principale e che nella monografia vengono trattati esclusivamente in relazione all'argomento principale. In questo caso, non sarebbe appropriato creare indici semantici perché ciò potrebbe essere fuorviante per gli utenti.

Si può utilizzare RDF per modellare metadati semantici a diversi livelli, secondo il modello delineato, aggiungendo ulteriori relazioni tra una risorsa e i suoi contenuti, per gli argomenti principali, per quelli secondari e per le entità menzionate. L'obiettivo è quello di creare con RDF metadati semantici più articolati e la possibilità di definire chiaramente, con specifici URI, topici principali, topici secondari e relazioni tra di essi. Si può utilizzare RDFa Core 1.1.[61], che permette di specificare il tipo di relazioni tra Subject e Object attraverso i predicati identificandole con IRI.

Articolo proposto il 16 marzo 2014 e accettato il 28 aprile 2014.

ABSTRACT

AIB studi, vol. 54 n. 1 (gennaio/aprile 2014), p. 11-34. DOI 10.2426/aibstudi-9955.

MARIA TERESA BIAGETTI, Sapienza Università di Roma, Dipartimento di Scienze documentarie, linguistico-filologiche e geografiche, viale Regina Elena 295, 00161 Roma, e-mail mariateresa.biagetti@uniroma1.it

Sviluppi e trasformazioni delle biblioteche digitali: dai *repositories* di testi alle *semantic digital libraries*
L'articolo delinea l'evoluzione delle biblioteche digitali, dai repositories di testi alle biblioteche di seconda generazione basate sull'uso dell'architettura *service-oriented* e delle reti P2P e Grid, la cui caratteristica è l'integrazione in un unico spazio informativo di documenti provenienti da istituzioni diverse e la possibilità di ricerca federata tra diverse biblioteche digitali. Si prendono in considerazione le peculiarità delle biblioteche digitali semantiche di ultima generazione, che utilizzano gli elementi fondativi del Semantic web – RDF e le ontologie – e l'apporto dei collaborative *filtering systems*. Infine, si discutono alcune criticità presenti nel DELOS Model, il modello di riferimento per le biblioteche digitali, e si pone l'accento sulla necessità di realizzare metadati semanticamente più ricchi che possano meglio rispondere alle necessità di ricerca degli utenti, in particolare nell'ambito delle Scienze umane.

61 W3C, RDFa Core 1.1 (Recommendation 07 June 2012), <http://www.w3.org/TR/2012/REC-rdfa-core-20120607/> . In un successivo articolo, oltre a presentare le modalità di ricerca semantica nelle biblioteche digitali di seconda e ultima generazione, verranno approfondite le possibilità di realizzare metadati semantici arricchiti, qui succintamente delineate.

The transformation of digital libraries from text repositories to semantic digital libraries
The article describes the evolution of digital libraries from text repositories to second-generation libraries. This kind of library uses service-oriented architecture, P2P networks, and Grid, and is characterized by the integration of documents coming from different institutions, as well as by the possibility of federated search within multiple digital libraries. The author then takes into consideration the peculiarities of last generation digital semantic libraries, i.e. the application of the fundamentals of the Semantic Web – RDF and ontologies – and the use of collaborative filtering systems. In the final part the article outlines some critical issues in the DELOS digital library reference model, and the need for semantically rich metadata to support the research needs of users, especially the ones in the field of the Human Sciences.