L'*authority control* nel mondo dei metadati

José Borbinha
Biblioteca nazionale portoghese

Biblioteche digitali

La parola "metadati" è recentemente diventata di moda, in relazione all'esplosione di Internet e all'emergere di nuovi contenuti e servizi in qualche modo associati alle biblioteche, agli archivi, ai musei e alle organizzazioni ad essi collegate. Un nome dato a questo nuovo contesto è stato "biblioteche digitali"!
La Figura 1 illustra un punto di vista evolutivo del problema, dalla prospettiva della biblioteca tradizionale. Qui notiamo Internet come il fattore recente più rilevante nell'evoluzione della "biblioteca", dopo molti altri. Tra questi diamo particolare rilevanza all'introduzione del computer nelle biblioteche, che ha avuto un impatto nel catalogo digitale e nella definizione dei primi standard di descrizione bibliografica. Questo è stato seguito dai primi servizi di comunicazione dei dati (X.25, TELNET, BBS - Bulletin Board Systems, ecc.), che fornivano accesso remoto al catalogo e ad altri servizi bibliotecari normali. Verso la fine degli anni Ottanta emergono i personal computer e i CD-ROM, che portarono alla biblioteca digitalizzata, in grado di fornire accesso anche ai contenuti. Infine, abbiamo Internet e il World Wide Web, coi quali oggi lavoriamo.
Questa evoluzione ci ha portato al problema della definizione della "biblioteca virtuale" o, in termini più comuni, della "biblioteca digitale". Questo è diventato di recente un argomento caldo di discussione, contraddistinto da demagogia ma anche da molto lavoro impegnativo, sia concettualmente che tecnicamente. Ha anche attratto professionisti e comunità estranee al mondo tradizionale delle biblioteche, particolarmente dal mondo dell'ingegneria e dell'informatica.
Da un punto di vista tecnico generico, queste comunità hanno inteso la "biblioteca digitale" come un caso della classe specifica "sistemi d'informazione" come proposto nel sistema di classificazione dell'ACM (Association for Computer Machinery), schematizzato nella Figura 2 [1]. Un punto di vista simile è emerso da un incontro molto libero di DELOS, come appare ancora nella Figura 2 [2], che affronta il problema secondo una prospettiva più ampia. Per quanti sono interessati allo sviluppo di un punto di vista completo di queste attività, discussioni e visioni, due importanti risorse sono il D-Lib Forum [3] e il DELOS Network [4]. Ulteriori informazioni e discussioni su questo, viste dalla prospettiva di una biblioteca di deposito, sono presentate in [5] e [6].

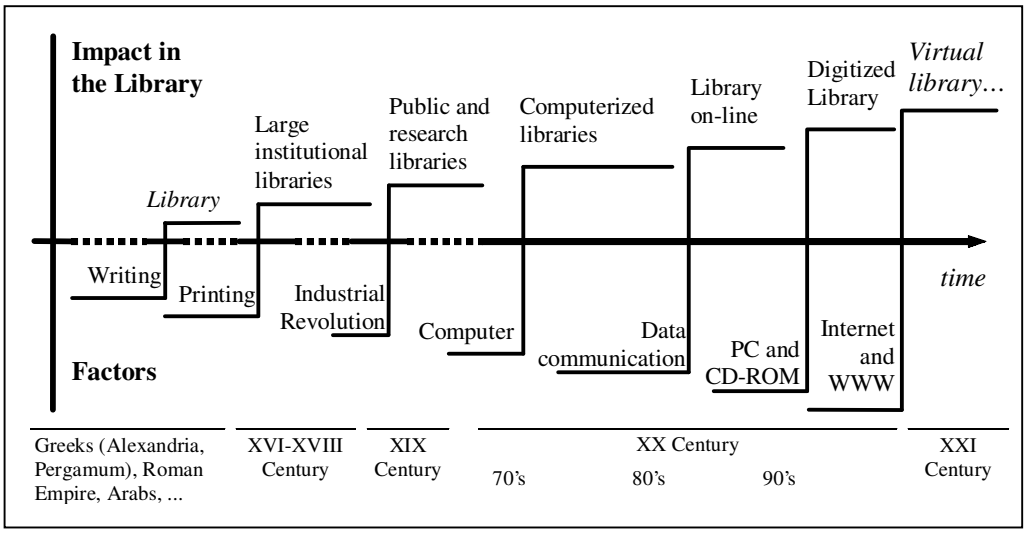

Figura 1: Biblioteche e tecnologia nel tempo.

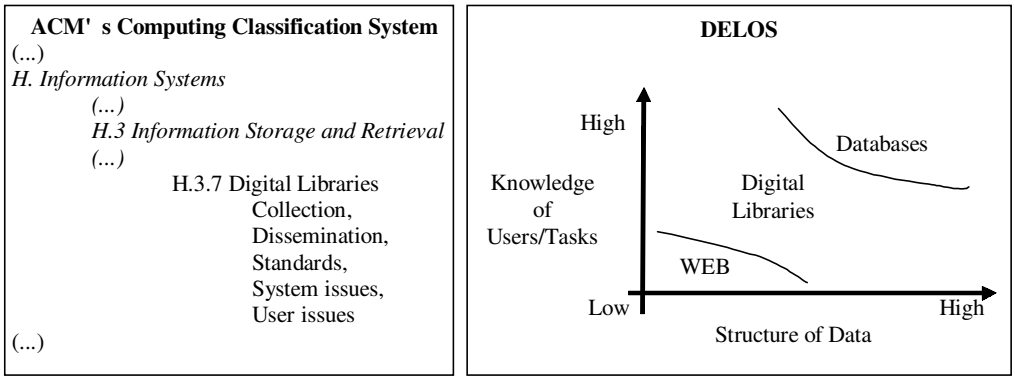

Figura 2: La "biblioteca digitale" secondo ACM e DELOS.

Nonostante tutte le riflessioni e gli sviluppi verificatisi negli anni recenti, dobbiamo accettare l'idea che non c'è una definizione unica e universale di "biblioteca digitale". La percezione dipende troppo dal punto di vista (questo potrebbe suonare non molto piacevole per le biblioteche tradizionali, ma dovremmo ricordare che questa confusione di concetti non è così strana, ad esempio, per archivi o musei). Questo è un fatto che diamo per scontato – e non è scopo di questa relazione discuterlo – ma è molto importante riconoscerlo, specialmente se andremo verso la definizione di modelli, procedure e standard comuni che ciascuno associa all'espressione "authority control". Tuttavia, e per lo scopo di questa relazione, lasciate che proponga una semplice definizione di "biblioteca digitale" come "gruppo di servizi libero o controllato, mantenuto da un'entità identificata, che rende possibile l'identificazione e l'accesso a risorse informative digitali documentarie, multimediali o di qualsiasi altro tipo".

Metadati e biblioteche digitali

Metadati è un termine coniato tempo fa da informatici e ingegneri nel mondo dei database per fare riferimento a informazioni strutturate che descrivono schemi di

database (p.e., il modo in cui i dati sono organizzati in un database). Un esempio di questa prospettiva è [7]. Il termine non è stato usato in questo modo nelle biblioteche digitali, ed è molto importante essere consapevoli di questo particolare. Nelle biblioteche digitali, i metadati sono stati di solito definiti, semplicemente come "dati sui dati" (che una volta registrati in un database significherebbe, dal punto di vista precedente, i dati in un database, mentre i metadati sarebbero le informazioni necessarie per descrivere l'organizzazione di quel database). In questo modo la "comunità Internet" ha assunto il termine dopo l'emergere del World Wide Web, e ora l'uso più comune del termine è in quest'area. Inoltre, dovremmo preferire la definizione di "informazioni strutturate su altre informazioni o risorse".
Tuttavia, anche in questo ambito ci sono alcune incomprensioni comuni sul termine. Ad esempio, dobbiamo essere molto attenti e sottolineare il fatto che i metadati si suppone facciano riferimento a informazioni codificate secondo uno schema specifico e non alla tecnologia che le gestisce, e nemmeno agli spazi concettuali che controllano il valore degli elementi dell'informazione. In questo senso MARCXML [8] o DCMES [9] non sono metadati, ma schemi di metadati, ad esempio, o definizioni di come esprimere metadati in quanto informazioni strutturate su altre informazioni o risorse. Nello stesso senso XML [10] in se stesso è solo uno strumento tecnologico e non metadati o anche uno schema di metadati. XML è un linguaggio dove possiamo definire schemi (usando un DTD – Document Type Definition, o più recentemente usando l'XML Schema language [11]). Inoltre, spazi autorevoli, come linguaggi di indicizzazione, schemi di classificazione, ecc., non sono metadati in se stessi, ma valori o regole per trovare il giusto valore da dare a elementi di metadati!
Nelle recenti attività delle biblioteche digitali e nella letteratura su di esse possiamo trovare diversi esempi di classi diverse di metadati, cioè:

- descrizione bibliografica delle risorse: descrizione bibliografica e identificazione delle risorse, come titoli, autori, termini di indicizzazione, classificazione, abstract, surrogati, ecc.;
- amministrazione delle risorse: informazioni amministrative sulla risorsa, come le informazioni sul processo di acquisizione e sui costi, i diritti, ecc.;
- conservazione delle risorse: requisiti tecnici o gestionali per la conservazione a lungo termine;
- descrizione tecnica e strutturale delle risorse: requisiti tecnici per manipolarle (sistemi e strumenti), ecc.;
- accesso, uso e riproduzione delle risorse: informazioni su come accedervi (indirizzi, password, ecc.), termini e condizioni per l'accesso e la riproduzione, ecc.;
- amministrazione dei metadati: informazioni su altre classi di metadati, come data di creazione, origine, autenticità, ecc.

La descrizione bibliografica delle risorse è una questione comune nelle biblioteche e negli archivi tradizionali dove, rispettivamente, la famiglia di schemi MARC [12] [8] e lo schema EAD [13] sono largamente usati. Anche il mondo esterno a questi ambiti tradizionali si sta movendo, creando modelli di descrizione che, una volta realizzati, potrebbero essere riusati a basso costo. Un esempio interessante di ciò è il formato di metadati descrittivi ONIX, definito da un consorzio di editori [14].
Più recentemente, sono stati identificati più requisiti per i metadati rispetto a quelli per la descrizione bibliografica. Sono rilevanti, ad esempio, gli sforzi per la descrizione tecnica delle risorse [15], nuovi approcci per la classificazione e le relazioni tra risorse [16] [17], per la conservazione [18] [19] [20], per la gestione dei diritti [9], ecc.

Altre iniziative rilevanti sono state intraprese per lo sviluppo di strutture generali allo scopo di coprire diverse classi di metadati. Un esempio interessante è la definizione, da parte della Library of Congress dello schema METS, con lo scopo di coprire metadati bibliografici, strutturali e amministrativi [21]. Un'altra iniziativa interessante è quella di MPEG (Moving Picture Expert Group) [22]. Particolarmente rilevante è MPEG-7 [23] e più in generale MPEG-21 [24], che riserva particolare attenzione ai fini della "Digital Item Declaration" (un pacchetto di metadati generico), "Digital Item Identification and Description" (identificatori, descrizione bibliografica e tecnica) e "Intellectual Property Management and Protection" (amministrazione, accesso e uso delle risorse).
A questo alto livello, simili a MPEG-21 sono anche i modelli di riferimento CIDOC [25], uno schema [o struttura] di mediazione che ha lo scopo di promuovere l'interoperabilità nei musei, usando metadati descrittivi eterogenei, MoReq [26], un modello per la gestione dell'archiviazione di registrazioni elettroniche e il ben noto FRBR promosso dall'IFLA [27]. Tuttavia questi non sono schemi di metadati specifici, ma direttive molto importanti per la loro definizione, così come le AACR sono state importanti per lo sviluppo di standard bibliografici, e di sistemi e servizi nelle biblioteche [28].

Metadati e tecnologia

Un'altra importante questione da considerare quando si parla di metadati è la relazione tra il piano concettuale e quello tecnologico. In senso generale un modello concettuale o uno schema di metadati dovrebbe essere indipendente da qualsiasi implementazione tecnologica. Ciò non è sempre vero, tuttavia, dal momento che a volte vediamo esempi dove, specialmente per il gusto dell'esempio e per una migliore comprensione (e per aiutare la sua immediata applicazione), i modelli sono accompagnati da specifiche soluzioni tecnologiche. Ciò è successo con MARC e l'ISO2709 [29], cosa che non ostacola l'effettiva definizione di MARCXML.
Andando avanti nella discussione, proporremo un modello di riferimento costituito da quattro distinte prospettive: concettuale, del contesto, del servizio e della tecnologia. La Figura 3 illustra questo modello.
La "prospettiva concettuale" è quella in cui vengono considerati modelli di riferimento generici. Qui, non abbiamo ancora registrazioni, database o file di dati, ma solo concetti e modelli su come le cose possono o dovrebbero essere fatte. Possiamo suddividere questa prospettiva in tre ambiti: modelli di riferimento generici, che si suppone definiscano un modello oggettivo *top-down*; schemi di metadati, che dovrebbero far riferimento a questioni o aree di applicazione specifiche (ma che dovrebbero comunque essere indipendenti dalla tecnologia) e implementazioni di metadati, dove infine sono affrontate le questioni tecnologiche (specialmente per la codifica).

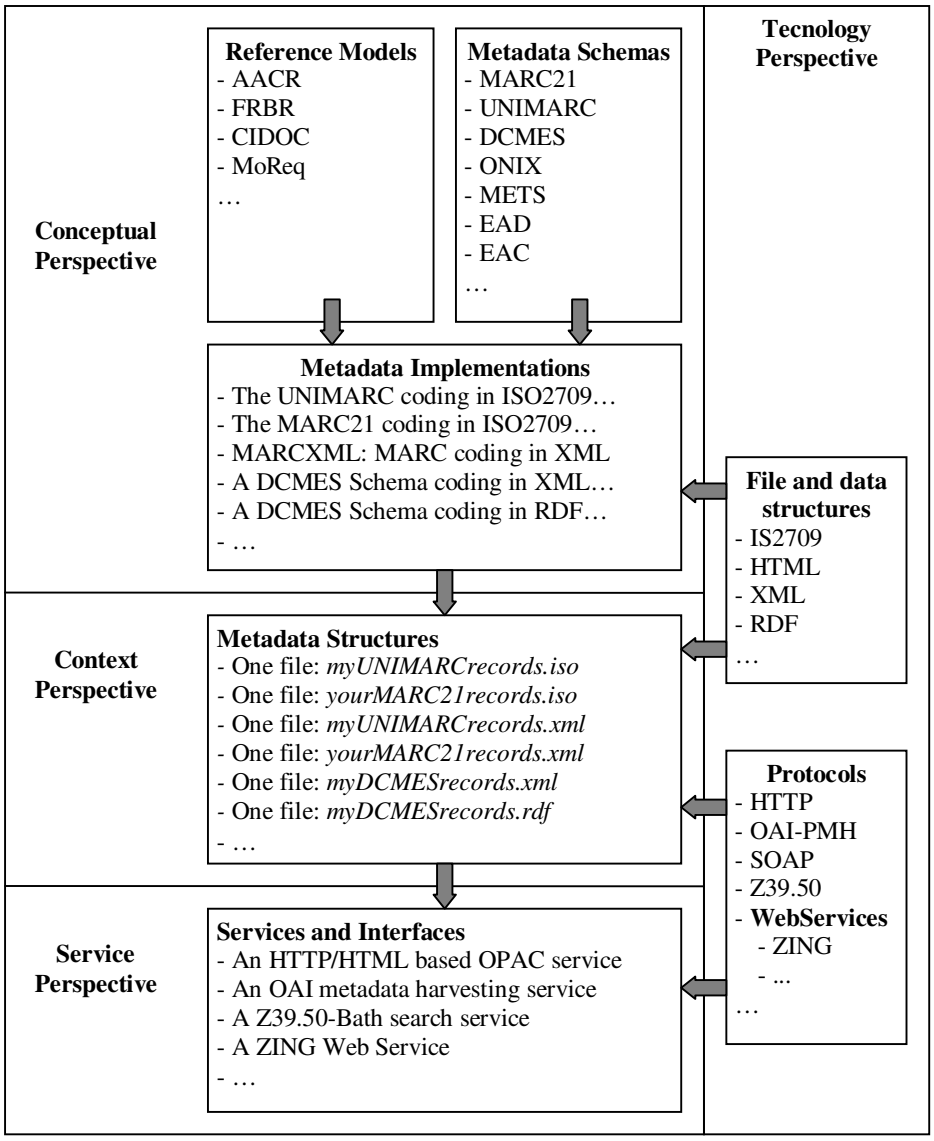

Figura 3: Prospettive multiple per il problema "metadati".

La "prospettiva del contesto" rappresenta la concretizzazione della "prospettiva concettuale". I modi in cui queste concretizzazioni vengono fatte dipende dalle opzioni o dai vincoli tecnologici, e anche dalla natura e dalle caratteristiche dei servizi locali. Ad esempio, in un contesto specifico possiamo decidere di trasportare ed esplorare un insieme di record in formato UNIMARC codificati in ISO2709, mentre in un altro contesto potremmo decidere di ottenere gli stessi risultati usando invece MARCXML (un semplice esempio di come ciò possa essere fatto può essere verificato in [30]). Il valore oggettivo delle informazioni trattate e il loro significato in entrambe le soluzioni è lo stesso, sono solo diverse le implementazioni tecniche. Ciò può essere dovuto alla tecnologia da usare nel servizio finale, o derivata dal passato o da nuovi sistemi con i quali ci si aspetta che le nuove soluzioni interagiscano.
Questo ci porta all'ultimo livello, la "prospettiva del servizio". Qui abbiamo a che fare principalmente con interfacce per esseri umani o per altri sistemi (protocolli). Di solito non è irrilevante per un protocollo qual'è il formato di codifica dei metadati da

trasportare, ma la tendenza è stata quella di renderli il più possibile flessibili. Un esempio di ciò è il protocollo OAI-PMH [31], che specifica *Dublin core* come formato di *default*, ma che si è evoluto per supportare qualsiasi altro formato esprimibile in uno schema XML. Ci aspettiamo che una generalizzazione completa di questo si otterrà con il concetto di Web Services [32], di cui ZING, la generazione successiva di Z39.50 è un esempio potenzialmente molto interessante [33].

I metadati nella società dell'informazione

È ora di porsi una domanda fondamentale: se i metadati sono una risposta, qual è, dopo tutto, la domanda? Quali sono i requisiti fondamentali della biblioteca digitale ai quali un concetto quale "metadati" si suppone forniscano una so luzione? Questi requisiti appartengono a tre grandi categorie:

- "eterogeneità di generi": i nuovi supporti informativi non sono semplici e stabili, come lo erano i libri a stampa, i periodici o i quotidiani. Una notevole eterogeneità e dinamismo di nuovi oggetti e modelli di supporto hanno caratterizzato la realtà dell'"editoria digitale". Per gestirli in modo tecnico ed economico, la biblioteca digitale deve prevedere e capire chiaramente ciascuna categoria di oggetti e modelli. Media, formati di dati, versioni, tipo, ecc., sono esempi delle caratteristiche che possono definire nuovi generi di risorse. I generi sono importanti per la definizione dei criteri di selezione per licenze, acquisizione e deposito delle risorse, indipendentemente dal loro soggetto e dal contenuto artistico o intellettuale. Per aiutare la biblioteca ad affrontare questi problemi abbiamo i concetti di metadati strutturali e tecnici, ad esempio;
- "interoperabilità": la biblioteca digitale fa parte del World Wide Web. In questo scenario, gli utenti si aspettano non solo di raggiungere la biblioteca da qualsiasi luogo, ma anche di raggiungere qualsiasi documento. Questo significa che gli utenti potrebbero non capire bene (e non accettare del tutto), se gli si dice che non possono usare un unico servizio per cercare allo stesso tempo in una biblioteca e in un archivio di film e avere accesso a libri e film creati da o su, ad esempio, Federico Fellini. Per essere in grado di offrire servizi di questo genere la biblioteca digitale, ora vista non solo come un'evoluzione della biblioteca tradizionale, ma come un servizio di livello concettualmente più alto, secondo la definizione già data, ha bisogno di essere progettata come servizio distribuito, o come aggregazione di servizi eterogenei (Figura 4). Ciò richiede la cooperazione da parte di biblioteche, archivi, musei e altre categorie di organizzazioni e attori specializzati e generali. Ancora una volta, la capacità di automatizzare questa interoperabilità è cruciale per i suoi costi e la sua efficienza tecnica, e comporta la definizione di requisiti per nuove categorie di interfacce e di metadati, definiti o semplicemente adottati da quegli attori. Cosa che tradizionalmente, è stata realizzata con mezzi, quali ad esempio Z39.50 [34], integrato recentemente da nuovi modelli e soluzioni che implicano registrazioni bibliografiche in XML [8] [30], si avvantaggiano grazie a strutture semplici come *Dublin core* [35], o forniscono grandi quantità di registrazioni per la raccolta mediante OAI-PMH [36] [37]. Questa tecnologia è stata specialmente concepita dalle comunità di biblioteche digitali, ma per il futuro dobbiamo iniziare a pensare in ambiti che riutilizzino soluzioni generali;
- "tecnologia": con lo sviluppo progressivo del Web semantico, e con la su a tecnologia che diventa sempre più generale e onnipresente, una parte importante dei

componenti e dei prodotti applicati nelle biblioteche digitali non saranno più specifici delle biblioteche stesse (le biblioteche tradizionali non sono molto abituate a questa generalità). Quei componenti saranno generici, specialmente per quanto riguarda le interfacce utente, la tecnologia dei database, i protocolli e i servizi Web. Questo significa che i metadati non sono un concetto specifico delle biblioteche digitali ma un concetto generale in qualsiasi sistema informativo (che in effetti è una "biblioteca digitale"). Di conseguenza, le comunità delle biblioteche digitali devono essere determinate nell'imporre i loro requisiti nella definizione di quei componenti (lavorando, ad esempio, con il World Wide Web Consortium, l'ISO, ecc.), ma anche disponibili a riutilizzare soluzioni che potrebbero essere state definite e diventate standard in qualunque altro luogo. Una regola aurea nel mondo della tecnologia dell'informazione è che può essere molto costoso fornire per la prima volta un nuovo sviluppo per un problema specifico ma, successivamente, il costo della generalizzazione della soluzione può diventare molto basso. Biblioteche, musei e archivi, che combattono sempre con ristrettezze di bilancio, devono prendere questa cosa in seria considerazione.

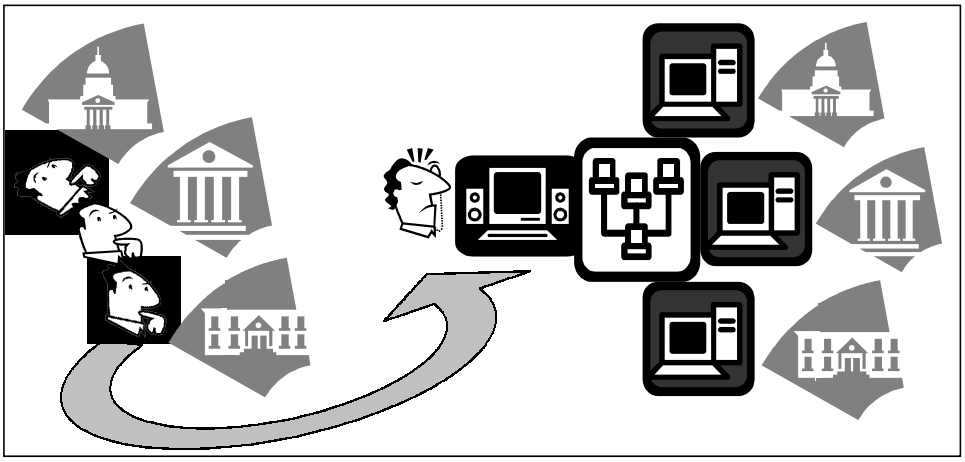

Figura 4: L'interoperabilità in un mondo in rete.

Le sfide

In parole povere, possiamo concludere che una visione essenziale della "biblioteca digitale" è stata quella di una naturale evoluzione di entità molto ben definite, con interfacce stabilite, per un concetto nuovo e poco definito, richiesto da un ambiente più dinamico. Questo avrà implicazioni importanti in alcune fondamentali riflessioni in biblioteche, musei e archivi, e l'*authority control* è solo uno degli esempi.
Le biblioteche tradizionali sono solite riconoscere diversi attori rilevanti per la descrizione bibliografica [14]. Il nuovo contesto rende necessario non solo mantenere questi concetti, ma anche estendere l'analisi per riconsiderare, ora, nuove questioni chiave relative, ad esempio, all'autenticazione, al possesso, al *copyright*, al controllo dell'accesso e all' *authority control* in generale.
Diversi progetti internazionali hanno analizzato questi problemi, tra cui il Progetto INTERPARTy [15], la DC-Agent activity [38], e più in generale il DELOS/NSF working group on Actors in Digital Libraries [39]. Nel contesto specifico degli archivi, l'EAC - Encoded Archiving Context [40] [41] è un lavoro interessante nel campo

dell'*authority description*, che dovrebbe essere seguito con attenzione da tutti. Questo è stato fatto, ad esempio, nel Progetto LEAF [42], il successore di MALVINE [43]. Questi progetti, di cui la Biblioteca nazionale portoghese è un partner attivo, sono interessanti dimostrazioni di come schemi e fonti eterogenee di metadati possono essere combinati in servizi comuni, con benefici rilevanti. Nel progetto LEAF, ci aspettiamo di dimostrare che avere a che fare con descrizioni eterogenee di *authorities* non deve essere sempre un problema. In effetti, potremmo anche trarre vantaggio da questa situazione, per migliorare altre descrizioni e per migliorare il richiamo nel lavoro di [identificazione/presentazione] delle risorse in MALVINE e in un altro progetto TEL [44].
Infine, penso che una importante lezione da trarre da questa discussione e da trasmettere a qualsiasi altro discorso più incentrato su questioni relative all'*authority control* sia questa: occupatevi di eterogeneità! La biblioteca digitale non può ignorare i nuovi centri di gravità, compresi i fornitori di risorse non specialistiche come la libreria online Amazon, i *gateway* come Yahoo e i servizi generici di recupero delle informazioni come Google. Non penso che le biblioteche dovrebbero ignorare questi e altri nuovi attori simili, che entrano in ogni momento nella società dell'informazione, dove alcuni potrebbero rappresentare nuovi partner potenziali e molto interessanti, capaci di apportare nuove e valide risorse o servizi. In scenari come questo la parola chiave per le biblioteche deve essere "adattamento", che significa capacità di interfacciarsi e di operare insieme per ottenere il meglio da ogni relazione senza imporre regole rigide troppo costose per gli altri partner (allontanandoli). La tecnologia è sufficientemente avanzata per gestire questo! Questo presupposto, se applicato all'*authority control*, significa che il problema potrebbe non essere più come concepire e mettere in pratica processi che portano a regole, descrizioni e formati unici, ma piuttosto essere capaci di capire le regole, le descrizioni e i formati usati da altri e di trarne il meglio per i nostri fini (ed essere anche capaci di dare il massimo ai nostri partner).

Note bibliografiche

[1] ACM. ACM' s Computing Classification System<http://www.acm.org/class/>.

[2] DELOS. Digital Libraries: Future Directions for a European Research Programme. Brainstorming Report. San Cassiano, Alta Badia - Itally. June 13-15, 2001. <htpp://www.iei.pi.it/DELOS/delo2/International/brainstorming.htm>.

[3] D-Lib Forum. <http://www.dlib.org>.

[4] DELOS. Network of Excellence on Digital Libraries. <http://www.ercim.org/delos/>.

[5] José Borbinha - Fernanda Campos - Fernando Cardoso. *Deposit collections of digital publications: a pragmatic strategy for an analysis*. Chapter 4 of *World libraries on the information superhighway: preparing for the challenges of the next millennium.* USA: Idea Group Press, December 1999.

[6] José Borbinha. *The digital library - taking in account also the traditional library*. Elpub2002 Proceedings. Berlin: VWF, 2002, p. 70-80.

[7] OMG. Catalog of OMG Specifications. <http://www.omg.org/technology/documents/spec_catalog.htm>.

[8] LOC. MARC Standards. <http://www.loc.gov/marc/>.

[9] DCMI. Dublin Core Metadata Initiative. <http://www.dublincore.org>.

[10] W3C. Extensible Markup Language (XML). <http://www.w3c.org/XML/>.

[11] W3C. XML Schema. <http://www.w3.org/XML/Schema>.

[12] IFLA. IFLA Universal Bibliographic Control and International MARC Core Activity (UBCIM). <http://www.ifla.org/VI/3/ubcim.htm>.

[13] LOC. Encoded Archival Description (EAD). <http://www.loc.gov/ead/>.

[14] EDItEUR. <http://www.editeur.org>.

[15] W3C. Synchronized Multimedia Integration Language. <http://www.w3.org/TR/REC-smil/>.

[16] W3C. Resource Description Framework (RDF). <http://www.w3.org/RDF/>.

[17] Topic Maps. Topic Maps Consortium. <http://www.topicmaps.org/>.

[18] CEDARS. Curl exemplars in digital archives. <http://www.leeds.ac.uk/cedars/>.

[19] NEDLIB. <http://www.konbib.nl/nedlib>.

[20] PANDORA. Preserving and Accessing Networked Documentary Resources of Australia.

[21] LOC. METS - Metadata Encoding & Transmission Standard. <http://www.loc.gov/standards/mets/>.

[22] MPEG. Moving Picture Expert Group. <http://www.cselt.it/mpeg>.

[23] Day, Neil; Martínez, José M. Introduction to MPEG-7. ISO/IEC working group JTC1/SC29/WG11/N4325. Version 3.0, July 2001.

[24] Bormans, Jan; Hill, Keith. MPEG-21 Overview. ISO/IEC working group JTC1/SC29/WG11/N4318. Version 0.2, July 2001.

[25] CIDOC. CIDOC Conceptual Reference Model. <http://cidoc.ics.forth.gr/>.

[26] IDA. Model Requirements for the Management of Electronic Records (MoReq) <http://www.cornwell.co.uk/moreq>.

[27] IFLA. Functional Requirements for Bibliographic Records. <http//www.ifla.org/VII/s13/frbr/frbr.htm>.

[28] AACR. Joint Steering Committee for Revision of Anglo-American Cataloguing Rules. <http://www.nlc-bnc.ca/jsc/>.

[29] ISO. ISO 2709: Documentation format for bibliographic information interchange for magnetic tape. ISO 1981.

[30] PORBASE. Protótipo de acesso por URN à PORBASE. <http://urn.porbase.org>.

[31] OAI. Open Archives Initiative. <http://www.openarchives.org/>

[32] W3C. Web Services Activities. <http://www.w3c.org/2002/ws/>

[33] LOC. ZING, Z39.50-International: Next Generation. <http://www.loc.gov/z3950/agency/zing/>.

[34] LOC. Z39.50 Maintenance Agency. <http://www.loc.gov/z3950/agency/>.

[35] OAF. Open Archives Forum. <http://www.oaforum.org/>.

[36] OAI. Open Archives Initiative. <http://www.openarchives.org/>.

[37] LOC. MARC Code Lists for Relators, Sources, Description and Conventions. <http://www.loc.gov/marc/relators/>.

[38] DCMI. Dublin Core Metadata Initiative. <http://www.dublincore.org>.

[39] DELOS. Reference Models for Digital Libraries: Actors and Roles. <http://www.delos-nsf.actorswg.cdlib.org/>.

[40] Encoded Archival Context (EAC). <http://www.library.yale.edu/eac/>.

[41] Cover Pages. Encoded Archival Context Initiative (EAC). <http://xml.coverpages.org/eac.html>.

[42] LEAF. Linking and Exploring Authority Files. <http://www.leaf-eu.org/>.

[43] MALVINE. Manuscripts and Letters via Integrated Networks in Europe. <http://www.cordis.lu/libraries/en/projects/malvine.html>.

[44] TEL. The European Library. <http://www.europeanlibrary.org/>.