Authority Control in the World of Metadata

José Borbinha
National Library of Portugal

Digital Libraries

"Metadata" became a recent buzzword related with the explosion of the Internet and the emerging of new contents and services by some way associated to libraries, archives, museums, and related organisations. A name given to this new paradigm has been "Digital Libraries"!

Figure 1 illustrates an evolutionary view of the problem, taken from the perspective of the traditional library. Here we point the Internet as the most recent relevant factor in the evolution of the "library", in the following of a series of others. From those we stress the generic introduction of the computer in the library, which had an impact in the digital catalogue and in the definition of the first standards for bibliographic description. That was followed by the first data communication services (X.25, TELNET, BBS - Bulletin Board Systems, etc.), providing remote access to the catalogue and to other common library's services. In the late 80's we had the emerging of the personal computer and the CD-ROM, which brought the digitized library providing now access to also the contents. Finally, and finally we had the Internet and the World Wide Web, with which we are working today.

This evolution brought us to the problem of the definition of the "virtual library", or in a more common term, of the "digital library". This became a recent hot topic of discussion, with some demagogy but also with lots of real serious work, both conceptual and technical. It attracted also professionals and communities from outside the traditional library's world, especially from Computer Science and Engineering.

From a generic technical perspective, those communities have understood the "digital library" as a case of a specific class of "information systems", as proposed early in the classification system of the ACM - Association for Computer Machinery, resumed in Figure 2 [2]. A similar view resulted from a brainstorming meeting reported by DELOS, as illustrated also in Figure 2 [12], which addresses the problem from a wider perspective. For those interested in developing a complete view of those activities, discussions and visions, two important resources are the D-Lib Forum [16] and the DELOS Network [13]. More information and discussion about this, taken from the perspective of a deposit library, is also present in [4] and [3].

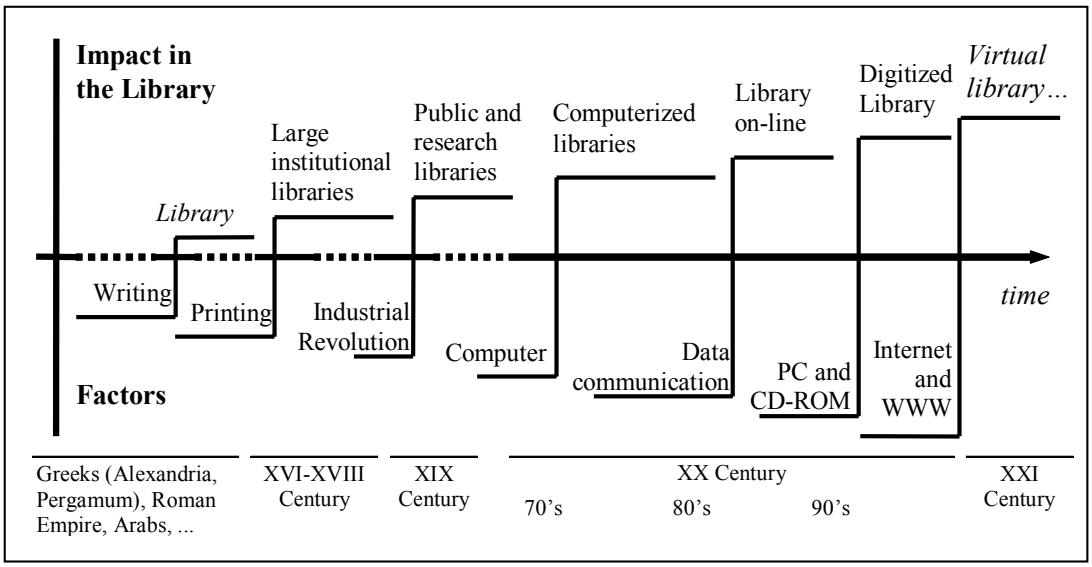

Figure 1: Libraries and technology among the times.

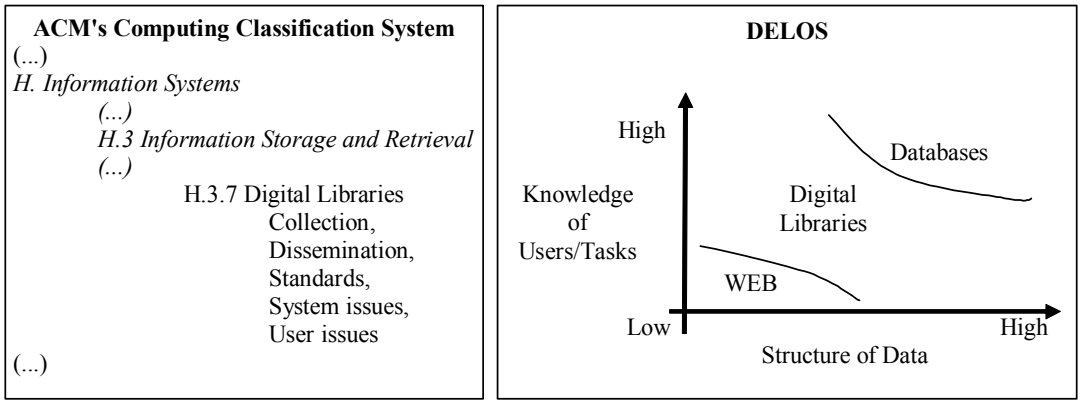

Figure 2: The "digital library" according to ACM and DELOS.

In spite of all the thinking and developments verified in the recent years, we have to accept that there is not a unique and global definition for the "Digital Library". The perception of it depends too much from the perspective taken (this might sound not very good to the traditional libraries, but we should remember that this fuzziness of concepts is not so strange to, for example, archives or museums). That is a fact that we assume, and is not the purpose of this paper to discuss it! But it is very important to recognise it, especially if our next steps are going to be the definition of common models, procedures and standards that everyone associates to the term "Authority Control".

Nevertheless, and for the purpose of our discussion here, let us propose a simple definition for Digital Library as "a free or controlled group of services, maintained by an identified entity, making it possible the discovery and access to documental, multimedia or any other possible classes of digital information resource artefacts".

Metadata and Digital Libraries

Metadata was a term coined a long time by the computer scientists and engineers from the database world to refer to structured information that describes database schemas (e.g., the way in a database the data is organized). An example of this perspective is [35]. This is not how the term has been used in digital libraries, and it is very important to be aware of this detail. In the digital libraries, we have defined usually metadata as simply "data about data" (which once stored in a database would mean, for the previous perspective, just the data inside a database, while metadata would be the information need to describe the organisation of that database). In this way, the "Internet community" took the term after the emerging of the World Wide Web, and now the most common usage for it is really in this area. Moreover, we should prefer for it the definition of "structured information about other information or resources".

However, even in this scope there are a few common misunderstandings around this term. For example, we must be very careful and stress that metadata is supposed to refer to information coded according a specific schema, and not the technology that handles it neither the conceptual spaces to control the values of the information elements. In this sense, MARCXML [26] or DCMES [11] are not metadata, but metadata schemas, e.g., definitions of how to express metadata as structured information about other information or resources. In the same sense, XML [40] in itself is just a technology, and not metadata or even a metadata schema. XML is a language where we can define schemas (by using a DTD - Document Type Definition, or more recently by using the XML Schema language [44]). In addition, authoritative spaces, such as indexing languages, classification systems, etc., are also not metadata in themselves, but values or rules to find the right values to give to metadata elements!

In the recent digital libraries' activities and literature, we can find several examples of different classes of metadata, namely:

- Bibliographic description of the resources: Bibliographic description and identification of the resources, such as titles, authors, indexing terms, classification, abstracts, surrogates, etc.;
- Administration of the resources: Administrative information about the resource, such as information about acquisition process and costs, rights, etc.;
- Preservation of the resources: Technical or management requirements for long-term preservation;
- Technical and structural description of the resources: Technical requirements to manipulate the resource (systems and tools), etc.;
- Access, usage and reproduction of the resources: Information about how to access the resources (addresses, passwords, etc.), terms and conditions for access and reproduction, etc.;
- Administration of the metadata: Information about the other metadata classes, such as data of creation, origin, authenticity, etc.

The bibliographic description of resources is a common issue in traditional libraries and archives, where respectively the MARC family of schemas [21][26] and the EAD schema [24] are widely used. The world outside these traditional scopes is also moving,

creating description models that, once in place, might be reused at low cost. One interesting example of that is the ONIX metadata descriptive format, defined by a publishers' consortium [17].

More recently, there were identified more requirements for metadata than just for bibliographic description. Relevant are for example the efforts for the technical description of the resources [42], new approaches for the classification and relation between resources [41][39], for preservation [6][32][36], for rights management [9], etc.

Other relevant actions have been the development generic frameworks aiming at covering several classes of metadata. One interesting example is the definition by the Library of Congress, in the United States, of the METS schema, aiming to cover bibliographic, structural and administrative metadata [27]. Another interesting purpose is that of the MPEG - Moving Picture Expert Group [31]. Especially relevant is MPEG-7 [10] and more generically MPEG-21 [5], which gives a special attention to the scopes of "Digital Item Declaration" (a generic metadata package), "Digital Item Identification and Description" (identifiers, bibliographic and technical description) and "Intellectual Property Management and Protection" (administration, access and usage of resources).

At this high level, similar to MPEG-21, are also the reference models CIDOC [7], a mediation framework to promote the interoperability in museums using heterogeneous descriptive metadata, MoReq [19], a model requirements for the management of archiving electronic records, and the well known FRBR - Functional Requirements for Bibliographic Records [20], promoted by IFLA. These are not specific metadata schemas, but very important guidelines for their definition, in the same sense as the AACR have been important to the development of bibliographic standards, systems and services in libraries [1].

Metadata and Technology

Another important issue that we must take in account when we discuss metadata is the relationship of the concept with the technology. In a general sense, a conceptual model or a metadata schema should be independent of any technological implementation. That is not always true, however, since sometimes we see examples were, especially for the sake of illustration and a better understanding (and to help on its immediate application), models are accompanied by specific technological solutions. That is what happened with MARC and ISO2709 [22], which did not obstruct the actual definition of MARCXML.

To proceed with this discussion we will propose a reference model of four principal perspectives: conceptual, context, service and technology. Figure 3 illustrates that model.

The Conceptual Perspective is where the generic reference models are considered. Here, we do not have yet records, databases or data files, but only concepts and models about how things can or should be done. We can subdivide this perspective in three scopes: generic reference models, which are supposed to define an objective top-down model; metadata schemas, which should be related with a specific issue or area of application (but that should still be independent of the technology); and metadata implementations, where finally technological issues are addressed (especially for coding).

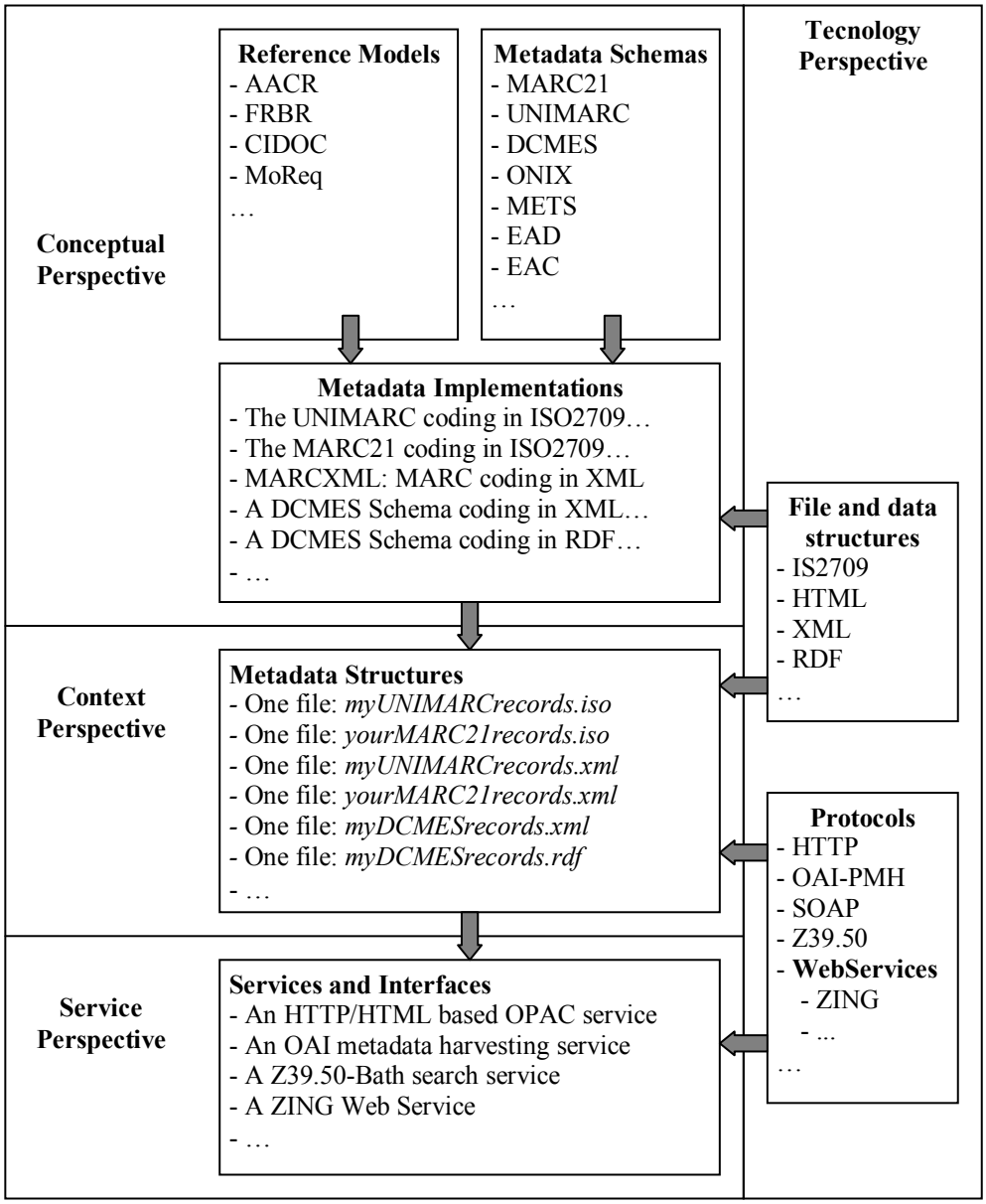

Figure 3: Multiple perspectives for the "metadata" problem.

The Context Perspective means the instantiation of the Conceptual Perspective. The ways those instantiations are done depend of technological options or constraints, as also of the nature and characteristics of the local services. For example, in one specific context we can decide to transport and explore a set of UNIMARC records coded in ISO2709, while in other context we might decide to achieve the same results but using MARCXML (a simple example of how that can be done through can be tested at [37]). The objective value and meaning of the information processed in both these solutions is the same, only the technical implementations are different. That might be dictated by the technology to be used in the final service, or by the legacy or new systems with which the new solution is expecting to have to interact.

This drives us to the last level, the Service Perspective. Here we deal mostly with interfaces, for humans or for other systems (protocols). Usually it is not irrelevant for a protocol what is the coding format of the metadata to be transported, but the tendency has been to make that as much flexible as possible. One example of that is the protocol OAI-PMH [34], which specified Dublin Core as its default format, but that has been evolving in order to support any other format possible of being expressed in an XML schema. We expect that a complete generalization of this will be achieved by the concept of Web Services [43], of which ZING, the next generation of Z39.50 is a potentially very interesting example [29].

Metadata in the Information Society

It is time to put a fundamental question: if metadata is an answer, what is after all the question? What are the fundamental requirements of the digital library to which a concept as "metadata" is supposed to provide a solution? We will find those requirements in three major classes:

- Heterogeneity of genres: The new information artefacts are not anymore simple and stable genres, as are the printed books, magazines or newspapers. A large heterogeneity and dynamism of new objects and models of artifacts have characterized the reality of the “digital publishing”. To deal with this in a technical and cost-effective way, the digital library must expect and understand clearly each class of objects and models. Media, data formats, versioning, type, etc., are examples of characteristics that can define new genres of resources. Genres are important for the definition of selection criteria for licensing, acquisition and deposit, independently of their subject, intellectual or artistic contents. To help the library to deal with those problems we have the concepts of structural and technical metadata, for example.
- Interoperability: The digital library is part of the World Wide Web. In this scenario, the users expect not only to reach the library from anywhere, but also to reach anything. This means that users might not understand very well (and not accept it at all), if they are said that they can not use a unique service to search on the same time in a library and in a film archive and access books and movies created by and about, for example, Federico Fellini. In order to be able to offer services of this kind, the digital library, now understood not only as an evolution of the traditional library, but has a conceptually higher level service, as we defined in the beginning, needs to be designed as a distributed service, or as an aggregation of heterogeneous services (Figure 4). This requires cooperation from generic and specialized libraries and archives, museums, and other classes of organizations and actors. Once again, the ability to automate this interoperability is crucial for its cost and technical effectiveness, bringing requirements for new classes of interfaces and metadata, defined or simply adopted by those actors. That has been done traditionally by means like Z39.50 [28], complemented recently by new models and solutions involving bibliographic records in XML [26][37], taking advantage of simple structures such as Dublin Core [11], or provide bulk of records for harvesting by OAI-PMH [33][34]. This is technology that was especially conceived by digital

libraries communities, but for the future we must start thinking in scenarios reusing generic solutions.

- Technology: As the Semantic Web develops, and its technology becomes more generic and ubiquitous, an important part of the components and products applied in digital libraries will be not specific of that anymore (traditional libraries are not very used with that generality). Those components will be generic, especially in what relates with user interfaces, database technology, protocols and Web Services. This means that metadata is not a concept specific of the digital libraries, but a general concept in any information system (which is in fact what a "digital library" is). Accordingly, the digital library communities must be effective in imposing their requirements in the definition of those components (working together for example with the World Wide Web Consortium, the International Organization for Standardization, etc.), but also open to reuse solutions that might have been defined and had became standards elsewhere. A golden rule in the actual world of the information technology is that it can be very expensive to provide a first new development for a specific problem, but after that, the cost of generalization of that solution can be very low. Libraries, museums and archives, which are always struggling with investment constraints, must take this is serious consideration!

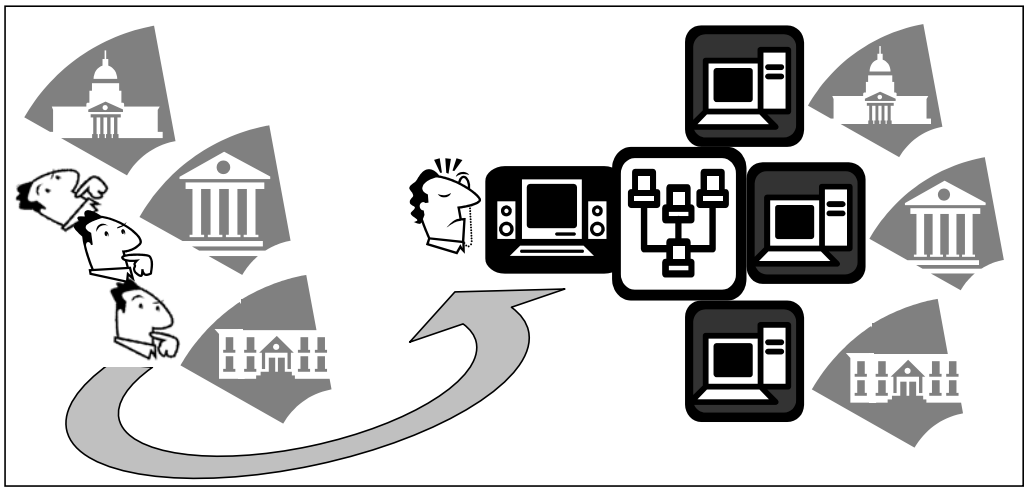

Figure 4: Interoperability in a networked world.

The Challenges

In simple terms, we can conclude that a main vision for the “digital library” has been that of a natural evolution of very well defined entities, with established interfaces, for a new less defined concept, required by a more dynamic environment. This will have important implications in some of the fundamental thinking in libraries museums and archives,, where authority control is just one of the examples.

Traditional libraries are used to recognize several actors relevant for bibliographic description [25]. The new paradigm makes it necessary not only to keep those concepts, but also to extend the analysis to reconsider now other key issues related for example with authentication, ownership, copyright, access control and authority control in general.

Several international actions have been analyzing those problems, namely the INTERPARTY project [17], the DC-Agent activity [11], and more generic the DELOS/NSF working group on Actors in Digital Libraries [14]. In the specific context of the archives, the EAC - Encoded Archiving Context [18][8] is an interesting work in authority description that should be followed with attention by everyone. That has been done for example in the LEAF project [23], a follow-up of MALVINE [30]. These projects, were the National Library of Portugal is an active partner, are interesting demonstrations of how heterogeneous schemas and sources of metadata can be combined in common services, with relevant benefit. In LEAF, we expect to demonstrate that having to deal with heterogeneous descriptions of authorities must not be always as a problem. In fact, we will even take advantage of that to improve other rich descriptions, and to improve the recall in resource discovery tasks in MALVINE and in other project, TEL [38].

Finally, I think that an important lesson to be learned from this discussion and to be transmitted to any discussion more centered on authority control issues is this: deal with heterogeneity! The digital library can not ignore the new centers of gravity, including popular resources providers as the on-line bookstore Amazon, gateways as Yahoo, and generic resource discovery services such as Google. I don't think that libraries should ignore these and other similar new actors entering at any moment in the Information Society, where some might represent very important potential new partners, bringing valuable new resources or services. In scenarios like this the key word for the libraries has to be "adaptation", meaning the capability to interface and interoperate in order to take the best from each relation without imposing strict rules that would be too costly for the other partner (keeping them away). The technology is powerful enough to deal with that! This assumption, when applied to scenarios of authority control, means that the problem might be not anymore how to conceive and put in place processes that drive to unique rules, descriptions and formats, but instead be able to understand the rules, description and formats used by the others and take from them the best we can for our purposes (and also be able to give the best of us to our partners).

References

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

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

[3] Borbinha, José. The Digital Library - Taking in Account Also the Traditional Library. Elpub2002 Proceedings, VWF Berlin, 2002, p.p. 70-80.

[4] Borbinha, José; Campos, Fernanda; Cardoso, Fernando. 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”, Idea Group Press, USA, December 1999.

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

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

[7] CIDOC. CIDOC Conceptual Reference Model. http://cidoc.ics.forth.gr/

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

[9] Creative Commons. http://creativecommons.org/

[10] Day, Neil; Martínez, José M. Introduction to MPEG-7. ISO/IEC working group JTC1/SC29/WG11/N4325. Version 3.0, July 2001.
[11] DCMI. Dublin Core Metadata Initiative. http://www.dublincore.org
[12] 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.
[13] DELOS. Network of Excellence on Digital Libraries. http://www.ercim.org/delos/
[14] DELOS. Reference Models for Digital Libraries: Actors and Roles. http://www.delos-nsf.actorswg.cdlib.org/
[15] DiTeD. Digital Thesis and Dissertations. <http://dited.bn.pt>
[16] D-Lib Forum. http://www.dlib.org.
[17] EDItEUR. http://www.editeur.org.
[18] Encoded Archival Context (EAC). http://www.library.yale.edu/eac/
[19] IDA. Model Requirements for the Management of Electronic Records (MoReq) http://www.cornwell.co.uk/moreq
[20] IFLA. Functional Requirements for Bibliographic Records. www.ifla.org/VII/s13/frbr/frbr.htm
[21] IFLA. IFLA Universal Bibliographic Control and International MARC Core Activity (UBCIM). http://www.ifla.org/VI/3/ubcim.htm
[22] ISO. ISO 2709: Documentation format for bibliographic information interchange for magnetic tape. ISO 1981.
[23] LEAF. Linking and Exploring Authority Files. http://www.leaf-eu.org/
[24] LOC. Encoded Archival Description (EAD). http://www.loc.gov/ead/.
[25] LOC. MARC Code Lists for Relators, Sources, Description and Conventions. http://www.loc.gov/marc/relators/
[26] LOC. MARC Standards. http://www.loc.gov/marc/
[27] LOC. METS - Metadata Encoding & Transmission Standard. http://www.loc.gov/standards/mets/.
[28] LOC. Z39.50 Maintenance Agency. http://www.loc.gov/z3950/agency/
[29] LOC. ZING, Z39.50-International: Next Generation. http://www.loc.gov/z3950/agency/zing/.
[30] MALVINE. Manuscripts and Letters via Integrated Networks in Europe. http://www.cordis.lu/libraries/en/projects/malvine.html
[31] MPEG. Moving Picture Expert Group. http://www.cselt.it/mpeg.
[32] NEDLIB. <http://www.konbib.nl/nedlib>
[33] OAF. Open Archives Forum. http://www.oaforum.org/
[34] OAI. Open Archives Initiative. http://www.openarchives.org/
[35] OMG. Catalog of OMG Specifications. http://www.omg.org/technology/documents/spec_catalog.htm
[36] PANDORA. Preserving and Accessing Networked Documentary Resources of Australia. http://pandora.nla.gov.au/
[37] PORBASE. Protótipo de acesso por URN à PORBASE. http://urn.porbase.org
[38] TEL. The European Library. http://www.europeanlibrary.org/
[39] Topic Maps. Topic Maps Consortium. http://www.topicmaps.org/
[40] W3C. Extensible Markup Language (XML). http://www.w3c.org/XML/
[41] W3C. Resource Description Framework (RDF). http://www.w3.org/RDF/
[42] W3C. Synchronized Multimedia Integration Language. http://www.w3.org/TR/REC-smil/
[43] W3C. Web Services Activities. http://www.w3c.org/2002/ws/
[44] W3C. XML Schema. http://www.w3.org/XML/Schema