The Digital Library

Taking in Account also the Traditional Library

José Luis Borbinha

National Library of Portugal

Abstract. This paper presents an analysis of the problem of the "digital library", taking in consideration the perspective and requirements of the "traditional library". The purpose is to help the approach between two extremes of the problem: the well-established communities of the traditional library and the communities working in new involving technological and business models. In that sense, it is proposed and described a model for the digital library based in the generic use cases of: publication; licensing; acquisition; registration; dissemination; searching; access; and preservation. These cases are also related with new emerging issues: genres; actors; identifiers and metadata.

1 Introduction

Traditional libraries are well established institutions with a long record of strong efforts in the definition of clear missions, scopes, procedures and standards. Meanwhile, the introduction of computers in libraries, and especially the Internet, has brought to our attention new terms for the library, such as the computerized library, the library on-line, the digitized library, the virtual library, etc. This evolutionary process has been defining the generic scope of the "digital library", with the emerging of a new community of researchers and practitioners.

These efforts have been followed by the traditional library, but the agenda has been imposed mainly by the computer science and engineering community. That is a logical consequence of the nature of the process, but we must be aware that, unless we want to promote strong disruptive scenarios, the changes must be conducted not only by the technological evolution but also by the clear understanding of the existing models and its requirements. This might be controversial, since one might argue that the requirement to observe the old models can be a strong constraint to innovation, especially in the scope of a context that will be, in the end,

clearly disruptive. That is correct, but we should recognize also that proposing new scenarios without the correct analysis and understanding of the established ones and especially without strategies for the transition might be not a clever and responsible attitude.

This paper intends to contribute the solution of this problem, by presenting a common model of the digital library taking in account both the perspectives, resulting in a model in the generic use cases of: publication; licensing; acquisition; registration; dissemination; searching; access; and preservation.

Finally, the paper makes describes the perspective and main initiatives related with this problem at the National Library of Portugal (BN - Biblioteca Nacional). This was also the perspective chosen for the references presented in the paper, which therefore are not intended to present the state of the art in their scopes, but mainly the projects or initiatives where BN, as a "traditional deposit library", has been participating or following with special interest.

2 A Generic View of the Problem

A simple introduction to the problem can be illustrated by **Fig. 1**. Here we stress the Internet as the last big factor in the evolution of the "library", in the following of a series of previous factors. From those we stress the generic introduction of the computer in the library, which had an impact in the digital catalogue and with it the definition of the first standards for bibliographic description. Than we had the first data communication services (telnet, X.25, etc.), providing remote access to the catalogue and also to other common library's services. That was followed by the personal computer and the CD-ROM, which brought the digitized library providing now access to also the contents.

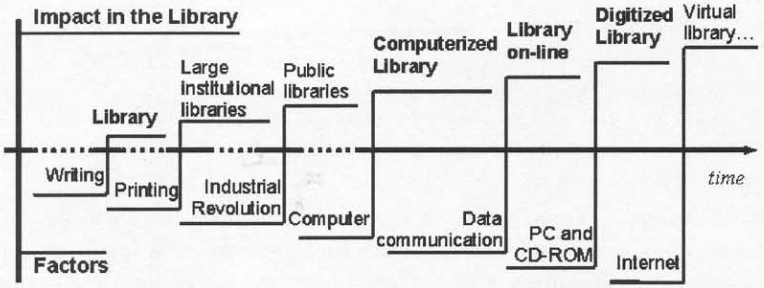

Fig. 1. *An evolutive perspective of the "Library"*

This evolution brought us to today, facing the problem of the definition of the "virtual library", or in a more common term, of the "digital library". What should it be a digital library?

This became a recent hot topic of discussion, with lots of demagogy but also with lots of real serious work, both conceptual and technical. For those interested in developing a complete view of those activities, discussions and visions, two important resources are the D-Lib Forum [12] and the DELOS Network [39].

From a generic technical perspective, the "digital library" has been seen as a possible specialization of the "information system", as proposed early in the classification system of the Association for Computer Machinery, resumed in **Fig. 2** [2].

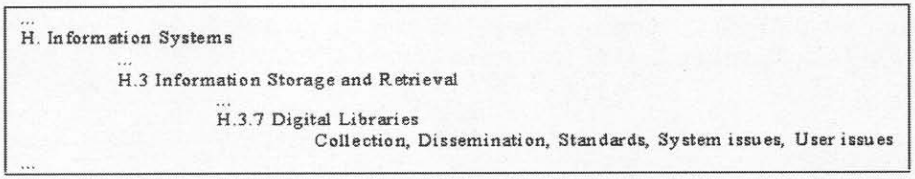

Fig. 2. *"Digital Libraries" in the ACM Computing Classification System (January 1998)*

A more recent proposal, resulting from a brainstorming report and reported by DELOS, is the one illustrated in **Fig. 3** [8]. This faces the problem from a wider perspective, but also less defined. In fact, when compared with the technical and conceptual images that we have of the traditional databases or information systems, we have to admit that the images we have been envisaging for the digital library are much broader and fuzzier.

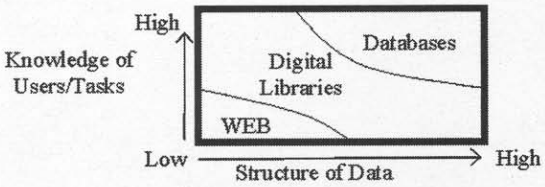

Fig. 3. *The problem of the "digital library" according DELOS*

The main difference for these different perspectives comes from the very instant of the definition of the solution. While for the development of a traditional database or information system we start with a clear identification of user requirements and usage scenarios, in an answer of specific needs felt in a specific context, in most of

the cases where we face a "digital library's challenge" those questions can simply not be put in the same way. In fact, libraries are almost by definition entities that have to be ready to deal not only with new incoming genres of contents, but also with unexpected users' requests. That has been true in the traditional libraries, and will be even truer in the "digital library", making it related also with the reality of the World Wide Web!

In simple terms, we can conclude that a main vision for the "digital library" has been that of a natural evolution of a very well defined entity, with established interfaces, for a new less defined concept, related with a more dynamic attitude, as sketched in **Fig. 4**.

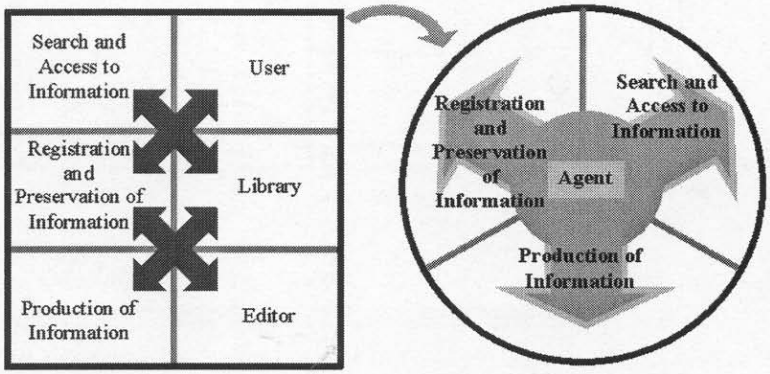

Fig. 4. *From the traditional model to the digital paradigm ...*

3 An Analysis of the Problem

This paper proposes to analyse this problem of the "digital library", but from what we might call a conservative perspective taken from the traditional libraries. The purpose is to try to contribute has a path-finder in the process of evolution of the perspective of these traditional libraries toward of a new digital paradigm.

The first step for the discussion of that analysis is given by the **Fig. 5** (an earlier but more detailed description of this model can be found in [3]).

The case of **publication** is related with the availability of a manifestation of a work. In the traditional paradigm those works are made available by traditional entities, the publishers. Now, as a result of the redefinition of the business models, the case might be more complex for the library, since it will be asked to interface with new actors. Examples of those new actors include the creators themselves, especially those ones that might prefer to take advantage of new business and

dissemination models for their works, not forcibly depending from immediate selling of the works.

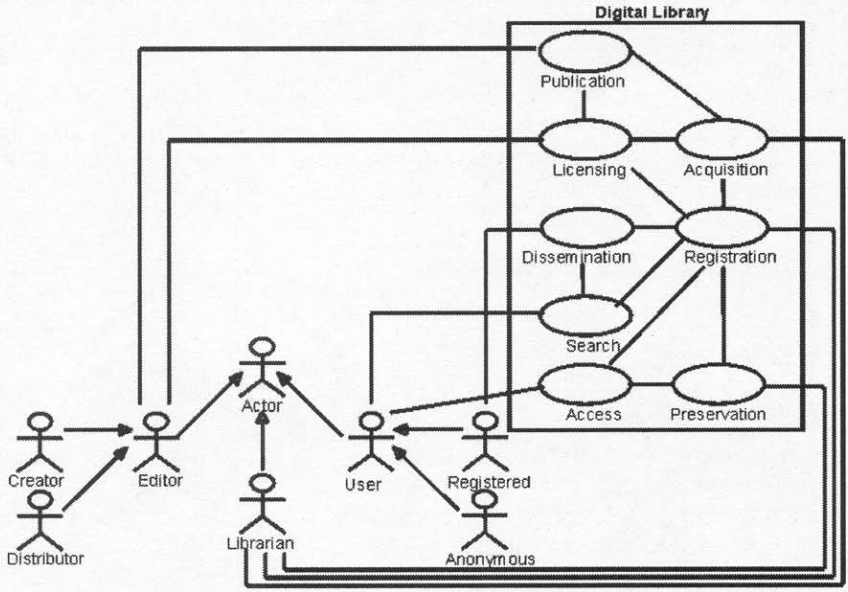

Fig. 5. *A generic model (use case) of the problem of the digital library*

The **acquisition** of a traditional publication is the set of actions of assuming the ownership of an item of a publication. In a traditional scenario the physical ownership of an item is a main requirement for a good quality of service (this implies even the acquisition of more than one item for high demanded resources). Services of inter library loan and remote copy make it possible for a user's library to access information resources from other libraries, but those are usually costly and slow. In a digital scenario it is possible for a library to acquire one digital copy of a resource, and make it easily accessible to its users. This is true in theoretical terms, since in the reality the technical complexities of the resource or the requirements for its local installation and access might be a great problem (file systems, database management systems, viewers, etc.).

A model that has been becoming popular to minimize this problem is the licensing of the remote access to the digital resources (a model with success in library's networks, where big consortia can manage to impose strong positions to the big publishers). In this case the libraries don't acquire resource items, but the right for their users to access them remotely. Interesting examples of this model are the ACM Digital Library [1] and the IEEE Xplore [15]. However, this model raises

some concerns in the long term: what happens to the investment made in the past by the library if it decides not to continue the subscription or if the publisher ceases to offer the service? To answer to this, some publishers provide to the library copies of the resources, but it is not sure that such will be technically compatible in the long term (problem of preservation).

The **licensing** is the negotiation of the rules for the access or the usage of an item of a publication. In the traditional scenario these terms are directly related with the model of acquisition of the items (buying and item makes it possible for the library to provide access to it to anyone) and the copyright laws (for private or charged copies). Everyone now that it is possible to abuse from this scenario, but along the time the market adapted itself to that and those possible abuses have been tolerated.

In the digital scenario the problem is much more complicated, essentially due to the ubiquity and easiness of the digital access and reproduction. In fact, in both the models of acquisition of ownership of items and licensing of remote access it can be very difficult for a publisher and a library to enforce licensing rules to avoid the illegal replication of the resources. This is an actual and widely discussed problem!

The **registration** is related with the collecting or production of metadata to describe a publication, its components and its items. That comprises the traditional tasks of indexing and cataloguing, including now new tasks specific of the digital paradigm. Some tasks can be automated, reusing embedded information provided with the publications or using special software applications for automatic analysis of the contents. However, it is necessary to address now also new requirements, such as the installation and de-installation of some of the publications (especially for those ones requiring executable software), the access (decryption keys, passwords, metadata for legal terms and conditions of use, etc.), preservation (file formats, logical structures of hypermedia publications, etc.), etc.

The potential for **dissemination** and announcement of the acquisition or licensing of a manifestation of a work or a service is one of the most relevant aspects of the Internet. The challenge here for the library is similar of any other organization embracing an "e-business" strategy.

The OPAC (On-line Public Access Catalogue) has been a fact since the "networked library". The new relevance now for the **search** in the library comes from the opportunities for interoperability of its systems with other external resource discovery services, augmenting the user's possibilities in the searching of information about a work or in generic serendipity tasks.

The **access** to a resource can be made easier now (there is no need anymore to pick it from the shelves), but the need to prevent the abuse in the usage of licensed material can be an important problem.

Finally, the generic problem of **preservation** is the keeping of the good shape of the physical, logical and intellectual contents of the items of a publication, as also of

its access forms. In the traditional paradigm this is usually related with the physical medium (the paper), but now the problem is much more complex. In a digital paradigm the preservation is raised from three perspectives: physical preservation, related with the instability of the media; logical preservation, associated with the need to assure format conversions when original formats become obsoletes or too expensive to maintain; and intellectual preservation, a concern mainly related with format conversion resulting in changes in the layout, presentation, or interaction with the publication.

4 New Emerging Issues

The discussion makes it possible to identify new emerging issues, namely: genres (what?); actors (who?); identifiers (where?); and metadata (how?).

The reality of the "digital publishing" has been characterized by a large heterogeneity and dynamism of objects and models. To deal with this in a cost-effective way, deposit institutions will need to identify clearly each object and model in order to understand their specific requirements, update their knowledge, and adapt their behavior just like any other technological player in the digital world. Media, data formats, versioning, type, etc., are examples of characteristics that can be combined o define genres (what?). Genres are important for the definition of selection criteria for deposit and preservation guidelines, but also for the immediate licensing.

Traditional libraries are used to recognize several actors relevant for bibliographic description [22]. The new paradigm makes it necessary the revising of those concepts, now with wider implications in other key areas related with authentication, ownership, copyright, access control and authority control in general (who?). Several international actions have been analyzing that problem, namely the INTERPARTY project [14], the DC-Agent activity [7], the DELOS/NSF working group on Actors in Digital Libraries [10], and the EAC - Encoded Archiving Context [21].

Traditional libraries are used with identifiers such as ISBN [18] and ISSN [19]. Now, the digital library needs to take also in account new established or emerging identification schemas (where?), such the URI and URN [37] concepts and their instantiations like the PURL [34], DOI [13] and OpenURLs [28].

Finally, we have now the generic problem of "metadata" (how?), with all the new perspectives illustrated in **Table 1**. The bibliographic description of resources is a common problem in traditional libraries and archives, where the MARC family of schemas is widely used [23, 16], or the EAD schema [21]. But the world outside these traditional institutions is also moving, creating valuable descriptions that can be reused at low cost, such as the definition of the new ONIX format, proposed by a publishers' consortium [14], relevant also for the administration of the resources.

Table 1. Typical types or classes of metadata

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

Other relevant activities are the efforts in address that with also the technical description of resources (SMIL [39], RDF [38] or Topic Maps [36], or the understanding the problem of preservation (NEDLIB [27], CEDARS [5], PANDORA [31], etc.). A generic framework aiming at covering a wide area of metadata are the activities of the Moving Picture Expert Group [26], especially of MPEG-7 [6] and more generically of MPEG-21 [4], 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). Other interesting activity has been the definition by the Library of Congress, in the United States, of the METS schema, aiming to cover bibliographic, structural and administrative metadata [24].

5 The Problem at BN

At BN the problem of the digital library as been addressed at four major levels: development of a consistent technical framework and know-how; systems and technology to give answer to the new requirements and deal and manage heterogeneous schemas of metadata; development of a strategy for digital deposit and preservation, and development of a wide program of digitalization of the printed resources.

The challenge of the metadata is the ability to provide OPAC services based on the integration and interoperability with heterogeneous services, both internal and external. BN is in charge of the national union catalogue PORBASE [32], an effort of 150 libraries traditionally using UNIMARC bibliographic databases. BN is assuming a leading position in helping those libraries to access not only traditional sources of metadata by traditional means (Z39.50 [25]), but also bringing to them new models and solutions, such as the ability to process records in XML [23, 33], deal with Dublin Core [7], or share them by OAI [30][29]. An important landmark in this area will be hosting of the IFLA's UNIMARC Program by BN after the end of 2002 [17].

Another important line of activities is the digital deposit. As a traditional deposit institution, BN started the study of the problem of the deposit of digital publications with the project NEDLIB - Networked European Deposit Library [27], which produced the structure of the problem presented in **Table 2** [3]. Deposit institutions should define deposit criteria for genres they can support and in real scenarios. Three of those scenarios at BN are the project DiTeD - Digital Dissertations and Thesis [11], the deposit of on-line periodic publications, and the deposit of digital monographs.

Table 2. Main functional requirements for the management of deposit of digital collections.

Use Cases	Main Scenarios	Related Requirements
Acquisition	Delivery by the publisher	
	Capture by the library	
	Harvesting by the library	
Verification	Medium integrity	
	Content integrity	Logical integrity, Authentication
Registration	*Metadata*	Bibliographic description, Installation and de-installation, Preservation, Access
Preservation	*Physical preservation*	Medium refreshing, Medium migration
	Logical preservation	Format conversion, Emulation
	Intellectual preservation	
Access	*Conditions of use*	Local access, Remote access

In parallel with these activities, there are also important efforts at BN in the digitization of its most relevant printed works and manuscripts. A specific project is

aiming a creating nearly one million pages by the end of 2003. All the deposited and digitized works will have a normal bibliographic UNIMARC record in PORBASE.

Concerning the problem of the identification for these works (both deposited and digitized), BN decided to develop an Internet based PURL space in the domain <purl.pt>.

This is a very simple schema, in the form of <http://purl.pt/x>, where "x" is a growing integer, starting at 1 (therefore, <purl.pt/1> is the URN for the first registered digital work, which happens to be a digitized copy of the first printed edition, from 1572, of "Os Lusíadas", the classic work of the national poet Luiz de Camões).

Finally, BN is also involved in other important initiatives in the areas of resource discovery and access, such as the international projects LEAF [9] and TEL [35].

References

1. ACM. **ACM Digital Library**. http://portal.acm.org
2. ACM. **ACM's Computing Classification System**. http://www.acm.org/class/
3. 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.
4. Bormans, Jan; Hill, Keith. **MPEG-21 Overview**. ISO/IEC working group JTC1/SC29/WG11/N4318. Version 0.2, July 2001.
5. CEDARS. **Curl exemplars in digital archives**. http://www.leeds.ac.uk/cedars/
6. Day, Neil; Martínez, José M. **Introduction to MPEG-7**. ISO/IEC working group JTC1/SC29/WG11/N4325. Version 3.0, July 2001.
7. DCMI. **Dublin Core Metadata Initiative**. http://www.dublincore.org
8. DELOS. **Digital Libraries: Future Directions for a European Research Programme. Brainstorming Report**. San Cassiano, Alta Badia - Itally. June 13-15, 2001. http://www.iei.pi.it/DELOS/delo2/International/brainstorming.htm.
9. DELOS. **Network of Excellence on Digital Libraries**. http://www.ercim.org/delos/
10. DELOS. **Reference Models for Digital Libraries: Actors and Roles**. http://www.delos-nsf.actorswg.cdlib.org/
11. DiTeD. **Digital Thesis and Dissertations**. <http://dited.bn.pt>
12. D-Lib Forum. http://www.dlib.org.
13. DOI. **Digital Object Identifier**. http://www.doi.org/
14. EDItEUR. http://www.editeur.org.
15. IEEE. **IEEE Xplore**. http://ieeexplore.ieee.org/.
16. IFLA. **IFLA Universal Bibliographic Control and International MARC Core Activity (UBCIM)**. http://www.ifla.org/VI/3/ubcim.htm

17. IFLA. **Statement on UNIMARC.** http://www.ifla.org/VI/3/u-statement.htm
18. ISBN. **The International ISBN Agency.** http://www.isbn.spk-berlin.de/
19. ISSN. **International Standard Serial Number.** www.issn.org/
20. LEAF. **Linking and Exploring Authority Files.** http://www.leaf-eu.org/
21. LOC. **Encoded Archival Description (EAD).** http://www.loc.gov/ead/.
22. LOC. **MARC Code Lists for Relators, Sources, Description and Conventions.** http://www.loc.gov/marc/relators/
23. LOC. **MARC Standards.** http://www.loc.gov/marc/
24. LOC. **METS - Metadata Encoding & Transmission Standard.** http://www.loc.gov/standards/mets/.
25. LOC. **Z39.50 Maintenance Agency.** http://www.loc.gov/z3950/agency/
26. MPEG. **Moving Picture Expert Group.** http://www.cselt.it/mpeg.
27. NEDLIB. <http://www.konbib.nl/nedlib>
28. NISO. **OPEN URL Standards Committee AX.** http://www.niso.org/committees/committee_ax.html
29. OAF. **Open Archives Forum.** http://www.oaforum.org/
30. OAI. **Open Archives Initiative.** http://www.openarchives.org/
31. PANDORA. **Preserving and Accessing Networked Documentary Resources of Australia.** http://pandora.nla.gov.au/
32. PORBASE. **Base Nacional de Dados Bibliográficos.** http://www.porbase.org.
33. PORBASE. **PORBASE Access by URN.** http://urn.porbase.org
34. PURL. **Persistent Uniform Resource Locator.** http://www.purl.org
35. TEL. **The European Library.** http://www.europeanlibrary.org/
36. Topic Maps. **Topic Maps Consortium.** http://www.topicmaps.org/
37. W3C. **Naming and Addressing: URIs, URLs, ...** http://www.w3.org/Addressing/
38. W3C. **Resource Description Framework (RDF).** http://www.w3.org/RDF/
39. W3C. **Synchronized Multimedia Integration Language (SMIL) 1.0 Specification.** http://www.w3.org/TR/REC-smil/