Proceedings of the 5th WSEAS Int. Conf. on SIMULATION, MODELING AND OPTIMIZATION, Corfu, Greece, August 17-19, 2005 (pp146-151)

An Infrastructure for a National Digital Library

JOSÉ BORBINHA
INESC-ID – Instituto de Engenharia de Sistemas e Computadores
Rua Alves Redol 9, Apartado 13069, 1000-029 Lisboa
PORTUGAL
http://www.dei.ist.utl.pt/~jlb

Abstract: This paper describes the infrastructure of the National Digital Library in Portugal. The requirements emerged from the definition of the resources to be manipulated and the services to be supported. It was given a special attention to scalability, as also to community building open standards, open solutions, and reusable and cost effective components. The structural metadata for the content resources is METS, and generic bibliographic metadata format is UNIMARC, EAD or Dublin Core. The resources identifiers are processed and resolved as simple but very effective PURL identifiers, supported by a specialized resolution service. The storage for immediate access is provided by open file systems, such as LUSTRE. The storage for long-term preservation it supported by ARCO, a locally developed solution. All the components can run on GNU/Linux systems.

Key-Words: **Digital Libraries, Information Systems, Structural Metadata, Descriptive Metadata, Interoperability, Digital Preservation, Digitization, Open-Source.**

1 Introduction

The National Library of Portugal (BN - Biblioteca Nacional) is a patrimonial library. Digital technology has been presenting new challenges to these institutions (mainly national libraries), mainly related with digitization; digital publishing; digital preservation; creation, processing, quality control and exchange of metadata; services for resource discovery, access and interoperability; etc. Each of these areas brings new expectations, but also new requirements for new skills and competencies, which BN has been identifying and mastering during recent years in trials and pilot initiatives. As a result, valuable expertise has been built, resulting in the actual initiative for the National Digital Library (BND - Biblioteca Nacional Digital).

A key part of the BND initiative is the development of a technical infrastructure, here described. This action started in September of 2003, and the first stable versions of the services have been released during 2005. This paper describes the main requirements and the global architecture for all of that.

The paper starts with a description of the main technical and strategic requirements, followed by the description of the architecture.

The most important elements of the overall model are the structural metadata (the "digital binding" for the information objects); the descriptive metadata (the description of the information objects); the names and identifiers (the references to the objects); and finally the repositories (storage and preservation). These issues are presented and discussed in the subsequent points, including the description of the main services to be supported by the related infrastructure.

2 Requirements

The BND is an initiative with four simultaneous areas of activity:

- Development of new digital resources
- Services for deposit and registration of digital resources
- Services for resource discovery and access
- Services for long-term preservation

Common to these areas is the building of a central infrastructure, which main technical and strategic requirements follow.

2.1 Technical Requirements

Creating digitized copies of manuscripts or printed items is one of the most interesting areas of actual investment for patrimonial libraries. As a result, multiple digitization projects are bringing for general access invaluable and so far inaccessible treasures.

Besides these digitized objects, libraries have to consider also nowadays the new original and digital born objects, ranging for example from on-line journals and newspapers, published in a wide range of models, until innovative web sites representing new genres of resources, yet to be named.

This scenario leads us to the main requirements for the architecture of BND: it has to be a framework to support services for deposit, registration, storage (for access and also for preservation), resource discovery, and access, all of this for a potentially large diversity of genres of information objects (in size, types, technical characteristics, conditions of use, etc.). Other issues for main requirements are cost, maintenance and scalability (also in dimension and in time).

2.1.1 Requirements of Digitized Objects

BND has promoted several digitization projects, covering a wide range of originals: newspapers, journals, posters, drawings, printed books, incunabula, manuscripts, maps, etc.

Most of those images are digitized in high quality, with 24 bits in color depth and resolutions from 300 to 600 dots per inch (dpi). These images are registered in master files in TIFF format, representing each one from a few kilobytes until a few gigabytes.

Each digitized image can be also part of a specific work. Therefore, these works can range from one simple file of a few kilobytes until a collection of multiple files, comprising possibly several gigabytes in total.

Most of these digitized works are in the public domain, since one of the main purposes of these digitization actions is to promote the access to that class of works. The high-quality master images are appropriate for reproduction and local fruition, but are not suitable for access by the Internet. For that, we produce lower resolution copies, in PNG, JPEG or GIF formats (reducing also the color depth).

The digitized images rich in printed text are also submitted to automatic optical character recognition (OCR). This produces immediate text for automatic indexing, but for the works of high relevance, that text is also corrected by humans and made available as alternative transcript copies (the actual OCR technology stills producing a high level of errors for most of these originals, the major created before the end of the XIX century, making this human correction very expensive).

As a result of this process, we have in the end of it, for each digitized work, a potential list of multiple copies. Those copies can range from a very simple object (just one image) until very complex objects, for books made for example of multiple chapters or parts, and digitized initially in hundreds or even thousands of image files. Figure 1 shows an example of an on-line record of one of those resources, representing a digitized book with copies ("Exemplar") in PDF, TXT and JPEG, besides a private master copy in TIFF.

To manage this complexity, we need therefore to record also the logical structure of these works. For that purpose we started by developing a specific XML schema, named DPCreator [1][2], but during 2004 we changed to the METS [13], an alternative emerging schema with growing usage in digital libraries.

Until the end of 2005 the BND will reach a total of nearly three thousand titles and one million of digitized images, all in more than 50 terabytes. These images are produced not only locally, but also by other digitization projects in the country that deposit their results at the BND, as a safeguard.

Fig. 1. The on-line record (home-page) of a digitized printed work, available in multiple digital copies.

2.1.1 Requirements of Digital Born Objects

Besides the digitized works, BND is also concerned with the new works, born digital. The deposit and registration of these objects represents a real social need, to assure their registration and dissemination, and especially their preservation, an issue under the scope of any patrimonial library. This means for BN the requirement to expand its scope as a deposit library, covering not only the traditional printing world but now also the new digital paradigm [4].

National libraries are generally mandated to maintain deposit collections of published documents, usually for the purpose of preservation of cultural heritage. Through this mission, those institutions are supposed to guarantee the long-term availability to those manifestations of intellectual works.

Depending of the country, the deposit framework is defined usually based in Legal Deposit (a system legally enforced, whereby authors, publishers or other agents must deliver one or more copies of every publication to the deposit institution), Voluntary Deposit (a system usually based on agreements between the deposit institution and the publishers or authors, under which those agents deliver one or more copies of each publication for preservation) or Acquisitions (a system where the deposit institutions have to take the initiative to identify, select and

acquire the publications relevant for deposit, according to their mission and strategy).
Portugal has a Legal Deposit law for printed works, but BN is not in favor of any change in this law toward mandatory scenarios for the deposit of digital born and online published resources. However, it is much in favor of any model that will give the right to harvest (acquire) those resources, as also to promote self-deposit by their creators (voluntary deposit).
In this sense the BND will have to consider the deposit and storage of a wide range of digital born objects, from simple files (but possibly in several formats) until complex web sites. The requirements from this are on techniques and technology for web harvesting; services for voluntary deposit; and metadata for description, logical structures and access control.

2.1 Strategic Requirements

Another important requirement is stability. BND is not anymore an experimental project, pursuing short-term results, but it has to be stable and trustworthy. That requires thinking seriously in proven models and long-term sustainable strategies for the developments and convergence with the traditional existing services in the library. Also, it has to be all done in accordance with the human and financial resources that BN will be able to commit for these new operational areas, taking especially in account long-term implications and dependencies.
Finally, we must take in consideration that the challenge of developing new services and collections of digital resources are not specific of any particular library, but an actual generic issue for any library, archive, and even in museums. This generic challenge must be a motivation for cooperation, technology reuse and service's integration (interoperability).
As a result of these generic requirements it was taken the strategic decision for open standards and open source software solutions.

2.2 Remarks

This paper is focused on the generic technological issues and the generic technical framework of BND. We must stress that several specific issues are not deeply addressed here, such as workflows, processes for deposit (harvesting and voluntary deposit scenarios), registration (creation of descriptive metadata, management of identifiers, etc.), storage (an issue were long-term preservation issues are very relevant) and access (especially models for terms and conditions).
Also, important but not here addressed, is the issue of the deposit of complex digital born resources, such as for example newspapers published by sophisticated content management systems. We believe that the requirements of those scenarios will be able to be supported by the infrastructure here described, since the main problems will be found mainly in the external workflows, with implications in the creation and maintenance of the descriptive and structural metadata. However, it is not our intention to demonstrate that or discuss it now.

3 System Architecture

The main requirements of this problem lead us to design the architecture for the BND, which main generic components are shown in a very simple way in the Figure 2.
The main services that this infrastructure is expected to support are, from the professional and management perspective, the deposit and registration of resources, their storage in the appropriated repositories, as also the need to manage and resolve their identifiers.
From the public perspective, the main services are the search and browsing (resource discovery), as also the access to the resources.

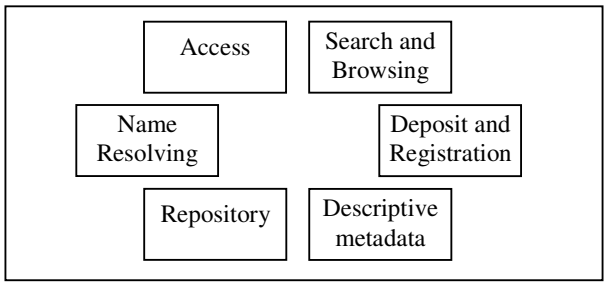

Fig. 2. A simple generic model of the main components of the architecture of the BND.

4 Deposit and Registration

The cases of deposit and registration comprise mainly the creation of metadata for the resources, the registration of that metadata, and the submission of the resources to their appropriate repositories.

4.1 Structural Metadata

All the objects sent to the repository are structured according to the METS schema. The METS metadata records can vary in detail according the object type. For example, a specific profile is used for the self-deposited digital born objects, for which we create automatically a simple but very useful record, with the descriptive and technical information. For harvested on-line resources (typically websites) a limited structural map is also automatically created. On the other side, for digitized objects it is possible, in most of the cases, to create very rich detailed structural maps and descriptive information.
The creating of structural metadata is supported by a framework named ContentE [6]. This can be used to

create very rich METS files, with structural descriptions for digitized objects, as shown in Figure 3. ContentE exists in two versions: as a standalone tool, for human operation, with the interface shown in Figure 4 (to edit the automatically created metadata of the digital born objects.); and as web service (offering functionalities to other applications and services). ContentE is released by BN as an open-source tool, being already used in Portugal by several libraries and archives.

In the METS files it is possible to include or make reference to descriptive metadata. The METS community defined also the METSRights schema for terms and conditions, which is also used in BND.

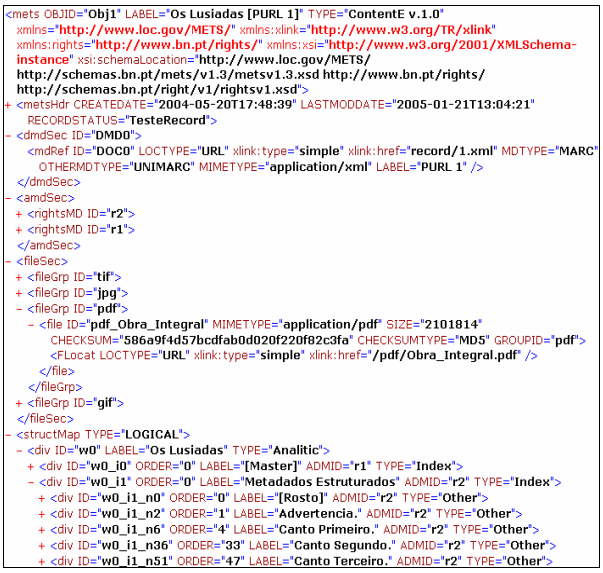

Fig. 3. An example of a METS structural file, created by ContentE for the work available at http://purl.pt/1.

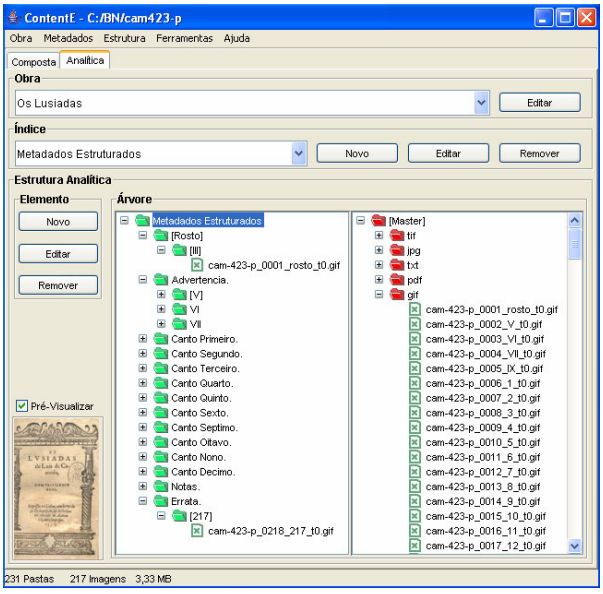

Fig. 4. The ContentE tool in its standalone version.

4.2 Voluntary Deposit

For Voluntary Deposit services we developed the DEPTAL framework [5]. DEPTAL is a collection-centric solution that can manage collections of documents in multiple copies and types, where each object is organized as an HTML site. It can manage also users and groups of users, it supports authority control (subjects, authors, etc.), and can interoperate by interfaces such as OAI-PMH [17], Z39.50 and SRU [14], and customized web services. DEPTAL recognizes descriptive metadata (UNIMARC, Dublin Core), and organizes the information objects in METS.

DEPTAL is especially focused for institutional repositories [11], being already used not only by services of the BND but also by several universities in Portugal.

4.2 Resource Description

The digital objects with bibliographic characteristics have standard descriptive records in PORBASE, the national union catalogue. PORBASE is a traditional bibliographic catalogue, based in UNIMARC [10]. It holds more than one million of bibliographic records, and half of a million of authoritative structured records of authors, families, organizations and subjects.

Other resources with archival characteristics, especially if they were already registered in the BN archiving system, can be processed in EAD [12].

Complex and unusual objects or collections, such as deposited or harvested groups of "blogs", archived mailing lists, etc., which traditional cataloguing or description would be not immediately effective, can be registered in Dublin Core [7].

5 Identifiers and Name Resolving

The Name Resolving service is based on the PURL concept [18], which ensures that the URL made for these objects will still be valid in the future. This service has three main components:

- An HTTP proxy that translates each PURL into the real URLs where the object must be accessible (which can change dynamically from access to access, or from user to user);
- A web user interface for the human access and back-office maintenance of the service;
- A web service interface used by other applications for resolution of identifiers and also management of the information.

6 Repository

The Repository has two main objectives: to assure the safe storage and preservation of the resources; and to assure the access to the resources.

Budget constraints, prior experiences and long-term maintenance concerns limited our solutions to the use of commodity hardware and free software like GNU/Linux.

The use of commodity hardware, like ATA disks, is considered a possibility thanks to the standardization of technology that allows inspection of environmental and operation conditions of these devices. This includes temperature sensors, working errors, log of failures for IDE disks (through the SMART interface), etc. This ability to watch the environment and detect errors before they can affect the data gives us the minimum assurance needed on such a critical project.

A final requirement was that the system had to be thought out to allow also the use of heterogeneous hardware configurations, to make it possible an easy and natural long-term scalability of the storage space.

After these considerations, we decided to design a solution that would fulfill all these requirements.

The preservation repository is implemented by a local solution named ARCO. This is basically a GRID solution, where terabytes of data can be stored in clusters built of commodity components [9][8].

One ARCO installation is a uniform storage space of several storage nodes and one metadata server. Each resource object is stored under a directory, without any data encodings or transformations.

The objects are also replicated in two nodes, providing mirror security. Each copy is always stored entirely in one storage node, allowing the full recover not only after a failure of a node but even in the catastrophic scenario of the lost of the metadata server.

The access repository uses the LUSTRE file system for its low level data storage [16]. LUSTRE is a highly scalable POSIX compliant file system designed for high-performance computing using also clusters of commodity hardware and GNU/Linux. It is a very scalable system, providing a fast access if necessary (depending of the characteristics of the hardware).

7 Deposit and Registration

The services for deposit and registration support the workflows for the verification, cataloguing or description, and storage of the objects.

Besides the deposit of the digitized works and of the digital born works developed by BN itself, BND has been promoting several specific projects, such as the self-deposit of thesis and dissertations [3], the deposit of periodical publications, in specific scenarios [4], or associated to other initiatives such as LOCKSS [15]. This includes also cases for the deposit of full collections, such as the Gutenberg collection, deposited at BN [19].

8 Search and Browsing

The service for search and browsing offers several indexes for browsing, an OPAC for the search in the digital collection (using the descriptive metadata, as shown in Figure 5). The indexes are built automatically by the PURL service, using the subject indexing information, authors and dates in the descriptive metadata, as shown in Figure 6.

The search engine, MITRA, is based on LUCENE [20]. This is a "Google like", but more "intelligent" search engine, able to index not only the full content of a wide range of text rich formats, but also the UNIMARC and Dublin Core metadata records (other schemas, such as EAD, are being added).

MITRA can index web sites and objects in any XML schema. This provides innovative added value, since it indexes the content of the objects as also their metadata structures, bringing together the best from the two worlds.

Other search and interoperability interfaces are provided by SRU, Z39.50 [14], and OAI-PMH [17].

Fig. 5. The OPAC (On-line Public Access Catalogue) at the BND.

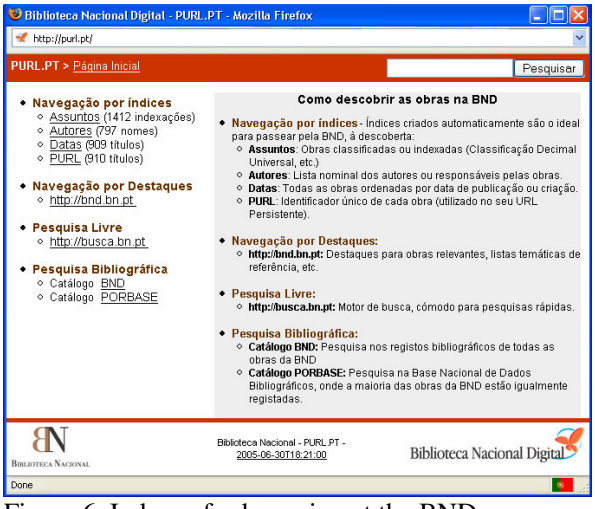

Figure 6. Indexes for browsing at the BND.

9 Access

The access service must provide the access to the information objects, interoperating with the search and browsing services. This includes a user workspace, where registered users can maintain private virtual collections, made out of their own indexes of references for the objects they've visited and marked.

Based on the profile of the user, the access service can provide also privileged access to special copies for some of the objects. This is made possible by the terms and conditions registered in the METSRights structure of each object and of each of its components. An example of an advanced service is, for example, the opportunity to order a collection on a CD-ROM or DVD, choosing probably from the different available versions (such as a higher resolution for the images, in JPEG or even in TIFF format).

The service that allows the users to order a virtual collection on a CD-ROM or DVD also uses the information represented in the METS structural metadata. The structural map of the object can be presented to the user has a tree, making it possible to choose only the relevant parts of the object, if desired. Additionally, all the access services use the information registered at the METSRights structure to check whether the objects, or some of their parts, are available or not for access or reproduction.

10 Conclusions

The development of the BND presents a unique class of challenges. This paper reports work in progress according with a pragmatic but efficient and flexible strategy that has been proved to be feasible. The results reached so far, comprising working versions of all the components and services, give us the necessary conviction to expect a successful project.

Several of the components, such as ContentE, MITRA and DEPTAL, are already available (and widely used in Portugal) for non-commercial purposes to who will be interested in reusing them.

References:

[1] Amorim, H., Borbinha, J. Digital Binding for Multiple Manifestations of a same Work, in Workshop on Generalized Documents 2001 / ECDL2001 – Fifth European Conference on Research and Advanced Technology for Digital Libraries, (4-9 September 2001), Darmstadt, Germany.

[2] Amorim, H., Borbinha, J. Digital Binding of Multiple Manifestations of Collections of Literary Works, in Proceedings of the ELPUB2002 (6-8 November 2002), Karlovy Vary, Czech Republic. VWF, Berlin, 182-193.

[3] Biblioteca Nacional. DiTeD – Depósito de Teses e Dissertações Digitais. <http://dited.bn.pt>

[4] Borbinha, J. The Digital Library - Taking in Account Also the Traditional Library, in Proceedings of the ELPUB2002 (6-8 November 2002), Karlovy Vary, Czech Republic. VWF, Berlin, 70-80

[5] Borbinha, J., Machado, J. Digital Library Components: DEPTAL and MITRA. ECDL2005 - Ninety European Conference on Research and Advanced Technology for Digital Libraries, (18-23 September 2005), Vienna, Austria (to be published)

[6] Borbinha, J., Pedrosa, G., Penas, J., ContentE: Flexible Publication of Digitised Works with METS. ECDL2005 - Ninety European Conference on Research and Advanced Technology for Digital Libraries, (18-23 September 2005), Vienna, Austria (to be published)

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

[8] Han-fei, et al. ARCO – Moving digital library storage to grid computing, in ICEIS 2004, 6th International Conference on Enterprise Information Systems (14-17 April 2004), Porto, Portugal.

[9] Han-fei., Almeida, N., Trezentos, P., Villate, J., Amorim, A. A Distributed Data Storage Architecture for Event Processing by Using the Globus Grid Toolkit, in Lecture Notes in Computer Science, Volume 2658 (2003) 267 - 274

[10] IFLA. UNIMARC: An Introduction. <http://www.ifla.org/VI/3/p1996-1/unimarc.htm>

[11] Johnson, R. Institutional Repositories - Partnering with Faculty to Enhance Scholarly Communication. DLIB Magazine, Nov. 2002. <http://www.dlib.org/dlib/november02/johnson/11johnson.html>

[12] Library of Congress. EAD - Encoded Archival Description <http://www.loc.gov/ead/>

[13] Library of Congress. METS – Metadata Encoding and Transmission Standard. <http://www.loc.gov/standards/mets/>

[14] Library of Congress. Z39.50 Maintenance Agency <http://www.loc.gov/z3950/agency/>

[15] LOCKSS - Lots of Copies Keep Stuff Safe. <http://lockss.stanford.edu/>

[16] LUSTRE – Scalable Clustered Object Storage. <http://www.lustre.org/>

[17] OAI-PMH. Open Archives Initiative <http://www.openarchives.org/>

[18] OCLC. PURL - Persistent Uniform Resource Locator <http://purl.oclc.org/>

[19] Project Gutenberg. <http://www.gutenberg.org/>

[20] Project Lucene <http://lucene.apache.org/>