

DEPTAL: A Framework for Institutional Repositories

José Borbinha¹, Jorge Machado²

¹INESC-ID – Instituto de Engenharia de Sistemas e Computadores, Rua Alves Redol 9,
Apartado 13069, 1000-029 Lisboa, Portugal
jlb@ist.utl.pt

²Escola Superior de Tecnologia e Gestão de Portalegre, Lugar da Abadessa, Apartado 148,
7301-901 Portalegre, Portugal
jmachado@ext.bn.pt

Abstract. This paper describes DEPTAL, an open and flexible framework for institutional repositories reusing open-source technology. DEPTAL is a collection-centric system that manages collections of documents in multiple copies and types. It can manage also users and groups of users, it supports authority control (subjects, authors, etc.), and can interoperate with other systems by interfaces such as OAI-PMH, Z39.50, SRU, web services, etc. DEPTAL recognizes descriptive metadata such as UNIMARC and Dublin Core, and organizes the information objects as HTML sites, with descriptions in the METS structural schema, making it very easy to backup and export those objects. For searching, it interoperates with MITRA, a search engine based on LUCENE, which was extended with new features to index not only the full content but also to recognize the structured metadata.

Introduction

The National Library of Portugal (BN) is a deposit library. The legal deposit law in Portugal mandates the universities to deposit the printed thesis and dissertations at BN. BN received so far nearly 30.000 of these documents, from all the scientific areas, which are all registered in PORBASE, the National Union Bibliographic Database [1].

With the purpose to replace the circulation of the printed thesis, BN has been proposing to the universities the voluntary deposit of digital copies, using the service DiTeD [2], part of BND, the National Digital Library Initiative [3][17]. To support this service and its specific requirements, BN developed an open source digital library management system called DEPTAL [4], which comprises the main functions that we can find nowadays in other similar frameworks for institutional repositories. DEPTAL has been freely distributed to universities and other institutions in the country, with the purpose to create a national network of interoperable repositories of technical and scientific documentation, even if only the thesis and dissertations are the concern of BN. Therefore, DEPTAL was designed with specific requirements to interoperate with PORBASE and BND, as also with any other open system. Another important requirement was also to make it easily adapted to multiple languages.

DEPTAL was designed as a collection-centric system that manages collections of documents in multiple copies and types. It can manage also users and groups of users; it supports authority control of indexing subjects, authors, and organizations; and it can interoperate with other systems by interfaces such as OAI-PMH [5], Z39.50 or its simplest version SRU [6].

Bibliographic descriptions are also objects in DEPTAL, just like the documents. DEPTAL recognizes UNIMARC [7], the descriptive metadata format used by almost all the libraries in Portugal (but it can be easily extended to process any other bibliographic format), and Dublin Core.

Finally, all the information objects in DEPTAL are structured in METS [8], making it compatible with the overall framework of the BND.

DEPTAL Architecture

DEPTAL is a client/server architecture, all coded in JAVA (Figure 1).

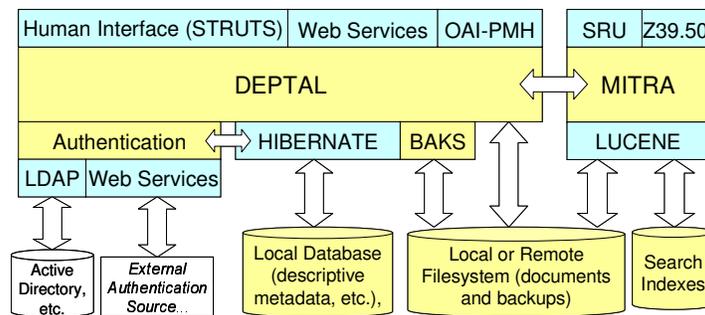


Fig. 1. DEPTAL architecture.

The system has "web based" interfaces for managers and users, developed using the Apache STRUTS Web Application Framework [9] and TOMCAT [18]. The preferable database management system (DBMS) is MYSQL [10], but DEPTAL can be installed with any other DBMS with ODBC interface and supported by HIBERNATE [11], an object/relational persistent and query service for JAVA.

Users can use the DEPTAL interface to browse or search in indexes, access the collections or documents that they have the right to access, as also to submit new documents. New documents submitted by external users are usually hold for approval for the administrators. Administrators can manage all the information of the system, such as to configure collections; manage users' accounts; deposit new documents; backup and export documents for archival (using the BAKS module, part of DEPTAL, or any other possible alternative); etc.

All the documents in DEPTAL are structured and stored as a web site, with a homepage linking to the contents, and including copies of all its metadata in XML files. All of this makes an object, described in METS. This makes it very flexible and easy to export and backup those objects, as also to support other management tasks.

DEPTAL is a multilingual framework, not only in the user and administration interfaces, but also in the handling of the contents. Content objects and their metadata can be stored in multiple languages. By omission, the system is configured for two languages, English and Portuguese.

Information Objects in DEPTAL

The information objects in DEPTAL are bibliographic descriptions and content documents. Bibliographic descriptions are metadata records describing documents with copies stored or not in a DEPTAL storage space. This means DEPTAL can be used as only a bibliographic database, with references to external digital and/or physical documents.

The content documents in DEPTAL can be partial or complete. Partial documents are explicitly identified to make it clear that more complete instances of that same document exist somewhere else (in the same DEPTAL repository or not). This is very relevant for collections of digitized copies of printed documents, such as thesis, books, papers, etc. In some cases, due to copyright restrictions, those documents can not be made freely available to the public, so we might have to have the complete copy available only to specific users or in a restricted domain (the local network, for example), and have only a partial copy freely available. Also, to save costs, it can be decided to digitize only part of the original documents, building an information object useful for search and dissemination, but which is not the complete document!

As a result of these requirements, each document in DEPTAL can have therefore more than one copy. Also, each copy can have more than one file, of more than one MIME format (MS-Word, PDF, Postscript, ASCII, RTF, HTML, etc.).

Objects Management

DEPTAL is a collection-centric system. Any information object in DEPTAL belongs to one or more collections. Collections can be organised in trees, which can be created for schools, departments, laboratories, sections, areas, research groups, or even for individuals. Collections are associated to their own users, or groups of users, and administrators. They can be open for submission by anonymous users (useful for systems like DiTeD), or can be closed for submission only by registered users.

A typical workflow to create new information objects in a DEPTAL system starts with the submission of a bibliographic description and the upload of the files of the content objects. In this step the user can reuse information from controlled authority files, namely the author's names and indexing terms (DEPTAL can manage multiple indexing languages).

Objects submitted by users that are not administrators are usually hold for approval. The administrator of the collection can exchange emails with the user (which might be typically the author of the object). If in the end the submission is approved, all these messages are also stored as part of the information object, as administrative metadata. To make it easier to manage specific collections, it is also

possible to give to a user or to a group of users, the privilege to have a document immediately approved in that collection (the case for individual collections).

The requirement to support anonymous submissions is very important for the DiTeD service. As a national service, it can accept deposits from users from everywhere, which is very convenient. Even if it requires a further more careful approval process, since it is necessary to confirm the identity of the user, its formal relationship with the document, etc., it is a very convenient function.

It is also possible to create an information object importing a bibliographic record coded, for example, in MARCXML [12], and with UNIMARC contents (UNIMARC is the bibliographic format most used in Portugal). In this case the information is parsed to copy the necessary fields to the DEPTAL internal metadata structure, but the original record is also kept in the system as an XML file. This is a very important feature, since it makes it possible to synchronize DiTeD with PORBASE, having more than 30.000 bibliographic descriptions of thesis in both the systems. BN runs a DEPTAL server, to manage its own collection of digital thesis and dissertation, which can import information objects from the other DEPTAL systems in the country. This central server interoperates also with PORBASE, exchanging UNIMARC records.

Structural Metadata

An information object can range from a simple MARXML file with a UNIMARC record, to a very complex object comprising several copies of a document, rights and administrative metadata, etc. To aggregate all of this, DEPTAL uses METS as the schema for structural metadata.

Each information object is stored has a web site, which includes also all the administrative metadata and the METS file. This makes it very simple to copy objects from DEPTAL without the need of any transformation, while the access to the METS file makes it possible to implement more complex services.

These objects, represented as HTML sites and described in METS structures, are also perfect structures for preservation, making it straight the submission to the deposit and preservation service at the BND.

Searching and Interoperability

DEPTAL provides its own searching interface (Figure 2), but we provide also an extra search engine, MITRA [13], based on LUCENE [14] (Figure 3). Other search and interoperability interfaces are provided by Z39.50, SRU and OAI-PMH services.

For searching, it interoperates with MITRA, a search engine that extends LUCENE with new features to recognize also structured metadata. MITRA can be configured to index any kind of XML schema, as also as it can index any usual web site. This provides important added value, since it can index the content of the objects, as also other metadata structures that might be part of them, such as Dublin Core records, bringing together the best from the two worlds.

The screenshot shows the DiTeD search interface. The top navigation bar includes links for 'Estadísticas', 'Sobre o DiTeD', 'Como Participar na Iniciativa DiTeD', 'Contactos', 'Links Relacionados', 'Instituições Participantes', and 'Perguntas Frequentes'. The search bar is set to 'Português' and shows 'Pesquisa Avançada' with 660 results. The results are displayed in a table with the following columns: PURL NCB DITED, Título do Documento, Data, Género, Autores, and Instituição(/Instituições) Editora(s).

PURL NCB DITED	Título do Documento	Data	Género	Autores	Instituição(/Instituições) Editora(s)
1086522 29487	Touros de morte em Barrancos **	2001	Tese de Mestrado	Tuleski, Valéria Eneida Ruviano	ISCTE-Instituto Superior de Ciências do Trabalho e da Empresa
1086241 29491	A habitação no grande Porto **	2001	Tese de Doutoramento	Matos, Fátima Loureiro de	UP-Universidade do Porto
1086174 29497	Paredes exteriores de edifícios em pano simples **	2001	Tese de Mestrado	Alves, Sérgio Gabriel Quádrão da Mota	UP-Universidade do Porto
1085753 29488	L'immaginario urbano di Lisbona	2001	Tese Final de Licenciatura	Cioriani, Laura	Istituto Universitario di Architettura di Venezia
1086048 29492	A fortaleza de Mazação **	2001	Tese de Mestrado	Matos, João Manuel Barros	Universidade de Évora
1086193 29493	Planeamento colaborativo em contexto de regeneração urbana **	2001	Tese de Mestrado	Alves, Sónia Cristina Nunes, 1973:	UP-Universidade do Porto

Fig. 2. An example of an installation of DEPTAL for the DiTeD system at BN.

The screenshot shows the MITRA search interface. The search bar contains the word 'morte' and the search button is labeled 'pesquisar'. The results section shows 'Resultados: 0 - 10 de cerca de 3 para morte. (0.19 segundos)'. The first result is 'Touros de morte em Barrancos-uma tradição da fronteira luso-espanhola ...' by Tuleski, Valéria Eneida Ruviano. The second result is 'A metáfora da morte na escultura contemporânea em Portugal, na 2 metade do séc. XX ...' by Silva, António Fernando Monteiro Pereira da. The third result is 'Study of the Resistance patterns of Uropatogenic E. coli isolated from animals to beta-lactams and beta-lactamase inhibitors ...' by Ramos, Christian Gomes da Silva.

Fig. 3. An example of a search result in DiTeD, a DEPTAL system, using MITRA.

Conclusions

DEPTAL is being used at BN for more than one year, as also at more four places in the country. At BN it is used to host several collections. The first public stable version will be released in middle 2005, as open-source.

So far DEPTAL has proven to be highly flexible, due to its very simple and modular architecture. Future plans include extending it with more interoperable features, such as interfaces for external authentication services others than LDAP [15], import and export of more descriptive metadata schemas, processing of SCORM objects [16], extend the BAKS module for more powerful version control, etc. In the scope of BND, DEPTAL will be used in new services to support multiple scenarios for voluntary deposit, a long awaited target of BN [18].

References

1. PORBASE. Base Nacional de Dados Bibliográficos. <<http://www.porbase.org>>
2. DiTeD. Digital Thesis and Dissertations. <<http://dited.bn.pt>>
3. BND. Biblioteca Nacional Digital. <<http://bnd.bn.pt>>
4. DEPTAL. Repositório Institucional. <<http://deptal.bn.pt>>
5. OAI-PMH. Open Archives Initiative - Protocol for Metadata Harvesting <<http://www.openarchives.org/>>
6. Z39.50. Z39.50 Maintenance Agency. <<http://www.loc.gov/z3950/agency/>>
7. UNIMARC <<http://www.unimarc.info>>
8. METS. Metadata Encoding and Transmission Standard. <<http://www.loc.gov/standards/mets>>
9. STRUTS. The Apache STRUTS Web Application Framework. <<http://struts.apache.org/>>
10. MYSQL <<http://www.mysql.com/>>
11. HIBERNATE. <<http://www.hibernate.org/>>
12. MARCXML. MARC 21 XML Schema. <<http://www.loc.gov/standards/marcxml/>>
13. MITRA. <<http://mitra.bn.pt>>
14. LUCENE. Apache LUCENE. <<http://lucene.apache.org/>>
15. LDAP. Lightweight Directory Access Protocol. <<http://www.ietf.org/rfc/rfc2251.txt>>
16. SCORM. The Sharable Content Object Reference Model. <<http://www.adlnet.org/index.cfm?fuseaction=scormabt>>
17. BND: the architecture of a national digital library. José Borbinha, Nuno Freire, João Neves. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, JCDL 2004, Tucson, AZ, USA, June 7-11, 2004 (DOI: 10.1145/996356)
18. TOMCAT. Apache Jakarta TOMCAT. <<http://jakarta.apache.org/tomcat/>>
19. NEDLIB. NEDLIB - LB 5648: D1.2.2 High Level Design. José Borbinha. 1998. <<http://www.kb.nl/coop/nedlib/high-level-design/d122.pdf>>