

A model of multilingual digital library

Ana M. B. Pavani

LAMBDA – Laboratório de Automação de Museus, Bibliotecas Digitais e Arquivos. Departamento de Engenharia Elétrica
Pontifícia Universidade Católica do Rio de Janeiro
apavani@lambda.ele.puc-rio.br

Resumo

Este trabalho aborda o problema de bibliotecas digitais multilíngües. A motivação para tal biblioteca digital decorre da diversidade de línguas dos usuários da Internet, bem como da diversidade dos autores do conteúdo, de autores de livros eletrônicos para elaboradores de cursos. São discutidas as definições básicas de tal sistema, as especificações de sua funcionalidade e a identificação dos itens que ele comporta. Apresenta-se o impacto do multilingüismo em cada um dos aspectos anteriores. Um estudo de caso de uma biblioteca digital multilíngüe – no Sistema Maxwell, na PUC-Rio – é descrito nas últimas seções. Suas principais características são descritas e é mostrado o status atual de sua biblioteca digital.

Palavras-chave

Biblioteca digital multilíngüe; Metadados; Bases de dados multilíngües.

A model of multilingual digital library

Abstract

This paper addresses the problem of multilingual digital libraries. The motivation for a such a digital library comes from the diversity of languages of the Internet users as well as the diversity of content authors, from e-book authors to writers of courseware. The basic definitions of such a system, the specifications of its functionality and the identification of the items it holds are discussed. The impact of multilinguism in each of the former aspects is presented. A case study of a multilingual digital library - in the Maxwell System in PUC-Rio - is described in the last sections. Its main characteristics are described and the current status of its digital library is shown.

Keywords

Multilingual digital library; Metadata; Multilingual database.

INTRODUCTION

Digital libraries are available on the Internet. This means that users from all over the world can reach digital libraries and try to use their contents. There is no doubt that English is the most popular language on the Internet but other languages must be considered when digital libraries are the focus. Remember that digital libraries, and more generally digital collections, are repositories of culture, education^{1,2}, research results^{3,4,5,6}, science, law, etc.

The 100 most spoken languages of the world, when first language speakers are counted, can be found in⁷. In descending order, the first 10 are Chinese (Mandarin), Spanish, English, Bengali, Hindi, Portuguese, Russian, Japanese, German (Standard) and Chinese (Wu). The various peoples and nations of the world feel their languages are symbols of their identities and struggle to preserve them.

When digital libraries are viewed as support to distance education⁸, it becomes clear that different languages have to be considered because they are important components of the educational process of each country.

In the context of the work that yielded this article, a multilingual digital library is defined as:

Definition 01: Multilingual digital library

A multilingual digital library is a digital library that has all functions implemented simultaneously in as many languages as desired and whose search & retrieve functions are language independent.

The implementation of such a concept allows the user to choose the language for the interfaces and to access cataloging information in any language (of the digital library) from the interface in the chosen language regardless of the language(s) of the library items.

The following sections present the definitions of all the important attributes of a multilingual digital library based on Definition 01, the functions of such a library, the data model of the database and the implementation in the Maxwell System⁹.

BASIC DEFINITIONS

The definitions presented in this section cover the items of the library and the language characteristics.

Definition 02: Content or title

A content or title is the logical definition of an item of the digital library. It is identified through a set of attributes that is presented in section 4 – Metadata Scheme.

Definition 03: Occurrence or instance

An occurrence or instance is the physical realization of a content or title. It is a digital object and has its identification through a set of attributes that is presented in section 4.

In a traditional library, we refer to titles and volumes, a library collection may hold many volumes of a title, as for example text books. In a digital library the separation of titles (contents) and volumes (occurrences, instances or objects) is necessary for three reasons. The first is the fact that a content may have instances in different digital formats. An example is a text that may be used as hypertext (html) and as a linear text (pdf); the former for online visualization (with animations, interaction, etc.) and the latter for printing¹⁰. The second comes from an eventual requirement of splitting a content in more than a file due to bandwidth characteristics. The third may come from rights management. It occurs when a digital library is entitled to operate with the number of ‘copies’ of a title that corresponds to the copyright paid¹¹.

Definition 04: Language of the content

The language of the content is the language in which the content is written and/or spoken.

Definition 05: Language of the catalog entry

The language of the catalog is the language in which the attributes of the content and its instances are written.

Definition 06: Language of navigation

The language of navigation is the language of all the interfaces and messages of the digital library system.

Definitions 04 - 06 yield the existence of 3 language control parameters all over the system - from database to programs with a special emphasis on the catalog (metadata). The language control parameters are described in subsection 5.2 - Language Control Parameters.

FUNCTIONALITY SPECIFICATIONS

In the case to be presented, the multilingual aspects of the digital library were implemented following a set of functionality specifications that were intended to make the system: comfortable to the user, flexible in terms of number and languages to support, wide and accurate in terms of the information to store and serve, and easy to operate in terms of feeding information.

Specification 01: Comfortable to the user

Digital libraries are created in countries and, in general, may be somehow specialized. It is expected that the users in the country where the library operates will want to use it in their native language. At the same time, users with other native languages than that of the country under consideration may need more international languages, as for example, English or Spanish.

In order to implement any set of languages (both in number and in nature) is necessary that:

- All program interfaces can appear in any of the chosen languages;
- All messages are shown in any of the chosen languages;
- All elements of the domain tables are shown in any of the chosen languages.

Specification 02: Flexible in terms of the chosen languages

Digital libraries in different countries and aiming at different sets of users may operate with distinct sets of languages. As an example, in Brazil a good set of languages is Portuguese, Spanish and English. In Italy or Taiwan the sets of languages would be different.

For the same system to be used in many countries, the number of languages must be any as well as the set of languages.

Another feature in this area is that all the languages do not have to be included at the same time. A fourth language, as for example French, could prove necessary and be added later on to the Brazilian set.

Specification 03: Wide and accurate in terms of the information

In order to fulfill its functions, the digital library must be accurate in terms of information. This requires that the languages be kept track at any moment of operation by the user; he/she must know the language(s) of the catalog entry and of the content, regardless the navigation language in use.

For the search to be effective, no matter the navigation language, the user may submit search arguments in any language and all points of access must be possible to be searched in all languages, regardless of the navigation language of the session.

In order to identify contents in original languages and corresponding translations into other languages, a strict translation control must exist.

Specification 04: Easy to operate

The digital librarian must control all cataloging (original languages and translations) and all the contents (original languages and translations). For data integrity to be achieved, translation interfaces for the cataloging and translation control applications must be available.

In order to make sure that the cataloging be kept under control, the translator must be able to translate but not to include items that have not been technically processed by the librarian.

The specifications yielded characteristics to both the database model and to the programs structures since both are language dependent. In terms of the programs and of the database, it does not make any difference which language is used first, since only the navigation language will determine how the interfaces are built in terms of texts.

The only control of original and translation is related to contents, as it is in traditional libraries. There is no restriction on the languages of contents, that is, a content can be in a language that is not a navigation language as it happens in a traditional library.

METADATA SCHEME

Digital libraries hold items that are of the same nature of traditional libraries as for example digitized books, maps, manuscripts, etc. Besides that, digital libraries are natural bases to distance learning as traditional libraries have been to face-to-face learning. In this model the digital library contains all the learning objects (LO's), besides links to other libraries, references, etc. To implement this model, all the LO's must be treated as library items before they are stored and, in a second step, related to courses.

In 1995, in the city of Dublin, a workshop was held to discuss the cataloging of digital library items. "The result of the first workshop – the Dublin Core Meta-data Element Set – represents a simple resource description record that has the potential to provide a foundation for electronic bibliographic description that may improve structured access to information on the Internet and promote interoperability among disparate description models. It's major significance, however, lies not so much in the precise character of the elements themselves, but rather in the consensus that was achieved across the many disciplines represented at the workshop."¹²

After this workshop, others were held and the Dublin Core Meta-data Element Set: Resource Page⁶ was published in 1997. Currently it contains 15 attributes that are considered the minimum set to describe an item of a digital library. Other metadata schemes are used, as for example the metadata specification of the Library of Congress that has almost 80 pieces from which, approximately, 20 are required¹³. The LC metadata specification, obviously, contains the 15 elements of DC set.

Information technology has played an important role in education for the last 10 years specially when distance learning is considered. The Internet has become an important tool to distance learning; WBE is the designation of Web Based Education and it is well known in the distance learning community. Since the digital contents used for education (LO – learning objects) are items of a collection, they must be identified to be made available.

A lot of discussion has been devoted to the definition of the metadata for the LO's. The most important forums for this discussion are the IEEE Learning Technology Standards Committee¹⁴ and the IMS Global Learning Consortium, Inc.¹⁵. Both organizations are seriously committed to the creation of standards for distance learning with the support of information technology and the identification of the LO's is one of the concerns.

The metadata scheme is one of the characteristics that the digital library & distance learning system has. Other characteristics are grouped into two main categories – the functions and the technological environment. Once the compliance to one of the metadata schemes is chosen, the other cataloging attributes must be defined from the specifications of the system functionality. This is necessary to assure the system operability since many information pieces are necessary to satisfy specifications of the local environment. The main points are presented in¹⁶.

It is interesting to remark that this is not different from the set of characteristics that are present in a e-business or e-commerce system too^{17, 18, 19, 20}. Digital libraries, WBE and e-commerce/e-business deal with items that must be identified, with sets of systems functions and technology standards. This makes them analogous in terms of problems and solutions.

THE MAXWELL SYSTEM DIGITAL LIBRARY – AN IMPLEMENTATION

A multilingual digital library is implemented and operating in the Maxwell System of the Department of Electrical Engineering of PUC-Rio. This system is the implementation of a digital library, a distance learning environment and all the administrative infrastructure to support both. In the case of this system, not only the digital library is multilingual, the whole system is multilingual. It is also multi-institution and multi-system (in each institution).

Although the Maxwell System can have as many languages as desired, the choice was to use Portuguese (the language of Brazil), Spanish (the language of our Latin American neighbors) and English (the most widely used language on the Internet). This set contains the 3 western languages with the largest number of native speakers⁷.

Figure 01 in the Appendix shows the opening page of the school/library environment of the system and figure 02 shows the corresponding page in the administrative environment. Both present the buttons for the user to choose the language to use the system.

Metadata

The implemented metadata scheme considered the specifications of both communities (WBE and libraries/digital libraries) since the Maxwell System Digital Library holds LO's besides articles, manuals, ETD's, administrative documents, etc. Many LO's are not electronic versions of originally paper supported contents; they are born digital objects - animations, simulators, interactive exercises, etc. that were developed for educational purposes and stored in the system digital library.

Because of the LO's, other attributes were added to the metadata set in the system – the ones of interest to the operation of the system and to the information that the university requires. For example, the ETD functionality of the digital library added new attributes to the cataloging such as examining committee, date of presentation, date of acceptance, funding agency, etc. The reasons and characteristics of the additional metadata are explained in¹⁶.

It is important to remark that there is flexibility in the cataloging. Some attributes are specific of some types of library items. Nevertheless, there is a minimum set that all items must comply with, both in terms of contents and of instances.

Language Control Parameters

In order to achieve the functionality specifications of section 3, three language parameters had to be specified according to definitions 04 - 06.

Language parameter 01: Language of the content

This parameter comes from definition 04. It is one of the metadata in all metadata schemes.

It is also used in traditional libraries since it is necessary to inform the library user when the record is retrieved to make sure he/she can understand its contents.

Language parameter 02: Language of the catalog entry

This parameter comes from definition 05.

In the Maxwell System Digital Library, contents and instances are cataloged in 3 languages (Portuguese, Spanish and English). Therefore, all information on contents and instances that is language dependent exists in the 3 languages and has a language code associated to it.

Language parameter 03: Language of navigation

This parameter comes from definition 06.

It is dependent of the choice of the user when he/she enters the system - there are buttons for each language on the opening page of both environments of the system, as shown in figures 01 and 02. It can be modified if the user returns to the opening page and chooses other language.

The programs and the database in all parts of the system use the 3 language control parameters.

Controlled Set of Keywords

The Dublin Core Metadata Specification requires 'subject' as one of the cataloging attributes. The contents of this field can come from traditional library sources as subject headings, like the Library of Congress Subject Headings that is used all over the world in many languages. They can also come from a controlled set of keywords. This was the choice in this implementation.

A multilingual controlled set of keywords was created. Each entry is identified by a number and a language code. English is mandatory with any other language. The number is used in the record and related to the vocabulary in the different languages by the use of a relational database.

All keyword searches are performed in the vocabulary and, once the number(s) is(are) identified, the records are searched for it(them).

In the process of inputting elements to the controlled set, all words are transformed to uppercase characters and accents (for example in Portuguese and Spanish) are eliminated. The transformation is performed by the program and requires no actions from the cataloger.

The same operation of changing to uppercase and eliminating the accents is performed in the search procedure, as it will be described in subsection 5.7 - Search, Retrieve and View in a Multilingual System.

Language Neutral Database

The database of the Maxwell System holds all the information that allow the management of contents, instances, users, courses, access control, etc. It is implemented in IBM UDB DB2 for AIX and, currently, has a little over 100 tables.

FIGURE 1
Opening page School/Library environment

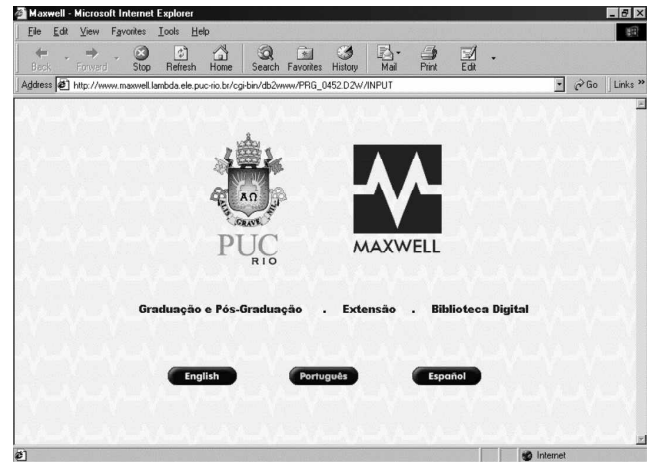
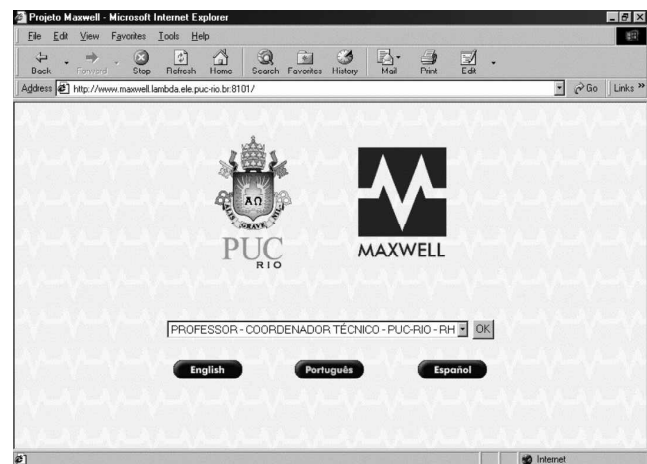


FIGURE 2
Opening page Administrative environment



In many, if not most, of these tables there is data that is written in a language, as for example types of users, functions of the users, titles of the contents, abstracts ('description' filed of the Dublin Core Metadata Specification). Other examples of information that are language dependent are keyword set, names of courses, syllabus of courses, etc. All the messages the system uses to communicate with users are, obviously, language dependent. Data that is not language dependent are the names of authors, names of students, electronic formats of objects, dates, etc.

Since the system is multilingual, the functionality of using the data in the right language, that is the language of the interface in use, must be implemented. For this reason, the database was designed to be neutral in terms language. The definition was that tables that have data that is language dependent must have the language code included in the primary key. This is used in all the domain tables. An example is the definition of types of users:

(APG, pt) = Aluno de pós-graduação
(APG, es) = Alumno de posgrado
(APG, en) = Graduate student

Obviously, the 3 lines have the same meaning.

Once the first row has been entered the database, the others can be created from the programs to translate information. Almost all functions of the Administrative, Librarian and Technical environments have functions for: Query, Input, Update, Delete and Translate.

Programs

A second characteristic of the multilingual digital library is the choice of the interface language. For this to be possible, the programs must be developed in a manner that all language dependent information be grouped in an area divided into sections, each one associated to one language. When the language control parameter specifying the navigation language is identified, the proper language section is displayed.

The messages that are displayed (after some processing is performed) are stored in a database table whose primary key is a composite key containing the message number and the language code. In this table there are more than 450 messages. All of them are in the 3 languages. An example of message in this table is:

(001, pt) = Identificação ou senha inválida.
(001, es) = Identificación o contraseña inválida.
(001, en) = Identification or password not valid.

The program contains the information that message 001 must be displayed. After the identification of the language control parameter of navigation language, an access is performed in the table, with the message number and the language code, so that the suitable sentence is chosen.

This is a convenient modeling since most messages are shared among many programs of the system.

Multilingual Catalog

The metadata of all contents and instances are captured through programs that store them in the database.

Titles, subtitles, abstracts, etc are language dependent while authors' names, cataloging dates, etc. are not. The fields that are language dependent must be associated with language codes too.

In the case of library items metadata, the cataloging is translated into the other 2 languages besides the original language of the cataloging. All language dependent data is translated. This does not imply translation of the content. A content in Portuguese, for example, may be cataloged in the 3 languages. The multilingual cataloging is available so that search mechanisms can find the content; if the content language is not suitable to the user, a translation can be requested.

Titles have a double representation in the database in all the languages. One is for displaying purposes - it has all the accents of the language it is written in and is written with upper and lowercase characters; the other has no accents and is written in uppercase characters for searching purposes. The search programs perform the same changes.

Since language codes are part of the primary keys, any number of languages and any language can be used. This makes the system suitable to use in countries other than Brazil.

When a content is translated into another language, this new item is cataloged independently (a new content identification number is generated) because the cataloging attributes are different, for example a collaborator/ translator must be included, dates are different, etc. There is a database table controlling the relations between original language items and translated items.

Search, Retrieve and View in a Multilingual System

The terms search and retrieve are well known and, since its first version in 1988, Information Retrieval Service Definition and Protocol Specifications for Library Applications (Z39.50) [21] has specified minimum requirements for these functions in terms of the end user.

The Maxwell System was implemented with the functionality mentioned above in terms of search and retrieve; it also added the view function. A brief description of the 3 functions follows.

• Search

This part of the Maxwell System is still in its initial stage. As a first step, 3 search possibilities were implemented: by author, by title and by keywords. There two new ideas in the implementation of these search possibilities.

The first new idea applies to the 2 functions that are language dependent (title and keywords) in which the search is performed in all the languages of the system, no matter what is the interface language in use. The purpose of this characteristic is to fulfill Specification 01 (section 3).

Regardless of the type of search to be performed, the other characteristics are the usual ones (Z39.50): the user defines the type of search, writes the search argument and submits the search; the return from the system is a page with the type of search, the search argument, the number of found items and data that makes it possible to minimally identify an item.

The second new idea is related to this identification information. It displays the main author's name, the title, the database identification number and the language code of the catalog entry in all the languages there are catalog entries. Thus, when a content is cataloged in the 3 languages, there will be one line of identification data for each one. Then, the user chooses the language to retrieve the catalog 'card'. This allows the user to be informed of the content even if its language is not the one he/she is using to navigate. If there is real interest, which can be found out from the catalog information, a translation can be requested.

A page with the return information after a search is shown in figure 03 in the appendix. The interface language is English.

• Retrieve

Once the user receives a page with the return information from the search, he/she can retrieve the catalog 'card' in the chosen language.

Two pages with the 'cards' are shown in figures 04 and 05 in the appendix. Figure 04 shows a Spanish interface and an English catalog information for a Portuguese written content. Figure 05 shows the corresponding catalog information from an English interface.

FIGURE 3
Result of search of contents by keywords

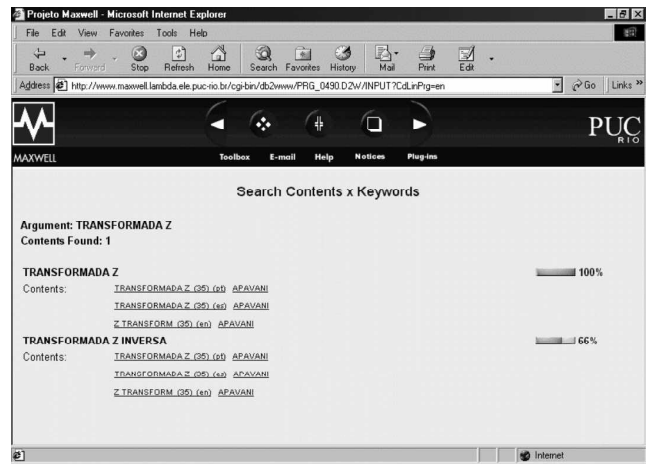


FIGURE 4
Retrieved catalog entry – Spanish interface, English cataloging and Portuguese content



FIGURE 5
Retrieved catalog entry – English interface, English cataloging and Portuguese content



• View

This function is executed after the user finds out the instances of a content. Figure 06 shows a page where the instances of the content are listed. To view one of them, the user clicks on the corresponding link. Figure 07 show the result of this operation, the pdf instance was chosen.

The first 2 functions follow the Z39.50 specification but add the multilingual characteristic and the third introduces, in the sequence operation sequence, the access to the electronic realization of the content.

Current Status of the Digital Library

The Maxwell System Digital Library contains many different types of digital objects. They differ in the nature of their contents from administrative information about registration to interactive exercises to teach Portuguese. In terms of electronic format there are .html, .doc, .ppt, .pdf, .exe, .xls, .swf, etc. objects.

The overall information about the library is shown in table 1.

ACKNOWLEDGEMENT

This work was partially financed by Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) and IBM Brasil.

FIGURE 6
List of instances of retrieved content

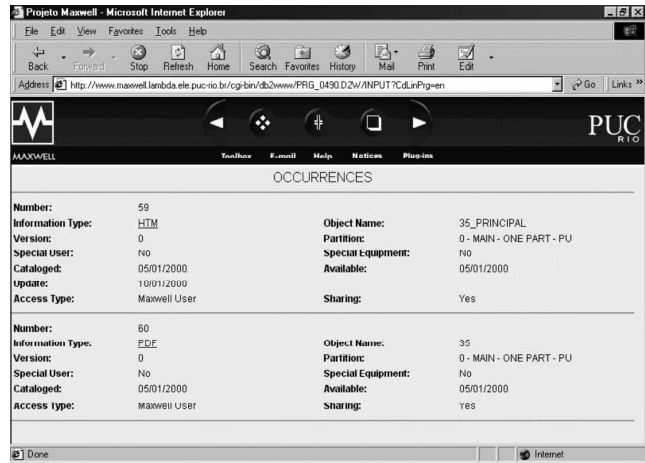


FIGURE 7
View of .pdf instance of retrieved content

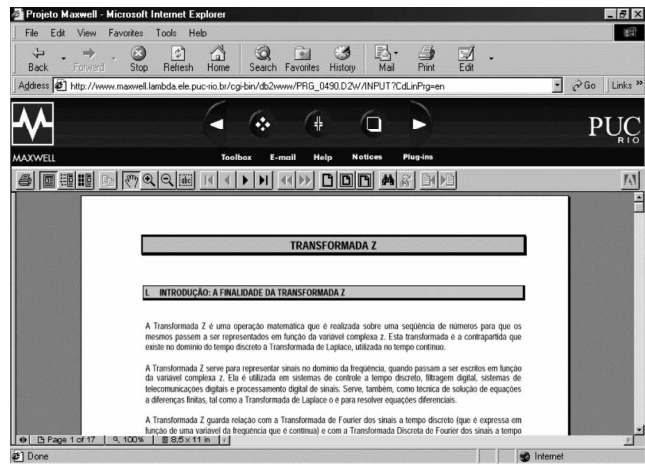


TABLE 1
Maxwell Digital Library - General Information on Nov 19, 2001

Type	Number
Contents	2102
Authors	633
First authors	429
Contents by author	4.9
Contents with collaborators	373
Instances	3894
Contents with instances	2073
Average instances/content	1.88
Contents under development	29
Authors' contents	1219
Teaching contents	817
Administrative contents	40
Technical contents	22
Internal reports	4

REFERENCES

1. PAVANI, A. M. B.; LUKOWIECKI, A. L. S. Digital libraries and sharing of course contents: the Maxwell project. In: INTERNATIONAL CONFERENCE ON COMPUTER AND ENGINEERING, 1999. **Proceedings...** [S. l. : s. n., 1999?].
2. FOX, E. A. *et al.* Networked digital library of these and dissertations. *D-Lib Magazine*, Sept. 1997
3. SULEMAN, H. *et al.* Networked digital library of theses and dissertations: bridging the gaps for global access. *D-Lib Magazine*, v. 7, n. 9, 2001. Pt 1: Mission and progress.
4. UNESCO ETD GUIDE - 2001. Disponível em: < <http://etdguide.org/> >
5. SULEMAN, H. Atkins *et al.* Networked digital library of theses and dissertations: bridging the gaps for global access. *D-Lib Magazine*, v. 7, n. 9, 2001. Pt 2: Services and research
6. Top 100 Languages by Population. Disponível em: < <http://www.sil.org/ethnologue/top100.html> >. Acesso em: jan. 2001.
7. PAVANI, A. M. B. Digital libraries and distance education. In: REENEN, Johann van (Ed.). *Digital libraries and virtual workplaces: important initiatives in Latin America in the information age.* [S. l.] : Organization of American States, 2001. Cap. 7.
8. MAXWELL SYSTEM. Rio de Janeiro : Pontifícia Universidade Católica, Departamento de Engenharia Elétrica, [2001?]. <Disponível em: <<http://www.maxwell.lambda.ele.puc-rio.br/>> .
9. PAVANI, A. M. B.; LUKOWIECKI, A. L. S. Engineering distance learning: development of course contents. In: INTERNATIONAL CONFERENCE ON ENGINEERING EDUCATION, 2000, Taiwan. **Proceedings...** [S. l. : s. n. 2000?].
10. BORGES, K. S. *Bibliotecas digitais: um sistema de controle de empréstimos e devoluções de objetos digitais.* 2000. Dissertação (Mestrado) - Pontifícia Universidade Católica do Rio Grande do Sul, Faculdade de Informática. [Porto Alegre], 2000.
11. DUBLIN CORE META-DATA ELEMENT SET: resource page, OCLC. Disponível em: <http://purl.org/metadat/dublin_core/>. Acesso em: maio, 2000.
12. LIBRARY OF CONGRESS DIGITAL REPOSITORY DEVELOPMENT: core meta-data elements. Disponível em: <<http://lcweb.loc.gov/standards/metadata.html/> >. Acesso em: maio, 2000.
13. DRAFT STANDARDS FOR LEARNING OBJECTS METADATA. [S. l.] : IEEE Learning Technology Standards Committee, 2000. Disponível em: < <http://ltsc.ieee.org/doc/wg12/scheme.html/> >. Acesso em: fev. 2000.
14. IMS LEARNING RESOURCE META-DATA INFORMATION MODEL, version 1.0. 2000. [S. l.] : IMS Global Learning Consortium, 2000. Disponível em: < <http://www.imspjct.org/metadata.mdinfo01.html/> >. Acesso em: abr. 2000.
15. Pavani, A. M. B. Some considerations on the cataloging of learning objects in a digital library. In: NTERNATIONAL CONFERENCE ON ENGINEERING AND COMPUTER EDUCATION, 2000, **Proceedings...** [S. l. : s. n. 2000?].
16. EAN INT . *European article number.* Disponível em: <<http://www.ean.be/> > .
17. EAN UK. *European article number.* Disponível em: <<http://www.e-centre.org.uk/> > .
18. UCC. *Uniform Code Council.* Disponível em: <<http://www.ucc-council.org/> >
19. UDDI. *Universal description, discovery and integration of business for the web.* Disponível em: <<http://www.uddi.org/>>.
20. INFORMATION retrieval service definition and protocol specifications for library applications - ANSI/NISO Z39.50. Disponível em: <<http://lcweb.loc.gov/z3950/agency/> > .