Libri, 2003, vol. 53, pp. 254–261
Printed in Germany · All rights reserved

Libri
ISSN 0024-2667

Long-Term Preservation of Electronic Theses and Dissertations in Algeria

YAHIA BAKELLI AND SABRINA BENRAHMOUN
CERIST Research Centre on Scientific and Technical Information, Algiers, Algeria

In accordance with a decree issued by the Ministry of Higher Education and Scientific Research in Algeria in August 2000, an electronic copy of every Master's and PhD thesis defended in all academic institutions must be deposited at the CERIST Research Centre. Deposit is a condition for getting the diploma. CERIST is then entrusted with the mission to build a database of Algerian theses and to update the national inventory of current theses and research. However a serious problem of archiving and preserving these Electronic Theses and Dissertations (ETDs) has emerged. From December 2001 to November 2002 a great number of ETDs has been deposited and constitutes a set of more than 1000 floppy disks and 100 CD-ROMs. What guarantees that these digital materials deposited by students are preserved and safeguarded? What guarantees that the content of these materials are preserved and accessible at any time regardless of machine, operating system and software. This paper explores the problem of the long-term conservation and preservation of electronic theses in the Algerian context, and shows how international recognised standards and techniques for setting up and organising the local ETD's archives may be applied.

Introduction

The CERIST Research Centre on Scientific and Technical Information of Algiers was created in 1985, with the main mission of designing and implementing the National Information System. Within this framework, great importance is given to recording and making available academic literature through the production of national bibliographic databases and national union catalogues. Examples of these bibliographic products are the Algerian Scientific Abstracts, Algeriana, CAT (Algerian Theses Catalogue) and FNT (National Theses Repository), all of which can be searched through the Academic Research Network and CERIST Web sites (see Figure1).

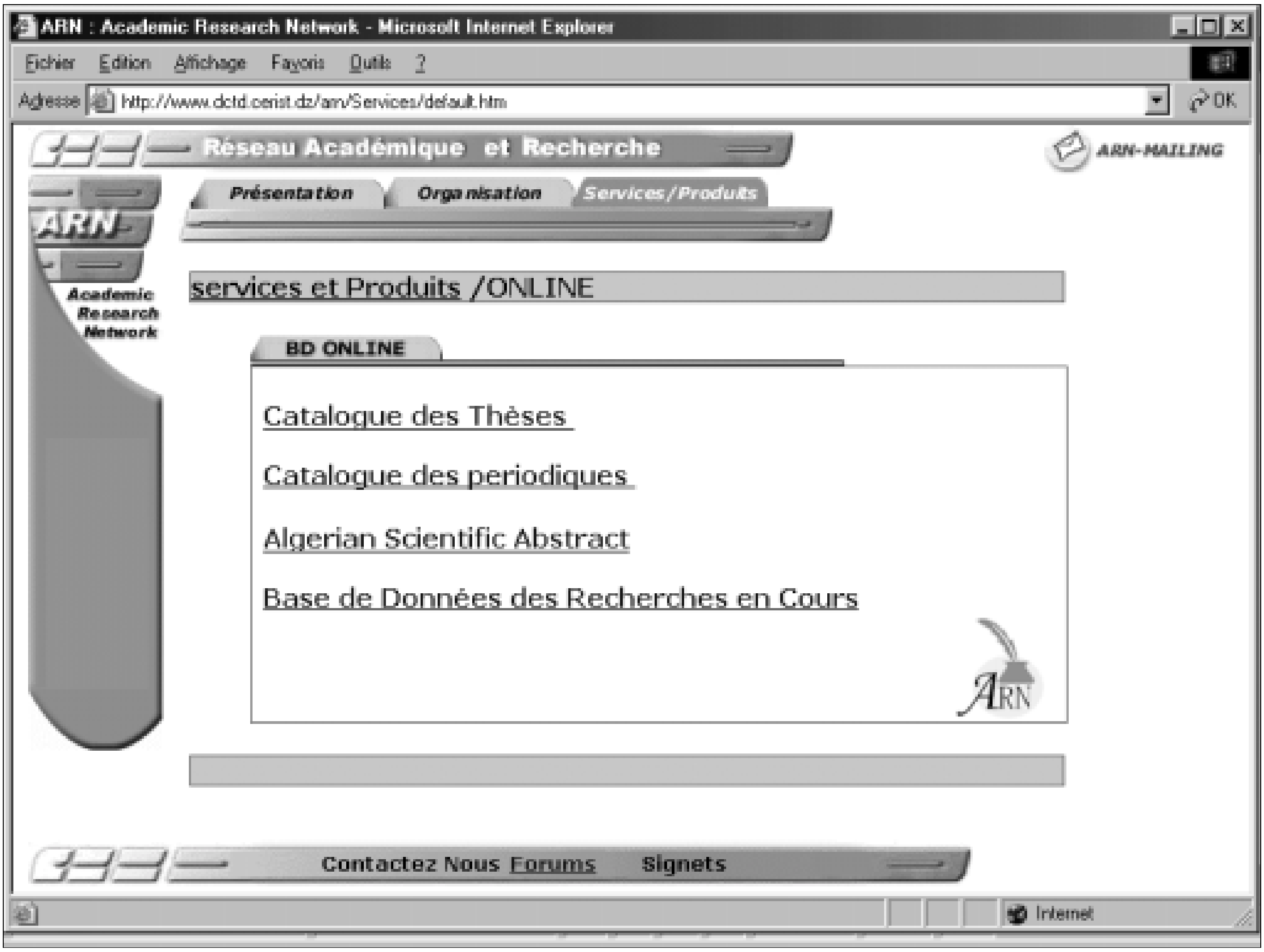
These initiatives give CERIST more and more competencies and know-how in terms of academic data acquisition and processing. However, now there is a need to go beyond bibliographic records, because we know that local users and scholars want to obtain full text and digital content. Within this context and according to an official decree issued in August 2000 by the Ministry of Higher Education and Scientific Research, an electronic copy of every Master's and PhD thesis defended in every academic institution must be deposited at CERIST. The submission of these copies is a condition for getting the diploma. The Centre has been entrusted with the mission of creating a national Electronic Theses and Dissertation (ETD) system and of updating the Current Researches Database (BDRC), a national inventory of ongoing theses and academic works.

Modules for acquisition, control, inventory, recording and processing have been launched. The archiving and the delivery modules are not yet available. The delivery subsystem is under construction but it doesn't seem to be a major problem given the experience of CERIST in hosting academic Web sites. However ETDs files are simply saved in hard disks without a professional archiving plan. Indeed there is a serious need to design the "digital archiving subsystem".

This text is a reviewed and updated version of a communication presented to the ETD2003 conference, Berlin, May 2003.

Yahia Bakelli & Sabrina Benrahmoun. CERIST Research Centre on Scientific and Technical Information, 03 rue des freres Aissou, Ben Aknoun, Algiers (Algeria). Tel: 21321916205-08. Fax: 21321912126. E-mail: ybakelli@mail.cerist.dz

Figure 1. Online access to Algerian bibliographic files created by CERIST.


ARN : Academic Research Network - Microsoft Internet Explorer
Fichier
Edition
Affichage
Favoris
Outils
?
OK
Réseau Académique et Recherche
ARN-MAILING
Présentation
Organisation
Services/Produits
ARN
Academic Research Network
services et Produits /ONLINE
BD ONLINE
Catalogue des Thèses
Catalogue des periodiques
Algerian Scientific Abstract
Base de Données des Recherches en Cours
ARN
Contactez Nous
Forums
Signets
Internet

Adresse http://www.dctd.cerist.dz/arn/Services/default.htm

This paper describes how the ETD system operates and how files are saved, and discusses how we might guarantee that the digital materials deposited by students are preserved for the long term, and how international standards, rules and techniques of digital archiving can be applied to the CERIST system.

CERIST ETD system

By March 2003, a collection of 1463 electronic objects has been collected:

- 1269 floppy disks
- 194 CD-ROMs

An analysis of the statistics of submitted ETDs between October 2001 and March 2003 shows that on average, a minimum of 54 theses is submitted monthly. Right now there are no statistics on the distribution by disciplines, but according to the inventory register, we can establish the following linguistic distribution:

- Arabic theses dominate with 1161 Floppy disks and 97 CD-ROMs.
- French theses number 108 floppy disks and 97 CD-ROMs.

Acquisition of ETDs

The acquisition of ETDs is based mainly on the legal deposit procedure recommended by the High Education Ministry Decree. The electronic version of a thesis is submitted to the library of CERIST in two different ways:

a) By the student himself.

b) By one of the CERIST representatives dispersed over Algeria. Thanks to these regional and local representatives of CERIST, students from universities situated far from Algiers (capital city of Algeria, where the CERIST's Library is located) have the possibility to submit their theses without necessarily moving to Algiers.

Up to now most theses are deposited using the first mode. Of course the most important universities and high academic institutions are concentrated in and around Algiers.

As the first step, a thesis is submitted to the librarian both in electronic and printed versions. Floppy disks and/or CD-ROMs are the digital media submitted by students. Some theses are contained only in one floppy disk; others take more than one disk (but never more than one CD-ROM). The student then fills out one input datasheet (printed datasheet) based on the UNIMARC format.

Control and inventory

The second step is to check the integrity of the electronic object. The librarian must see if the floppy disk contents are readable and not infected with viruses. He also checks that what is contained in the printed version is contained on the electronic copy. Then the thesis title is recorded into a chronological inventory registry.

Bibliographic recording

Then the librarians check the datasheet filled out by the student and complete it according to UNIMARC rules. The data is input into a database called "Depot" using the "SYNGEB" software (developed by CERIST). Indexing of theses is an operation done by a subcontractor at the FNT service (National Repository of Theses). Currently these bibliographic operations are done at two separate stations, one for records in Latin languages theses and other for records related to Arabic theses.

Periodically a set of these bibliographic records is exported to another CERIST database: the BDRC (on going Research Database) in order to update the information about theses being defended.

Conversion and storage of files

Students are usually using MS Word and adopt DOC as the document format of their theses. But the CERIST library has decided to adopt the Portable Document Format (PDF). So it is necessary to convert the deposited files into PDF. This is a mechanical operation using Adobe Acrobat 4.0. But it often takes time because most theses are rarely submitted in one single file. The librarian must convert each file separately then merge them into one unique file. PDF files are then uploaded and saved in two different folders: the "ARN-A" for files in Arabic and the "ARN-F" for files in French (and Latin languages).

Anomalies of submitted digital objects

In fact, librarians frequently detect anomalies concerning the integrity of the digital objects. These anomalies concern both physical and content/logical aspects.

a) Main examples of physical anomalies are:

- Interruption of the uploading process (from the floppy disk to the hard disk). This is certainly caused by the bad quality of floppy disks.
- The presence of a virus infection.

b) Main examples of content anomalies are:

- Absence of the cover page of the thesis.
- Lack of few parts or chapters of the thesis (Table of Contents, bibliography, etc.)

An April 2003 survey conducted at the CERIST Library shows that:

- Among 196 floppy disks in the Arabic theses collection, 43 disks have some of these anomalies.
- Among 108 floppy disks in the French Theses collection, 24 disks have some of these anomalies.
- Among the French Theses collection of 97 CD-ROMs, 31 contained incomplete theses.

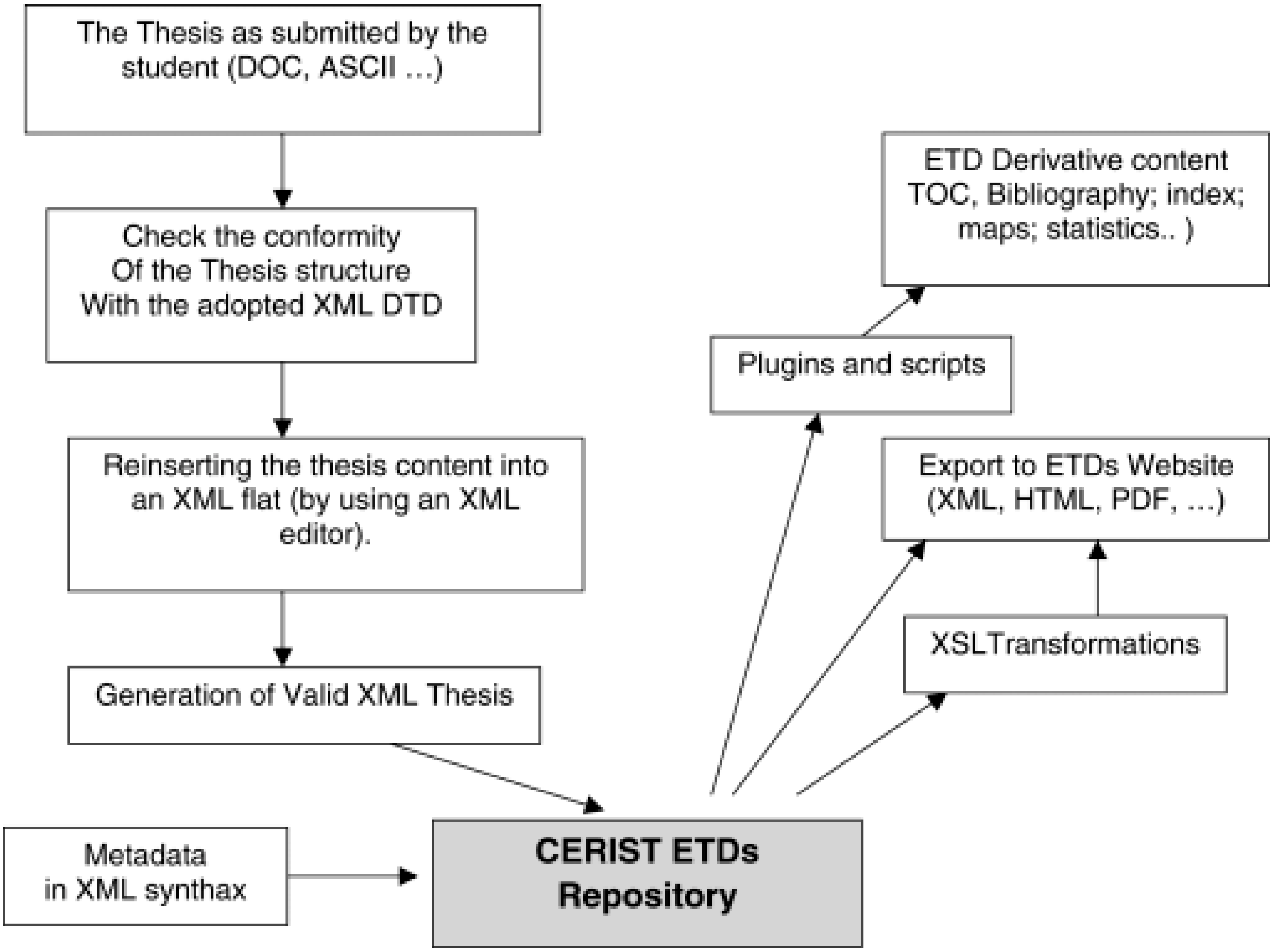
This means that 5% of the deposited floppy disks and 32% of the submitted CD-ROMs cannot be integrated directly and archived into the current collection. Some reparation operations must be done before processing can proceed. These operations consist of two main kinds of actions:

a) Repairing faulty disks and healing infected files.

b) Digitising the missing chapters/content from the printed copies.

These actions must be managed in a way not to complicate the whole process, not to increase the

Figure 2. Process of Archiving ETDs content on XML: Long-term preservation of data and value-added possibilities. [Suggested by Y. Bakelli and S. Benrahmoun (CERIST, Algiers May 2003)]


The Thesis as submitted by the student (DOC, ASCII ...)
Check the conformity Of the Thesis structure With the adopted XML DTD
Reinserting the thesis content into an XML flat (by using an XML editor).
Generation of Valid XML Thesis
Metadata in XML synthax
CERIST ETDs Repository
Plugins and scripts
ETD Derivative content TOC, Bibliography; index; maps; statistics..)
Export to ETDs Website (XML, HTML, PDF, ...)
XSLTransformations

cost of the work and not to affect the quality of the archived files.

Current experimentation: Toward a professional digital archiving plan

With regard to the way the current CERIST ETD system is operating, we must note that theses files are stored in PDF format on two separate hard disks. Original floppy disks and CD-ROMs as submitted by students are kept in boxes and closets. Moreover and even if bibliographic records are entered using the UNIMARC style, they must be saved into a proper format of ".THE" generated by SYNGEB software. These observations lead us to ask some questions regarding the future use of these stored files:

a) Is the decision to adopt PDF as an archiving format a good one? Of course PDF is highly recommended as a delivery format but what would guarantee the independence of the archived files from future Adobe business plans?

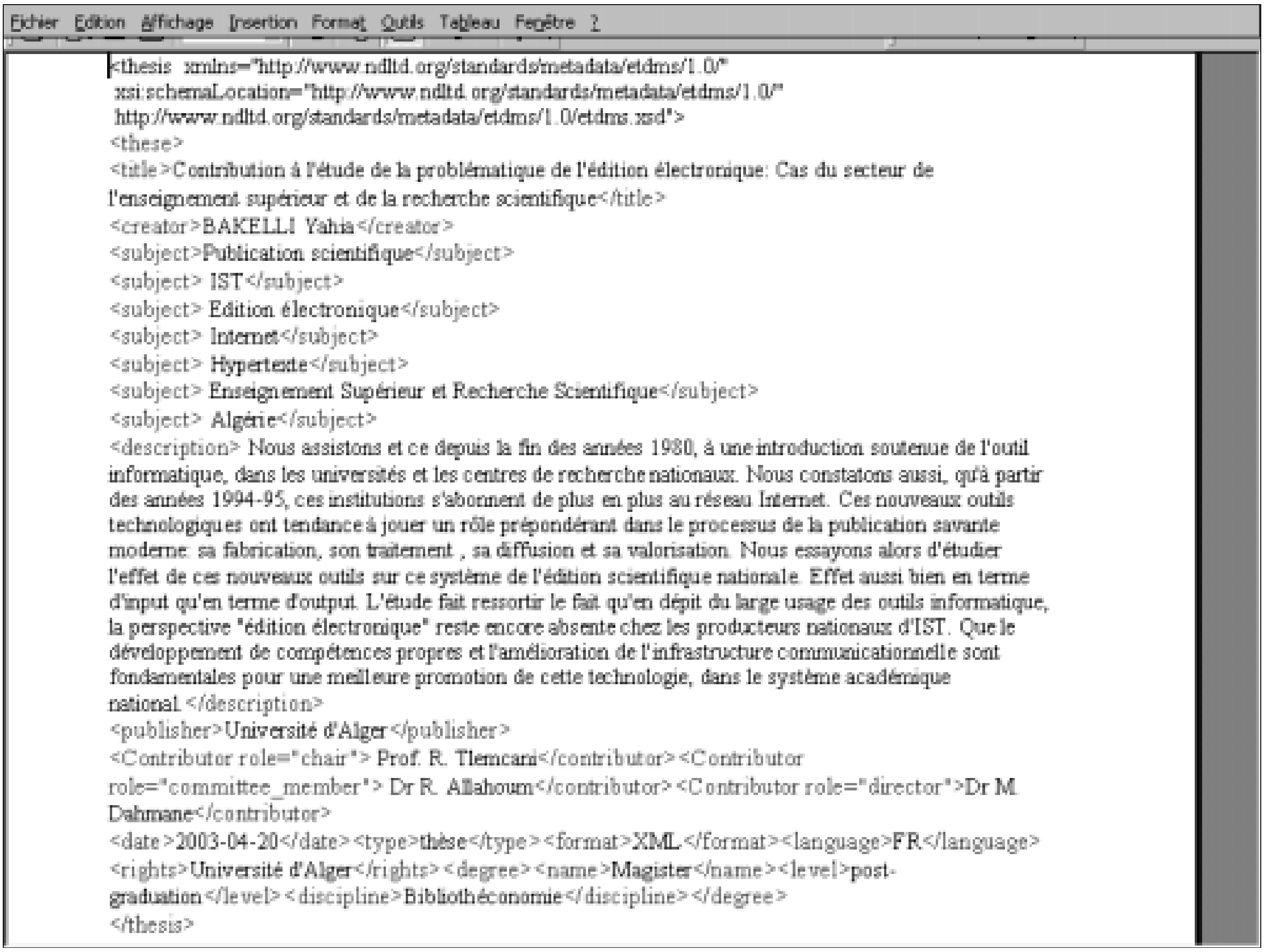
b) How, and at what cost, does such a decision support future manipulations of the content of the theses?

c) What would be the cost of converting bibliographic records each time we need to export them, for ETDs Internet delivery, exchange with other ETD systems, etc?

d) What guarantee is there that original disks deposited by students and stored in boxes and closets will be easily used in future?

e) Do the current machines allocated to this ETD System have enough disk space and memory to contain the mass of files being submitted day after day?

Figure 3. Generation of an ETD-Metadata Standard in XML Format.



```
<thesis  xmlns="http://www.ndltd.org/standards/metadata/etdms/1.0/"
 xsi:schemaLocation="http://www.ndltd.org/standards/metadata/etdms/1.0/"
 http://www.ndltd.org/standards/metadata/etdms/1.0/etdms.xsd">
<these>
<title>Contribution à l'étude de la problématique de l'édition électronique: Cas du secteur de
l'enseignement supérieur et de la recherche scientifique</title>
<creator>BAKELLI  Yahia</creator>
<subject>Publication scientifique</subject>
<subject> IST</subject>
<subject> Edition électronique</subject>
<subject> Internet</subject>
<subject> Hypertexte</subject>
<subject> Enseignement Supérieur et Recherche Scientifique</subject>
<subject> Algérie</subject>
<description> Nous assistons et ce depuis la fin des années 1980, à une introduction soutenue de l'outil
informatique, dans les universités et les centres de recherche nationaux. Nous constatons aussi, qu'à partir
des années 1994-95, ces institutions s'abonnent de plus en plus au réseau Internet. Ces nouveaux outils
technologiques ont tendance à jouer un rôle prépondérant dans le processus de la publication savante
moderne: sa fabrication, son traitement , sa diffusion et sa valorisation. Nous essayons alors d'étudier
l'effet de ces nouveaux outils sur ce système de l'édition scientifique nationale. Effet aussi bien en terme
d'input qu'en terme d'output. L'étude fait ressortir le fait qu'en dépit du large usage des outils informatique,
la perspective "édition électronique" reste encore absente chez les producteurs nationaux d'IST. Que le
développement de compétences propres et l'amélioration de l'infrastructure communicationnelle sont
fondamentales pour une meilleure promotion de cette technologie, dans le système académique
national. </description>
<publisher>Université d'Alger</publisher>
<Contributor role="chair"> Prof. R. Tlemcani</contributor><Contributor
role="committee_member"> Dr R. Allahoum</contributor><Contributor role="director">Dr M.
Dahmane</contributor>
<date>2003-04-20</date><type>thèse</type><format>XML</format><language>FR</language>
<rights>Université d'Alger</rights><degree><name>Magister</name><level>post-
graduation</level><discipline>Bibliothéconomie</discipline></degree>
</thesis>
```

These are actually more predictable problems than questions. We argue that the archiving module of the CERIST ETD system must be redesigned in order to overcome these constraints. Some procedures and tools must be integrated into this ETD system in order to make the media safe for all time and their files permanently readable and independent from the evolution of machines, platforms and software. Also the ETD content must be archived in a way so that it can be manipulated and reused directly without need for preliminary operations. It must also be saved in an economic way that makes it possible to deliver the same content in different forms and contexts (e.g. full text database, OPAC, Internet portal, digital library) and for different user profiles. This is what we are calling a dynamic archiving of content (see Figure 2).

Other international experiments and programs tell us of the necessity to distinguish between two main levels:

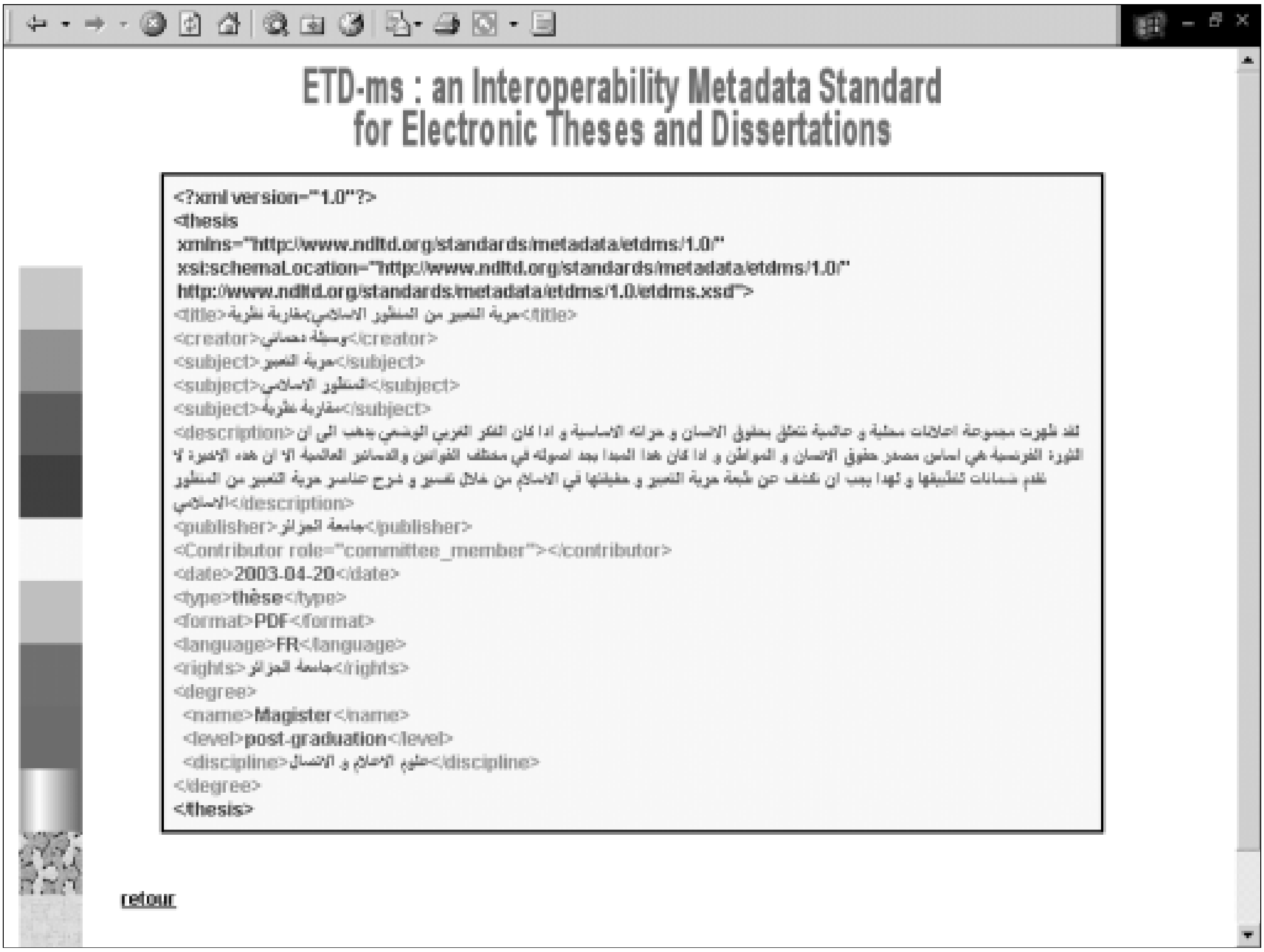
a) Conservation of the digital object itself.

b) Preservation of data and content of the ETD.

Currently we are operating three sets of experimentation with a sample of submitted electronic theses. The sample is about 430 objects (30% of the whole collection). Tests and experimentation concern these two levels. However we are mainly focusing on the second one, the preservation of content.

- The first category of tests concerns the media used for backup of the data. Several technologies exist in the market, so the aim of this test is to identify, for the CERIST ETDs case, the difference in terms of quality of

Figure 4. Application of ETD-Metadata Standard for Arabic texts.



ETD-ms : an Interoperability Metadata Standard
for Electronic Theses and Dissertations

```
<?xml version="1.0"?>
<thesis
 xmlns="http://www.ndltd.org/standards/metadata/etdms/1.0/"
 xsi:schemaLocation="http://www.ndltd.org/standards/metadata/etdms/1.0/"
 http://www.ndltd.org/standards/metadata/etdms/1.0/etdms.xsd">
<title>حرية التعبير من المنظور الاسلامي:مقاربة نظرية</title>
<creator>وسيلة دهماني</creator>
<subject>حرية التعبير</subject>
<subject>المنظور الاسلامي</subject>
<subject>مقاربة نظرية</subject>
<description>لقد ظهرت مجموعة اعلانات محلية و عالمية تتعلق بحقوق الانسان و حرياته الاساسية و اذا كان الفكر الغربي الوضعي يذهب الى ان
الثورة الفرنسية هي اساس مصدر حقوق الانسان و المواطن و اذا كان هذا المبدا يجد اصوله في مختلف القوانين والدساتير العالمية الا ان هذه الاخيرة لا
تقدم ضمانات لتطبيقها و لهذا يجب ان نكشف عن طبيعة حرية التعبير و حقيقتها في الاسلام من خلال تفسير و شرح عناصر حرية التعبير من المنظور
الاسلامي</description>
<publisher>جامعة الجزائر</publisher>
<Contributor role="committee_member"></contributor>
<date>2003-04-20</date>
<type>thèse</type>
<format>PDF</format>
<language>FR</language>
<rights>جامعة الجزائر</rights>
<degree>
  <name>Magister</name>
  <level>post-graduation</level>
  <discipline>علوم الاعلام و الاتصال</discipline>
</degree>
</thesis>
```

using a WORM, DVD or DAT solution. Also we aim to define the method of identifying the appropriate back-up for a given quantity of bytes and data, and the archiving system architecture we must adopt to optimise the archiving activity with a minimum of data loss risk.

- The second set of tests addresses the concepts of Refreshing, Migration and Emulation. Which technique is the most appropriate for the CERIST ETDs context? Given the limited budget of the organisation, the cost of the techniques is an important criterion. Moreover, regarding the relatively limited skills of librarians involved in the project, we must recommend the simplest technique. We believe that "Refreshing" seems the cheapest and most simple method.
- The third set of experimentation concerns the proper content of ETDs. How must the structure of submitted dissertations be reformatted in order to preserve them for the long term? Based on the belief that XML is the most suitable standard for such formatting, we are doing several actions to give answer to these questions: Do we have to opt for well-formed XML Files or Valid XML Files? The first kind of XML has the advantage of being simple to produce and economic for massive workflow chains but it has the disadvantage of decreasing the possibilities for automated manipulations later. The second kind of XML files is of course better but needs many manual corrections and takes more time before we can save the file into the archive.

Next we compare two existing XML DTDs:

a) The DiML developed at the Humboldt University of Berlin (http://edoc.hu-berlin.de/diml/), and which is adapted from the DTD developed since 1985 at Virginia Tech (http://etd.vt.edu/).

b) The TeiLite DTD as adopted by certain ETDs chains such those of the Presses de l'Université de Montreal (Canada) and Université Lumière Lyons2 (France) within the "Cyberthèses" chain.

This comparative study of DTDs is based on the following parameters:

- easy to interpret.
- appropriate for a wide range of disciplines (e.g. humanities, medicine technology, chemistry)

As digital content archiving concerns not only the full text of theses but also their metadata, we are generating metadata for the chosen sample of ETDs. In this way we decided to adopt the model of the ETD-MS of Virginia Tech, an adapted Dublin Core metadata which we generate in XML (see Figure 3). One of the most interesting results of this test is the demonstration of the feasibility of this standard not only for texts in Latin languages but also for Arabic texts (see Figure 4).

Another output of our experimentation is the design of a naming scheme for the dissertations collection to serve as the protocol of how files must be stored in directories. This scheme must take into consideration the adopted codification system. However we expect to go beyond the simple class of disciplines, “1”, for technology and pure sciences, “2” for medicine and life sciences and “3” for human and social sciences. In this way the URN handle-server technique is currently applied to the CERIST ETDs sample.

Conclusion

As the number of submitted electronic theses increases, more and more anomalies are being reported. This has made it necessary to think seriously about the risk of unavailability of content in the future. Having a clear archiving plan as soon as possible will avoid another more complex problem, that of “managing the retrospective” or the “past”. This last one is often very difficult to resolve without big costs.

Based on the current survey and experiments, international standards and concepts of digital archiving can be effectively applied to the local context of the CERIST ETD system. However an important effort is needed to identify for each concept or standard the most appropriate solution and application. Generally speaking there are two important parameters for such a decision: the cost and the facility (less required skills). Given that digital archiving cannot be separated from the other modules of capture, processing and delivery, we need to adopt one generic and exhaustive ETD chain, following the NDLTD and Cybertheses models. XML will constitute one major option of the CERIST ETD system. We are aiming at:

- involving the student in the process of archiving, stressing the principle that digital archiving is more and more facilitated and performed when it is observed in an early and upstream step of the chain. In this way we are trying to study the possibility of application and translating into Arabic of the “The Guide for Electronic Theses and Dissertations” hosted at the Web site http://etdguide.org.
- Introduce the digital archiving life cycle concept.

Moreover the analysis of international models of ETD archiving lets us extract some important lessons and trends:

- At the opposite of technique and standards issues, methodology and policy aspects of digital archiving still need to be developed. Such aspects are very important in our context where there is not enough budget, means and skills but there exists an urgent need to control complex and massive academic dissertations.
- The scarcity of literature in terms of archiving ETD experiences. International models of ETDs rarely describe their archiving models. This leads us to recommend the encouragement of initiatives like the UNESCO ETDs Clearinghouse to encourage international communication and creation of spaces where it will be possible to share experiences, tools and ideas.
- The OAI (Open Archives Initiative) concept seems to constitute one of the most important future trends. One of our future targets is to introduce an OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting, see http://www.openarchives.org/OAI/openarchivesprotocol.html) into the CERIST ETD system in order to link it with regional and international ETDs systems.

Bibliography

Bakelli Y. 2002. Digital Access to Scientific Content in Algeria: CERIST Initiatives. *The 11th International Conference for Science Editors*; August 24–28, Beijing (China); Poster session P01.

Bakelli Y. 2000. Model for an electronic access to the Algerian Scientific Literature: short description. In *Research and Advanced Technology for Digital Libraries: 4th European Conference, ECDL 2000, Lisbon, Portugal, September*. Proceedings. Lectures Notes in Computer Science, 432–36.

Cybertheses. URL: http://sophia.univ-lyon2.fr/CyberTheses/index.php [viewed 17 November 2003]

ETD-ms an Interoperability Metadata Standard for Electronic Theses and Dissertations. URL: http://www.ndltd.org/standards/metadata/ET-ms-v1.00.html [viewed 17 November 2003]

Guide for Electronic Theses and Dissertations (The). URL: http://etdguide.org/ [viewed 17 November 2003]

Humboldt University of Berlin edoc – DiML The Dissertation Markup Language (DiML). URL: http://edoc.hu-berlin.de/diml/ [viewed 17 November 2003]

Policies for digital preservation. 2003. *ERPANET Training Seminar Fontainebleau (France) 29–30 January*. URL: http://www.erpanet.org [viewed 17 November 2003]

Third International Symposium on Electronic Theses and Dissertations: applying New Media to Scholarship. 2000. March 16–18; University of South Florida, St. Petersburg, Florida. URL: http://etd.eng.usf.edu/conference/bios.htm [viewed 17 November 2003]

Virginia Tech ETD. URL: http: /etd.vt.edu/ [viewed 17 November 2003]

Editorial history:
paper received 7 July 2003;
accepted 12 November 2003.