Long-term preservation of ETDs in Algeria

Discussion Through CERIST Deposit System

Yahia Bakelli
Sabrina Benrahmoun
CERIST Research Centre on Scientific and Technical Information
ybakelli@mail.cerist.dz
03 rue des freres Aissiou, Ben Aknoun, Algiers (Algeria)
www.dst.cerist.dz/equipe-edition-electronique.htm

Keywords: Long term preservation, ETD, Algeria, CERIST, Archiving, XML.

Abstract

In Algeria, 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 Research Centre. This as a condition to get the diploma. CERIST is then entrusted with a mission to build a database of national theses and update of the national inventory of current theses and researches. However a serious problem of archiving and preserving of these ETDs is more and more arising. Thus, We have noticed that from December 2001 until November 2002 a great number of ETDs is deposited and constitute a set of more than 1000 floppy discs and 100 CD-ROMs.

What would guarantee that these digital materials deposited by students might be preserved and safeguarded? What would guarantee that the content of these materials might be preserved and accessible at any time regardless the machine, the operating SYSTEMs, and software. the user will be running in the future?

Through our paper we will clarify the problem of the long-term conservation and preservation of electronic documents in the Algerian context? And in which way we may apply an international recognised standards and techniques for setting up and organising the local ETD's archives?

Preface

The CERIST Research Centre on Scientific and Technical Information of Algiers was created in 1985, with the main mission of design and implementation of the National Information System. Within this framework, a big importance is given to the academic literature recording and availability. This was practiced through a production of national bibliographic databases and national union catalogues. Examples of these bibliographic products were Algerian Scientific Abstracts, Algeriana, CAT (Algerian Theses Catalogue), FNT (National Theses Repository), which can be interrogated through the Academic Research Network and CERIST websites.

These initiatives are giving CERIST more and more competencies and know how in a term of academic data acquisition and processing. However, now there is a need to go beyond the bibliographic records. Because we have to establish that local users and scholars are in need of obtaining 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 Research Centre. The submission of these copies is a condition to get the diploma. Then the centre is entrusted with a mission to create a national ETD SYSTEM, and to update the Current Researches Database (BDRC), which is a national inventory of current theses and academic works.

Right now modules of acquisition, control, inventory, recording and processing were launched. The archiving and the delivery modules are not launched yet. In fact the "Delivery subSYSTEM" is under construction and doesn't seems to pose big problems regarding the experience of CERIST in term of academic websites hosting. However we must establish that Theses files are simply saved in hard disks without a professional archiving plan. Indeed there is a serious need to design the "digital archiving subSYSTEM".

So what we aim through our current study is to understand how the ETD SYSTEM is operating and how these files are saved? And what would guarantee that these digital materials deposited by students might be preserved for a long term? Then how international standards, rules and techniques of Digital archiving can be applied to this SYSTEM?

The ETD System of CERIST: A New Chain In Need of Digital Archiving Module

As impact of this decree, an ETD chain is setting. Thus until the 13th of march 2003, a collection of 1463 electronic media is constituted. This quantity can be detailed as following:

- 1269 floppy disks
- 194 CDROMs

So 87% of media are floppy disks and 13% are CDROMs.

The Analysis of statistics of theses submitted between October 2001 and march 2003 shows that in average 54 theses are submitted monthly i.e. 54 digital media at least.

Right now there are no statistics of such a distribution by "disciplines", but according the inventory register, we can establish the following "linguistic" distribution:

- Arabic theses are dominating with 1161 Floppy disks and 97 CDROMs.
- French theses with 108 floppy disks and 97 CDROMs.

Acquisition of ETDs

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

1. The Student himself.
2. One of The CERIST representations distributed among the Algerian territory. Thanks to these regional and local representations of CERIST, students from universities situated far from Algiers (capital of Algeria, where the CERIST's Library is located) have possibilities to do the deposit of their theses without necessity to move to Algiers.

But we have to mention that up to now the first mode is still dominating as an acquisition source of ETDs. Of course most important number of universities and high academic institutions are concentrated in and around Algiers.

As a first step, a thesis is submitted to the librarian both in electronic and print versions. Student is then invited to fill out one input datasheet (printed datasheet) respecting the UNIMARC format. This datasheet is then put into limps (these datasheets must be checked by the librarian in a next step). 3"1/4 floppy disks and CD-ROMs are the digital media given by students. Some theses are contained only in one floppy disk other take more than one disk (but never more than one CDROM).

Control and Inventory

As a second step it's question to check the integrity of the electronic media. The Librarian must see if the floppy disk is running well and if is empty of viruses. Also he must check if all what's is contained in a printed version is contained on the given electronic copy. Then the thesis title is reported into an inventory registry following a chronological order.

Codification

After stamping both the electronic and the printed copies, a shelf code is assigned and reported in the cover of the print copy and in the label of the digital media. For Example the thesis coded: "THA.3.905" where:

- "TH" for thesis.
- "A" for thesis written in Arabic.
- "3" for thesis in the field of Human and Social Sciences.
- "905" to indicate a sequential number in the collection.

Bibliographic Recording

At this step librarians check the datasheet filled out by the student and complete it according Unimarc rules. The input is then done into a Database called "Depot" using a "SYNGEB" software (developed in the CERIST). Indexing of theses is an operation done by a subcontracting at the FNT service (National Repository o Theses). Currently these bibliographic operations are done into two separated 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 (Current Researches Database) in order to update the information about theses being defended.

Files conversion and Storage

Students are usually using MS Word tool and adopt DOC as edition format of their theses. But the CERIST library decides to adopt the PDF Format. So there is necessity to convert the deposited files into PDF. This is being a mechanical operation using Adobe Acrobat 4.0. But it often takes a time to do because one thesis is rarely given in one unique file. So librarian must do a conversion of each file separately then merge them in 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). These folders are currently saved in two separated machines: folder ARN-A is saved into a machine containing the Arabic "Depot" Database and ARN-F is saved into a machine where is built the French "Depot" database. Original floppy disks and CDROMs given by students are finally kept into boxes and cupboards.

Anomalies of Submitted Digital media

In fact, librarians are frequently detecting anomalies concerning the digital media integrity. These anomalies are concerning both physical and logical aspects.

a) Main examples of Physical anomalies are:

- The 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 (TOC, bibliography,..).

The survey we conducted in April 2003 at the CERIST Library shows that:

- Among 196 floppy disks constituting a part of the Arabic theses collection, 43 disks were concerned by theses anomalies.
- Among 108 floppy disks constituting the French Theses collection, 24 disks were concerned by these anomalies.
- Among the French Theses collection of 97 CDROMS, 31 CDROMs were contained an incomplete theses.

This means that 05% of the deposited floppy disks and 32% of submitted CDROMs cannot be directly integrated and archived into the current collection. Some reparation operations must be done before. These operations consist on two main kinds of actions:

a) Repairing faulty disks and healing infected files.

b) Digitising of missing chapters and content from the printed copies.

All these actions must be managed in a way to do not complicate the whole process, do not increase the cost of the work and do not affect the quality of the archived files.

Current Experimentations: Toward a Professional Digital Archiving Plan

Regarding the way in which the current CERIST ETD SYSTEM is operating we must notice that theses files are stored in PDF format in two separated hard disks. Original floppy disks and CDROMs are kept into boxes and cupboards as given by students. Moreover and even if Bibliographic recorded are entered respecting UNIMARC but they were saved into a proper format of ".THE" generated by the SYNGEB software. These observations lead us to ask some of questions regarding the future use of these stored files:

- Is the decision to adopt PDF as an archiving format a good one? Of course

PDF format is highly recommended as a delivery format but what would guarantee the independency of the archived files from future Adobe business plan?

- How and which cost do the future developments and manipulations of theses content will be facilitated under such a decision?
- What would be the cost of converting bibliographic records at each time we need to export them, for ETDs Internet delivery; exchange with other ETD SYSTEMs; etc?
- What guarantee that original disks deposited by students and stored in boxes and cupboards will be easily reused in future?
- Do the current machines allocated to this ETDs System have enough disk space and memory to contain a mass of files being submitted day after day?

Actually it consists more on predictable problems than on questions. So what we would like to argue is the fact that the archiving module of the CERIST ETD SYSTEM must be redesigned in order to avoid all these constraints. Some procedures and tools must be integrated into this ETD SYSTEM in order to make media safe all the time and their files permanently readable and independent from the evolution of machines; plate-forms and softwares. Also ETDs content must be archived in a way to be manipulated and reused directly without need to preliminary operations. It must also be saved in an economic way that gives possibility to deliver the same content in different forms and contexts (full text database; OPAC, Internet portal; digital library…) and for different user's profiles.

What we learn from international experiences is the necessity to distinguish between two main levels:

a) Conservation of the Digital media itself.

b) Preservation of data and content of the ETD.

Currently we are operating three set of experimentations into a sample of submitted electronic theses. The sample is about 430 media (30% of the whole collection). Tests and experimentations are concerning the two mentioned levels. However we are mainly focusing on the second one i.e. "preservation of content":

- First category of tests is concerning the media of backup. Of course several technologies exist in the market, so the aim of this test is to identify, for the CERIST ETDs case, the difference in term of quality using WORM, DVD or DAT solutions. Also we aim to define the method to identify the appropriate backup for a given quantity of bytes and data. And what's the archiving SYSTEM architecture we must adopt to optimise the archiving activity with a minimum of data loss risk.
- The second set of tests is concerning the concepts of Refreshing; Migration and Emulation. Which technique is the most adequate for the CERIST ETD Context? Of course and regarding the limited budget of the organisation, the cost of the technique will be important criteria. Moreover, and regarding the relatively limited skills of librarians implied in the project we have to recommend the simplest technique. So we are going on the assumption that "Refreshing" seems cheap and simple.
- The third set of experimentations is concerning the proper content of ETDs. How does the structure of submitted dissertations must be reformatted in order to make them preserved for a long term? Going on

the axiom that XML is the most suitable standard for such formatting we are doing several actions to give answer to several questions:

Do we have to opt for the well formed XML Files or Valid Xml Files? The first kind of XML has the advantage to be simple to produce and economic for a massive workflow chains but it presents the inconvenient to decrease possibilities of later automated manipulations. The second kind of XML files is of course better but needs many manual corrections and more time before the save of the file into the archive.

Now we are comparing two existed 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 in 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.
- appropriated for a wide range of disciplines (humanities; medicine technology, chemistry, and so on).

As the digital content archiving is concerning not only the full text of theses but also their metadata, we are generating a metadata of the chosen sample of ETDs. In this way we decided to adopt the model of the ETDMS of Virginia Tech. An adapted DC metadata which we are generating in XML format. 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.

As another output of our experimentations we are designing a "naming scheme" of the dissertations collection to serve as a protocol of how files must be stored in directories. This scheme must take into consideration the adopted codification SYSTEM (see section 2. c). However we have ambition to go beyond the simple class of disciplines[1]. In this way the "URN handle-server" technique is currently applied for the CERIST ETDs sample.

Conclusion

In parallel to an increasing mass of submitted electronic theses, anomalies are more and more reported. This is justifying the necessity to think seriously about the risk of unavailability of content in future. The importance of a digital archiving problematic may be argued by the fact that as the CERIST ETDs is just setting so having a clear archiving plan as soon as possible will give the opportunity to avoid another more complex problematic of "managing the retrospective" or the "past". This last one is often very difficult to resolve without big dispenses.

It appears from the current survey and experiences we are conducting that international standards and concepts of digital archiving are effectively applied for the local context of the CERIST ETDs SYSTEM. However an important effort is to do in order to decide for each concept or each standard the most appropriated solution and application way. In general two parameters are decisive for such a decision: the cost and the facility (less required skills). Regarding the fact that digital archiving cannot be separated from the other modules of capture, and processing and delivery, we need to adopt one generic and exhaustive ETD chain, in a model of NDLTD, Cybertheses, etc. The XML will constitute one major option of the CERIST ETDs SYSTEM. As a perspectives we are targeting to:

- Make the student implied in the process of archiving. This starting from the principle that digital archiving is more and more facilitated and performed when it's observed in an early and upstream step of the chain. In this way we are trying to study the possibility of application and arabisation of the "The Guide for Electronic Theses and Dissertations" hosted at the etdguide.org.
- Introduce the digital archiving life cycle concept.

Moreover the analysis of international models of ETDs archiving let us to extract some of important lessons and trends:

- At the opposite of "techniques" and "standards" issues, "methodology" and "policy" aspects of digital archiving still in a need of a development. Such aspects are very important in our context, where there is not enough budgets, means and skills in parallel to an urgent need to control a complex and massive academic dissertations.
- The rarity of literature in term of archiving ETDs experiences. International models of ETDs rarely described their archiving models. this leads us to recommend the encouragement of initiatives such the "UNESCO ETDs clearing house" to encourage international communication and creation of spaces where it will be possible to share experiences, tools and ideas.
- The OAI concept seems constituting one of the most important future trend. So one of our future targets is to introduce an OAI-PMH into the CERIST ETDs System in order to link it with regional and international ETDs SYSTEMs.

[1] "1", for technology and pure sciences, "2" for medicine and life sciences and "3" for human and social sciences.

Bibliography

1. Y. BAKELLI : Digital access to scientific Content in Algeria: CERIST Initiatives. . Hrsg.: IFSE: The 11th International Conference for Science Editors; August 24-28, 2002, Beijing (China); Poster session P01.. ISBN 3-540-41023-6.
2. Y. BAKELLI : Model for an electronic access to the Algerian Scientific Literature: short description. Hrsg.: SPRINGER: Research and Advanced Technology for Digital Libraries: 4th European Conference, ECDL 2000, Lisbon, Portugal, September 2000. Proceedings. 2000, S. pp.432-436 ,
3. Cyberthèses . Hrsg.: UNIV. LYON 2 : http://sophia.univ-lyon2.fr/CyberTheses/index.php.
4. ETD-ms an Interoperability Metadata Standard for Electronic Theses and Dissertations . Hrsg.: VIRGINIA TECH : http://www.ndltd.org/standards/metadata/ETD-ms-v1.00.html.
5. Guide for Electronic Theses and Dissertations (The) . Hrsg.: UNESCO: http://etdguide.org.
6. Humboldt University of Berlin edoc - DiML The Dissertation Markup Language (DiML) . http://edoc.hu-berlin.de/diml/.
7. Policies for digital preservation . Hrsg.: ERPANET: http://www.erpanet.org.
8. Third International Symposium on Electronic Theses and Dissertations: applying New Media to Scholarship . Hrsg.: University of South Florida, St. Petersburg, Florida.: http://etd.eng.usf.edu/conference/bios.htm.
9. Virginia Tech ETD . Hrsg.: Virginia Tech : http://etd.vt.edu.