Metadata Migration into *Greenstone*: *A case study in a small size library*

Luis Manuel Rocha
Department of TIC
Universidade Atlântica
Barcarena, Lisboa - Portugal
Luis-l-rocha@telecom.pt

Filipa Taborda
Department of TIC
Universidade Atlântica
Barcarena, Lisboa - Portugal
ftaborda@uatlantica.pt

Abstract — **As a case study, this paper focus on an information delivery problem which is how a small university library dealt with the problem of cataloging and delivering access to a digital collection of final works in all courses. Those collections are very useful to potential students and final students in all fields of studies.**
Documents were somehow already classified and arranged in an excel list. The aim of this work is the automatic migration from an excel list of classified documents into a digital library tool. The open source digital software *greenstone* was adopted in the university library because it fits the needs and the way the library is organized.

Keywords-digital library; XML; migration; greenstone

I. Introduction

A Digital Library is a powerful way to distribute and manage information in many distinct scenarios [1]. As such it was considered to solve the problem of a small size library. The problem was: how to manage the digital documents collection of final works in a way that students or prospective students can easily find work done by other students within a specific subject, year, course or teacher coordinator and mark. The list of the classified documents is kept in an excel list with the information of the correspondent document file name.

The study began with the selection of the open source tools most in use by other universities. The organization of the library, centered in the library responsible, and the need for keeping low the application support effort [2] were decisive in the choice of adopting *greenstone*.

The objective of this study was to build a *greenstone* collection with the digital documents referred above and migrate their correspondent metadata already resident in an excel file into the collection avoiding the extra effort of doing that one by one as it would be supposed to be done before compiling the collection. After finding the location of the metadata source file used by *greenstone* in the compiled collection and as it is stored as an XML file, [1] the problem of the migration was limited to the problem of finding an automating way to add and arrange data in the XML collection metadata source file.

For this remaining problem, a prototype, software application, was developed and tested. The original excel file is presented as input, registers are picked and after being formatted are included in the *greenstone*'s collection metadata source file.

Next to the introduction, a bibliographic revision is made on the subject standards and around the consulted sources to get ideas on the implementation. After that, a description of the methodology leading to the solution is made as well as the reference of some practical aspects related with the format of the target file. Finally the continuity of the project is described as future work and some conclusions were built out of this case.

II. Literature review

The concept of digital library was the first object of the study. We studied the aim of the subject, the standards and available tools. *Greenstone*, the chosen digital library software, was also object of study to get into the solution. The literature review presented here covers the concepts and some relevant practical aspects in use with g*reenstone*

A. About Digital Libraries

The concept of digital library has evolved between the research community, focused on solving their specific problems with emphasis on technology, and the library community focused on institutional and service issues of libraries [3].

Technological developments in recent decades have encouraged the use of such applications, has attracted interest from new authors and the emergence of new and more comprehensive definitions. For Michael Lesk [4] a Digital Library is an organized collection of digitalized information. David Bainbridge [1] encompasses the concepts of service and maintenance, whereas a digital library is an organized collection of information, focused on digital objects, text, video and audio, along with methods for access and retrieval, selection, organization and maintenance of the collection. The

European Commission in its Communication of the Commission to the European Parliament in 2005 [5] highlight the possible sources of digital objects that make up the library, digitized or born digital.

A digital library is not a "digitalized library". Its purpose is not to replace traditional books for their digital versions. Books have their own place and standards [1].

Another term that usually comes associated with this type of repositories is the virtual library. A virtual library does not contain documents, contains pointers to the locations, where those documents reside, showing similarities to a portal [1].

It is also noticeable the difference between an internet information repository equipped with a search engine, and a digital library, lacking the first for organization, careful revision, selection and classification of digital objects.

B. Preservation of digital objects

The typical volatility of digital information leads us to deeply reflect on the contents of a digital collection. The quality and a plan to maintain the collection must be thought out accordingly to each community and subjects [4].

An interesting view on these issues have Professor Moxley, Joseph M., South Florida University, Tampa, USA, he says: “A document that can be read over the course of several years by many people is preferable to a document available for a million years and read only by a few people” [6].

III. Digital library support

Analyzing the open source platforms that can be used to develop a digital library was the second step into the solution. Two of them showed up to be the most widely used, g*reenstone* and *D-space* [7].

It is not part of this work to develop a comparative study between the two platforms; it is worth to point out the main characteristics that determined the choice of g*reenstone* [2]:

- Infrastructure support: **D-space** has been designed to be used in institutions with centralized facilities and competent support. *Greenstone* can be installed on any platform; can be a simple laptop or a corporate system requiring only basic computer knowledge.
- Guidance (Librarian vs. Author): *Greenstone* interface is librarian oriented. It has an intuitive graphical user interface that allows selection of digital elements, enrichment and construction of a digital library. It does not allow the users (typically authors) to dynamically add objects to the collection. *D-space* is clearly author oriented, allows end users to submit their own objects and the respective metadata, requiring only the completion of three fields of the scheme. Also allows the design of a librarian interface but this process is not accessible to the ordinary end user.
- Metadata: With *greenstone* it is possible to adopt different metadata schemas in different collections, including the design of ad-hoc schemes. *D-space* uses only the Qualified Dublin Core for the whole library.
- Portability: In g*reenstone* is possible to export the library to an ordinary CD-ROM allowing it to run on all Windows environments (including versions 3.1x).

These platforms have clearly different orientations. The choice to use *greenstone* or *D-space* depends on the objectives of each project. In both, the compatibility of protocols ensures easy migration between them.

IV. Methodology

The methodology followed in this case can be organized in four stages:

- Understanding the real necessity of the library and the way it works. Adjusting a digital library solution to give technical support to the library’s information management need. At this point the platform was chosen according to the library way of working, the type and number of documents.
- After the decision about the proper digital library tool, the remaining problem was: How to migrate the excel list of already classified documents to g*reenstone*, as well as future ones, avoiding the need of documents manual classification in *greenstone*’s librarian interface.
- The solution first approach was a try in finding an easy adaptable module of the application in g*reenstone* community. This approach showed to be not easy to follow due to lack of knowledge in Pearl [8]. The chosen approach followed the way of working out the XML file containing the source metadata of the *greenstone’s* collection documents.
- Development of the prototype. The developed software uses as input an MsExcel list of document’s metadata and completes the correspondent *greenstone*’s metadata source file. The documents listed must be also in a *greenstone* collection, with the same file names, so that the correspondent metadata can be mapped automatically.

Further investigation and search in the *greenstone* community support, clarified and consolidated the option to develop a migration tool, capable to assure the creation of an appropriate XML code in the *metadata.xml* file, receiving, as input, the initial list of classified documents.

V. Migration

This section refers to the most relevant points considered in the development of the migration tool, its limitations and requirements to work properly. Tests realized are also described shortly.

A. Grennston’s classification metadata file

Document metadata in *greenstone* is supported by a specific structure, according to the chosen metadata schema, written in XML. The XML file, named *metadata.xm*l, is stored

Identify applicable sponsor/s here. If no sponsors, delete this text box. (sponsors)

in the directory where the collections objects are placed sfter being compiled.

The *metadata.xml* file is read during the compile process of the collection which is responsible for converting the original format of documents in the composed file format used internally by the platform. This composed or native format includes the metadata described in XML for each document [1].

B. Development tool

In order to maintain a technically simple approach, MS Visual Studio 2008 was chosen as the development platform. A project based on Windows Forms Application was created and the code was written in Visual Basic. This choice allows an easy development environment of an intuitive interface for users used in Windows operating system [9].

C. Metadata specification

A metadata schema based on Dublin Core specification was used to support this proof of concept since the university intends to adopt a variant of this standard.

D. Software Functionality

The user interface provides the user with the ability to specify the location of the source file as being the list of document's metadata in MS Excel format and the location of the file to be modified, *the metadata.xml* file.

Conversion takes place:

- Determines the size of the list of classifiers (rows x columns);
- Reads the first line to an internal list to build the field names of the metadata schema;
- For each row in the source file it builds the correspondent metadata structure in the destination file.

The application can be used with any metadata schema since the field names are obtained dynamically from the document classification file. Those names must map the chosen schema in the *greenstone* collection.

Fig. 1 shows an example of the XML structures created by the application. Among the labels <FileSet> </ FileSet> are placed the descriptions of all document in the collection. Among the labels <Description> </ Description> fields describe the Dublin Core metadata classification of one document. Each field is specified between the tags <Metadata> </ Metadata> showing its name as described in the attribute "name".

VI. Tests

The prototype was tested with the digital collection of this paper. Documents were classified under a Dublin Core metadata subset and imported to a *greenstone* collection with success. First documents were classified in an MsExcel file as done by the librarians. Next, with the Librarian Interface, documents were collected into *greenstone* and a collection was built without information in metadata arguments. The collection was compiled and the correspondent source file, *metadata.xml*, became available. Modifications to the source *metadata.xml* file were accomplished by running the prototype. To associate the new source file to the collection, using the librarian interface, the collection was opened and compiled again. The documents were then presented and indexed, as expected, by *greenstone* interface.

Other tests were carried out based on custom metadata structures confirming that the prototype supports other metadata sets.

There are some assumptions that must be taken in consideration when using the prototype successfully:

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE DirectoryMetadata (View Source for full doctype...)>
<DirectoryMetadata>
 <FileSet>
 <FileName>generating metadata file\.pdf</FileName>
  <Description>
   <Metadata mode="accumulate" name="dc.Title">Generating Metadata File</Metadata>
   <Metadata mode="accumulate" name="dc.Creator">Katherine Don</Metadata>
   <Metadata mode="accumulate" name="dc.Subject">metadata, greenstone, automatic</Metadata>
   <Metadata mode="accumulate" name="dc.Description">Greenstone users mailing list reply</Metadata>
   <Metadata mode="accumulate" name="dc.Publisher">greenstone-users mailing list</Metadata>
   <Metadata mode="accumulate" name="dc.Contributor">University of Waikato, New Zealand</Metadata>
   <Metadata mode="accumulate" name="dc.Date"> 2005-09-20 </Metadata>
   <Metadata mode="accumulate" name="dc.Type"> Text </Metadata>
   <Metadata mode="accumulate" name="dc.Language"> en </Metadata>
   <Metadata mode="accumulate" name="dc.Format"> PDF </Metadata>
   <Metadata mode="accumulate" name="dc.Identifier">generating metadata file.pdf</Metadata>
  </Description>
 </FileSet>
</DirectoryMetadata>
```

Figure 1. XML metadata structure created by the application

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE DirectoryMetadata SYSTEM "C:\ DirectoryMetadata.dtd">
<DirectoryMetadata>

   *** Here goes the FileSet metadata structures ***
</DirectoryMetadata>
```

Figure 2. Heading of *metadata.xml* file

- The initial classification list should be in MS Excel and the first line must represent the fields of the metadata schema in use by the *greenstone*'s collection. Subsequent rows in the list must contain the classification of the various documents.
- In the standard *greenstone* application directory, the file *metadata.xml* must exist, containing, at least, the heading definitions with the initial settings as represented in Fig. 2.
- The system in which the collection is built must provide access to the Internet; otherwise a local DTD (Document Type Definition) named *DirectoryMetadata.dtd* must be created as showed in Fig. 3 and the situation must be adequately described in the file header of the *metadata.xml* file, see Fig. 2.
- The document's name cannot contain periods ("."). When a period is found in a file name the application will assume the value after the period as being the file extension and thus the associated metadata will not be read properly.

VII. FUTURE WORK

Further work is on going to optimize and standardize the actual metadata set of the Library to get them compatible with national repository projects, namely the RCAAP (Repositório Cientifico de Acesso Aberto de Portugal). This and other projects normally are developed according the DRIVER 2.0 standards for metadata organization.

There are still problems in compiling pdf source files. The main source collection is made of pdf files and about 20% of them were not able to be compiled into *greenstone*. The plug-in available in version 2.83 has problems in leading with some characteristics of pdf documents that we haven't identified yet but the problem is under investigation.

```
<!ELEMENT DirectoryMetadata (FileSet*)>
<!ELEMENT FileSet (FileName+,Description)>
<!ELEMENT FileName (#PCDATA)>
<!ELEMENT Description (Metadata*)>
<!ELEMENT Metadata (#PCDATA)>
<!ATTLIST Metadata name CDATA #REQUIRED>
<!ATTLIST Metadata mode (accumulate|override)

                              "override">
```

Figure 3. Document Type Dedinition (DTD)

A Librarian manual will be prepared as demo videos to accelerate the knowledge of the librarian in the process of getting different collections into g*reenstone*. Librarians, as ordinary users, are well capable to use the librarian interface of *greenstone* and the prototype to get collections ready to be published. Preliminary acceptance and classification of digital documents can still be done as before downsizing the impact of the change.

VIII. CONCLUSIONS

Objectives were achieved; a digital library was created and automatically enriched with external metadata already stored in an excel file.

The library process of collecting and classifying the information remains as before. Teachers submit the documents and correspondent preliminary classification in excel, and the librarian completes the classification. Enrichment and compilation of the published digital collection is added to the process as a monthly activity.

This incoming activity will surely be well accepted. G*reenstone* has an easy learning curve and librarian and assistants have high motivation and expectations in producing and updating their own digital collections.

With *greenstone* facility of getting run locally from a CD or Hard-Disk, the solution delivered to the library is expected to return value in short time. Students will get easy access to the documents, librarian and assistants won't be overwhelmed in searching and getting paper documents and the University will broadcast easily its production.

The quick wins achieved by the decision of getting *greenstone* run locally was important to convince librarian and their assistants of the importance in using technology but was also naturally contested. Maintenance and support of those and other technology solutions must be well thought out under a whole enterprise strategy [10]. Most important is, when adopting solutions locally, to have them compatible with the standards so they can scale easily and get easily integrated with internal or external projects and that was taken in consideration.

REFERENCES

[1] D. Bainbridge, I. H. Witten. How to Build a Digital Library. San Francisco: Morgan Kaufmann, 2003

[2] H. Ian, D. B. Witten, "StoneD: A Bridge between Greenstone and D-Space," D-Lib Magazine , vol 11 issue 9, pp. 2-19, September 2005..

[3] C.L. Borgman, "What are digital libraries? Competing visions," in C. L. Borgman, Information Processing and Management, vol. 35, pp. 227-243, Los Angeles, USA: Elsevier Science Ltd, 1999.

[4] M. Lesk. Understanding Digital Libraries. San Francisco: Morgan Kaufmann, 2005.

[5] Comissão das Comunidades Europeias. i2010 : Bibliotecas Digitais. Comunicação da Comissão ao Parlamento Europeu, ao Conselho, ao Comité Económico e Social Europeu e ao Comité das Regiões. Bruxelas: Comissão das Comunidades Europeias, 2005.

[6] J.M. Moxley, "Universities Should Require Electronic Theses And Dissertations," Educause Quarterly, nº 3, 2001.

[7] S. Y. Tam, S. T. Wat, D. M. Kennedy, "An Evaluation of Two Open Source Digital Library Software Systems," University of Hong Kong, China, Mai 2007.

[8] D. Bainbridge, D. McKay, I. H. Witten, Greenstone Digital Library : Developers Guide. University of Waikato, New Zeland: New Zeland Digital Library Project, 2004.

[9] Microsoft, Microsoft Visual Studio Documentation 2008. Microsoft, 2008.

[10] B. Maizlish, R. Handler,. IT Portfolio Management step-by-step. John Wiley & Sons, 2005.