

**CREATION OF A DIGITAL REPOSITORY FOR THE MASTER THESIS OF THE
FACULTY OF JOURNALISM, LIBRARY AND INFORMATION SCIENCE OF THE
OSLO UNIVERSITY COLLEGE**

REPORT OF THE GROUP PROJECT

By:

Akhmad Riza Faizal
Eric Boamah
Kanokporn Nasomtrug
Liliana Melgar
Tsigereda Asfaw

International Master in Digital Library (DILL).
January 16th, 2008

INTRODUCTION

Between December 2007 and January 2008, some students of the DILL program stayed over in Oslo and decided to undertake a group project to learn some aspects of digital libraries. The aim was to enhance their understanding in digital library studies. For this reason, professors Ragnar Nordile and Tor Arne Dahl suggested the creation of a digital library involving all the master thesis of the Faculty of Journalism, Library and Information Science. It was also suggested to the team to use Greenstone as the digital library software.

The project has as objectives:

- To learn how to build a digital library, given a group of digital documents.
- To learn how to apply the software Greenstone for this purpose.
- To learn about the existing metadata and projects to manage ETDs: Electronic Theses and Dissertations.
- To get involved, as a group, in real projects that demand our initiative and cooperation.

The project doesn't pretend to be finished or concluded.

DESCRIPTION OF THE ACTIVITIES

Given the task to build a digital library using the master thesis mentioned, the group decided to do the following activities:

1. Study and read, individually, during December 15th and January 2nd about Greenstone, and about metadata standards existing to assign to ETDs.
2. Receive the materials and review them, taking decisions about how to treat them.
3. Download the Greenstone software and install it both in our own computers and in the Linux server at the HiO.

4. Try to find some main projects around the world that involve the cataloguing of E-theses or Electronic Theses and Dissertations (ETDs).
5. Take decisions about the metadata to use to describe the thesis.
6. Customize the Greenstone software according to the metadata chosen.
7. Upload the documents in Greenstone and describe them.
8. Customize and get the most of our documents in the Greenstone software.

THE SOFTWARE USED TO BUILD THE DIGITAL LIBRARY

There are some options to choose the software to build a digital library, some of them very specific for the purpose of describing ETDs. And although we didn't face the problem of choosing an specific software, because one of the main goals of the project was to know how to use Greenstone, we studied some of the main features of the software available. The most important ones are DSpace and ETD-db, because they are widely used. The options that we found are:

NAME	DEFINITION	DEVELOPERS	WHO USES IT	METADATA	OAI
DIGITOOL	Commercial software that enables academic libraries and library consortia to manage and provide access to digital resources, both those that are created for use within the institution and those that are collected and maintained by the library for the benefit of the public.	Ex Libris Group	? (7)	Customizable, some standards not, like EAD (Encoded Archival Description) (6)	Supported
DSPACE	Open-source platform for accessing, managing, and preserving scholarly works.	MIT Libraries and Hewlett-Packard (HP) Labs	-301 installations -in 50 countries In Norway: -Bergen Open Research Archive -Munin at the University of Tromsø See more in (1)	DSpace supports only the Dublin Core metadata element set with a few qualifications conforming to the library application profile (see DSpace Metadata)	Supported
ETD-db	Open source software. The ETD database is a series of web pages and perl scripts that interact with a MySQL database. These scripts provide a standard interface for web users and	Virginia Tech. Endorsed by the Networked Digital Library of Theses and Dissertations (NDLTD)	"Currently this is the most widespread E-theses package in use, in part due to the support it has from the NDLTD." (2)	A thesis-specific metadata set: ETD-MS standard set by NDLTD. (Based on Dublin Core)	Supported (with some specifications) (2)

	researchers, ETD authors, graduate school personnel, and library personnel to enter and manage the files and metadata related to a collection electronic theses and dissertations.				
EPRINTS	EPrints is an open source software for building repositories of research literature, scientific data, student theses, project reports, multimedia artefacts, teaching materials, scholarly collections, digitised records, exhibitions and performances.	School of Electronics and Computer Science, University of Southampton, UK.	241 known archives are running EPrints worldwide. (3)	Can be configured. The default set is has been designed as part of a collaboration between EPrints and library staff as a starting setup for an institutional archive.	Supported (Green OA Self-Archiving)
FEDORA	Fedora is a Linux-based operating system, free and open source.	Sponsored by Red Hat	It has users and ambassadors in more than 60 countries	Mainly METS, and also DC. See for more information: (4)	Supported
GREENSTONE	Greenstone is a suite of software for building and distributing digital library collections. It is open-source, multilingual software, issued under the terms of the GNU General Public License	New Zealand Digital Library Project at the University of Waikato, and developed and distributed in cooperation with UNESCO and the Human Info NGO.	Aprox. 60 users and developers who work with Greenstone in at least 32 different countries (5)	Any metadata set can be customized. By default: DC.	Supported

INSTALLATION OF THE PROGRAM

To install the Greenstone software at HIO we faced some problems:

Oslo University College server is working under Unix and the PCs in most of the computer rooms are under Windows XP. We have to set how to connect those two platforms for the greenstone. For installing the greenstone interface, we followed these steps to install the program in the Unix server. We summarized here the steps recommended by Tor Arne Dahl and Kjetil Iseli:

1. Use the ssh program to connect to the server. You can start ssh by choosing Start -> Programs -> Internet -> Telnet and ssh in Microsoft Windows. You can login using these parameters:

- host: bibin.hio.no
- username: gsdI
- password: gr3enstone9

2. Unpack the file `gsdl-2.80-unix.tar.gz`, which is located in the home directory. After the software is unpacked, you must change directory to `gsdl-2.80-unix/Unix/`. Run the install script with the command: `./Install.sh`

3. During installation, Greenstone will ask you some questions. The questions (Q) and the answers (A) you must provide, follow in the dialogue below. Just press Enter on the questions not mentioned in the dialogue (this means you should use the default answers).

Q: Enter directory to install Greenstone into. A gsdI directory will be created in this directory. [/home/gsdI]
A: /var/www/bibin

Q: Greenstone needs a valid cgi executable directory (normally called cgi-bin on unix systems) from which to run.
[...]
Enter "[1]" or "2"
A: 2

Q: Enter existing cgi executable directory [/usr/local/apache/cgi-bin]
A: /usr/lib/cgi-bin

Q: Please enter the web address of the /usr/lib/cgi-bin directory.
A: <http://bibin.hio.no/cgi-bin>

Q: In order for Greenstone to run, the /var/www/bibin/gsdI directory and all it contains must be accessible from the web.
A: 1

Q: Enter the web address of the /var/www/bibin/gsdI directory.
A: <http://bibin.hio.no/gsdI>

4. You will then be asked to choose a password for the administrator. We recommend that you choose the same as the password to the ssh account (gr3enstone9).

5. After installation you can access Greenstone from the <URL: <http://bibin.hio.no/gsdI>>.

library path: <http://bibin.hio.no/cgi-bin/library>
greenstone gliserver url: <http://bibin.hio.no/gsdI/cgi-bin/gliserver.pl>

gliserver.pl is not under cgi-bin, but under gsdI/cgi-bin

The installation run well, but we can not open The Greenstone Librarian Interface from PC in computer room because the PC run under windows XP. We have consulted this problems to Kjetil Iseli, and he suggested us to install the interface in each CPU we

used in S558. For installation we followed:

1. Download GLI:
<http://prdownloads.sourceforge.net/greenstone/gli-client-2.80.zip>
2. Unpack to H:\
3. Edit gli-client.bat in notepad
4. On the top, after "set GLILANG=en" line, enter the following:
set JAVAHOME=C:\progra~1\java\jre16~1.0_0\
5. Save the file
6. Go to start menu, run: cmd.exe
7. type H: <enter> then: cd gli-client-2.80 <enter>
8. type gli-client.bat and hopefully everything works.
9. Put these paths as value, library path: <http://bibin.hio.no/cgi-bin/library>
greenstone gliserver url: <http://bibin.hio.no/gsd/cgi-bin/gliserver.pl>
10. Enter the interface by submitting:

username: admin
password: gr3en

The greenstone librarian interface appeared, but we have to encounter another problem, only "gather" and "design" features of the interface were working, rest of the features were off. The solution for this problem was that we should continue our work on local installation, using c:\program files\greenstone\collect\ to access to collection and K:\ to share files/collections to share the collections between us. First, we assumed that we could install the interface in c:\ but we could not, because it is unwritable. So we decided to work on our own laptops because some of them (with XP) allowed installing the interface and could be accessed perfectly.

PREPARATION OF THE DOCUMENTS

The cds and floppy disks delivered by Tor A.D., had different contents. Some of them included the thesis in word files separated by chapters, others included the thesis report plus some kind of database or program; and some others only included software.

We decided to do the following:

1. To put the separate word files into a single document, to make it easier for the user who retrieves a thesis, and to keep it simple to assign the metadata. Later

on we discovered that Greenstone (the Demo collection) presents the documents separated by chapters and that they put all the separated files into one folder that is described by the metadata, which is assumed by default by all the separated files. But we kept them in a single document to manage it in an easier way.

2. We also decided to use PDF for the documents in the digital library instead of keeping them in Microsoft Word format (.doc) because:
 - Word changes versions constantly, we want it to be compatible with all computers
 - PDF is good for intellectual property preservation
 - For changing file format from Microsoft Word to PDF, we used PDFCreator which can be extract from G:\distro\PDFCreator.
3. We kept the pdf files into one folder that was imported into Greenstone and described with the metadata.
4. We kept the files with software, or with software plus report, in a different folder. Each subfolder was zipped to be described in Greenstone, or attached to the report (in the case when the sotware were annexes of the document).

METADATA

There are some options to choose metadata for describing ETDs:

- Dublin Core: used by the majority of the projects of ETDs' cataloguing.
- MARCXML or MODS: when used, it is necessary to use a cross walk to Dublin Core (8).
- Specific metadata sets: some projects use specific sets. Depending on the project, these sets are important to communication and exchange of information between the members. The most important one that we found is the NDLTD metadata set, based on Dublin Core (9) (10), and also promoted by UNESCO (11) and some other important institutions (12).

In Europe we found a project to catalog doctoral thesis, which uses Dublin Core with some minor adaptations (13).

In the group there was also some suggestion to follow the recommendations of the Universal Catalog (14), but we found that it is not appropriate for ETDs because it's not applied in that specific field.

In the greenstone librarian interface we used then the NDLTD Dublin core elements. The following metadata fields were used:

- 1) Title (dc.title).
- 2) Author (dc.creator).
- 3) Subject and keywords (dc.subject). Some of them were extracted from Bybsis.
- 4) Abstract (dc.description). In this case, abstract of the thesis and dissertations be describe as description of the files.
- 5) Publisher (dc.publisher).
- 6) Date/year (dc.date).
- 7) Type (dc.type). We used "Electronic Thesis and Dissertation" as

- recommended by NDLTD.
- 8) Format (dc.format). We used "application/pdf" as recommended by NDLTD (MIME controlled vocabulary), and "application/zip" for the software compressed.
 - 9) Language (dc.language).
 - 10) Thesis degree name (thesis.degree.name). We used "Master in LIS" as recommended by one teacher at HIO.
 - 11) Thesis level (thesis.degree.level).
 - 12) We added dc.relation to relating software files with pdf files under the name "Is Part Of" (dc.relation.ispartof).

CATALOGUING

After we set the metadata and uploaded the documents, the next thing we did was to create the indexes for searching and browsing facility in Greenstone. For search indexes, we focused on using Title (dc.title), Author (dc.creator), Subject and Keywords (dc.subject), Abstract (dc.description), Year (dc.date), Publisher (dc.publisher), and Thesis level (thesis.degree.level). For browsing indexes, we used AZCompact List from Greenstone facility with Title (dc.title), Author (dc.creator), and Subject and Keywords (dc.subject) for simplicity reasons.

After finishing those settings, we put the metadata to the files, specially to those consisting only in the pdf file.

Some decisions:

We considered the language to use to assign metadata. English or Norwegian were the options. The criteria were that, if this digital library was going to be in an international network (like NDLTD), or just to use locally; also we consider the recommendations of the standard adopted (NDLTD Dublin Core). We decided:

- To use Norwegian for names of institutions, titles, subject of files, and description (abstract).
- To use English for the field "DC. Type", because it was suggested as a string: "

For subjects: we copy them from Bybasis, in Norwegian. Some of them did not have subjects at all, and some of them had the Dewey classification number. We thought to search what subject corresponded to that number, but finally we decided to leave this for further work.

See this link to observe how they use DC in Bergen:

https://bora.uib.no/handle/1956/1822?mode=full&submit_simple>Show+full+item+record

Unfilled metadata: Due to time and problems in lack of knowledge of the software, and to some difficulties to read in Norwegian, the following fields were not filled:

- dc.contributor: it is supposed to be filled with the advisors. This data is not in the theses.

- dc.identifier: Greenstone assigns it automatically, in a tag which namespace is ex. (extracted metadata). This identifier is supposed to be assigned correctly when the digital library is online or in the server.
- dc.coverage: needs Norwegian to be understood.
- dc.rights: to be defined by the University.
- dc.relation: We were planning to use this element to connect the zip folders (mentioned before) which contain the software of the thesis, with the report in pdf. However, due to some difficulties faced with the program to do that, because we couldn't study it in depth, we couldn't finish this task. That's the description of what we did and what could be developed more:
 - a. Greenstone has a tab called "Format" where many things can be set. One of them is the displaying of the metadata, in the property called "DocumentText". Under this property some code could be written to decide which elements of the metadata want to be displayed. Some html needs to be written. We did it but it could be developed more.
 - b. In those properties we also have "RelatedDocuments", which we were trying to use to connect the zip file with the pdf. The code that we wrote is in the Greenstone files that we delivered, but it is not working right now, some more things need to be fixed.

Another problem that is present as the Greenstone folder is delivered, is that not all the documents "gathered" to the software and "enriched" with metadata, are displayed in the final "preview collection", especially those that are zipped. This is because of the plugins needed (in the "design" tab) for the system to recognize not only the zip file but the files inside it. One or two of the pdf files are displayed also with wrong characters.

RECOMMENDATIONS

1. Greenstone has a simple interface either in user or librarian interface. It is software easy to use, with some more advanced things that can be worked or studied more (as all the possibilities included in the "Format" tab of the Librarian interface). We couldn't study it in depth, and some more things need to be done in order to gain a better profit of the repository. Some good practices to follow: (17) (18).
2. For better work on electronic thesis and dissertation (ETD) in the future, we recommend to use Dspace as digital library software because it's more widely use and has better -also stable- documentation management than greenstone (19). For moving collections from Greenstone to Dspace see (20). It is also possible to use the "export" feature in Greenstone to sort the digital library in other kind of standards and formats.
3. Other projects that need to be taken into account are for example "DiVA", the Academic Archive Online (Digitala Vetenskapliga Arkivet). It is been using by some of scandinavian institution. See: <http://www.diva->

portal.org/ntnu/index.xsql?lang=en

4. For efficiency, HiO needs to develop policies for the students to know how to submit the theses. Some examples are in (15) and (16).

NOTE: We need to clarify that this is a very incomplete project, but we are also aware of the importance for HiO to work more on this starting to create a real digital library of ETDs available for HiO community, for Norway and Europe.

REFERENCES

- 1) <http://wiki.dspace.org/index.php/DspaceInstances>
- 2) <http://www.ariadne.ac.uk/issue38/jones/>
- 3) <http://www.eprints.org/software/archives/>
- 4) <http://wiki.dlib.indiana.edu/confluence/display/INF/Fedora+Metadata+Storage+Philosophy>
- 5) <http://www.ils.unc.edu/~sheble/greenstone/survey-report.html>
- 6) <http://www.loc.gov/standards/premis/tools.html>
- 7) <http://john.curtin.edu.au/aboutus/papers/kjh-lp-icau2003.html>
- 8) <http://www.exlibrisgroup.com/category/DigiToolOverview>
- 9) <http://metallogger.wordpress.com/2007/08/10/recommended-australian-digital-thesis-metadata/>
- 10) <http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html>
- 11) <http://www.etdguide.org/>
- 12) <http://www.ariadne.ac.uk/issue32/theses-dissertations/>
- 13) http://www.surffoundation.nl/download/ETD_LessonsLearned_Full-Report+Annex.pdf
- 14) <http://c2.com/cgi/wiki?UniversalCatalog>
- 15) <http://etd.vt.edu/submit/tutorials/submit/step01.html>
- 16) <http://etd.library.vanderbilt.edu/ETD-db/howsubmit.html>
- 17) <http://ibdigital.uib.es/gsd/cgi-bin/library>
- 18) <http://www.greenstone.org/examples>
- 19) <http://www.ariadne.ac.uk/issue38/jones/>
- 20) http://wiki.greenstone.org/wiki/gsdoc/tutorial/en/greenstone_to_dspace.htm

This file also can be downloading from:
<http://ahmadriza.wordpress.com/>