INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

E-books Digitization using Open Source Software – Lessons from Haramaya University Library Digitization Project, Ethiopia

Milkyas Hailu
Director, Department of Library and Information Services, Haramaya University, Ethiopia

Abstract:
e-books digitalization project at Haramaya University Library and Information Services has created 17,456 E- books and 300 local contents in which manually derived descriptors for the scanned books are used for indexing and segmenting the library contents using Greenstone Digital Library Software. During implementation, many lessons were learned on both technical and administration aspects of the project. This paper highlights some of the author's pertinent experiences during the implementation of the project.

Keywords: *Digitalization, Open Source Software, Greenstone, Ethiopia*

1. Introduction

1.1. Background

Ethiopia owns more than 1,700-year tradition of elite education linked to the Orthodox Church. But the advent of modern education in Ethiopia went back only to the turn of the 20th century. Adissababa University was founded in 1950 and two years later the then Imperial Ethiopia College of Agriculture and Mechanical Arts which is presently known as Haramaya University, was established (Saint 2004) .

Haramaya University's long year academic exercises and research practices in Ethiopia were basically aimed at addressing the pressing socio-economic problems of the country particularly, in agriculture. The then college's academic, research and extension programs were conducted with profound assistance of staff and material from the US-Oklahoma State University particularly between 1952 and 1968 (Arts 1957).

US-Oklahoma State University helped in the commencement of the four- year Bachelor of Science in Agriculture. The BSc program was followed by tremendous hands–on agricultural research undertakings as reflected in the existing published researches and reports documents from the period. However, most of these rare research documents, books and publications are not properly preserved and made accessible for scholarly research. Lack of concern and physical space contributed for the poor conditions of these resources. Hence, the project aimed at filling this gap by digitalizing these important resources.

In addition, the University Library and Information Services (LIS) progressively collected thousands of e-books from various sources and scanned its few popular books in order to commence digital content provision service since 2008. However, the initial digital content provision service was complex to use, the content was stored in the Library FTP Server and accessed from users/clients computer, where full text search of the materials was impossible and there was no title /subject browsing facilitates. Hence, it has become necessary to upgrade the digital content provision services of the LIS by allowing users to access full text searchable and accessible web based Library services. In addition, the availability and accessibility of various open access software's for digitalization purposes created an opportunity to embark this digitalization project.

1.2. Digitization

Digitization – whether "mass" or just "large-scale" – of books and other materials is not new. Whether for the purpose of preserving them for future generations or making them available to a much wider audience than could ever access the physical objects, some libraries, archives, museums and publishers have been scanning their older documents and pictures for many years. Thousands of libraries of all sizes have scanned images, cataloged them and made them available on the Web (Hahn 2006).

The expansions of ICT infrastructure particularly the Internet in Haramaya University motivated the LIS to initiate this project. There were two major ICT expansion projects in the University, introduction of broadband Internet access in mid to late 2000s (Mammo 2010) and two major Local Area Network expansion projects implemented between 2005 and 2012. The projects have significantly

increased number of Internet access points in the university particularly in offices, staff residences, computer laboratories, libraries even wireless zone has been established in recreational areas(VPSA 2014).

The expansion of the ICT infrastructure in the University has also brought new opportunities and threats to the LIS. Firstly, the LIS may explore this opportunity to automate and digitize its services so that more users can get the LIS service virtually. On the flip side, the popularity of ICT particularly, the Internet can pause a threat to the traditional LIS in such way students and faculties will be less incline to use the traditional Library services in line with (Howard 2007), putting the University LIS in less important position in the academics and research affairs of the university.

The implementation of the project aimed at helping the LIS to deliver web based access Library services to the University community. The ever increasing price of hard copy books, Stagnant Library budget in the face of increasing number of students, the need to preserve and make accessible local research results are also another contributing factors for commencing the digitalization project. Above all, the need to reclaim the important position of the Library in the face of growing ICT/Internet infrastructures in the University was the major objective of the project. It was envisaged that the execution of this project would help to increase the visibility of the university Library through adding values to the existing LIS services.

The Digitalization project took 11 months, from February 20, 2014 – January 20, 2015. During the project period, 17,456 full text books and 300 local research contents were digitalized using Greenstone Digital Library Software.

2. Digitization Process

The project followed three phases namely, Planning, Digitalization process and post digitalization process. The three phases followed in the project are summarized on the table hereunder.

Steps	Key Activities
Planning	• Identified major collection to be digitized in agriculture and other colleges of the University • Assessed the required Human and Physical resources to populate and digitalize the materials • Decided on the metadata standards. • Defined methods and timing of quality control.
Digitalization Process	• Availed required equipment • Team Formed (5-calatogers were trained on data enriching) • Scan books and documents with quality control • Configuring Greenstone for individual catalogers (collection creation, plugin, populating materials) • Selected materials to be digitized. • Classified books using Library of Congress Classification scheme. • Quality control of the materials (conditions of the material including cleanness, file size etc.) • Troubleshooting errors
Post-digitization processes.	• Migrating collection from cataloger's computer to main server • Design user Interface for each collection • Assessment and evaluation of the project. • Official Inauguration

Table 1

Quality control and review process was conducted by a person other than the ones doing the original digitizing work. This is in line with the recommendation from previous literatures (Jenn Riley & Kurt Whitse, 2005).

3. Digitization using Open Source Software (OSS)

The term Open Source software refers to a computer program that is distributed under one of a number of licensing arrangements. The OSS scheme requires the specific software's source code to be availed and accessed. It also permits the receiver of the program to modify the source code freely and align it with their own needs provided that, if they distribute the software modifications they create, they do so under an open source license (Myeza 2010). The open source model helped to develop software's in many industries and business. It also offers librarians, the capability to create the software that is standards compliant, interoperable, extensible and scalable. Many of this software's help customers to find information quickly, conveniently, no matter where that information resides, it also gives the freedom to use, change or distribute the way you want(Hasan 2009).

Open source software has been commonly used for building digital libraries in many academic and research institutions(Oak et al. 2013). It presents a system for the construction and presentation of information collections. More importantly, it helps in building collections with searching and metadata-bases browsing facilities and easily maintained, augmented and rebuilt automatically (Tramboo 2012).

3.1. Reasons for selecting Open Source Software

(Castagné 2013) outlined five aspects of open source software as the major reason for selecting open source software as political, economical, social, technical and legal. On the Political aspect, Open source software's allow everyone to use, study, modify and distribute the software, regardless of person's status, wealth, social background etc (Vimal 1998). From the Economical perspective, upfront costs of commercial software's is not included in the open source software cost users get full version of the product, no time limited trials(Pankaja & Raj 2013). Commercial software's require huge investment at the initial stage and additional payment needs if the user wishes to update the software and users have no ownership on the software, it only allows work with the application.

Open source software has also social and technical influence(Singh & Phelps 2009). The social influence is vivid in its development and maintenance process where people collaborate and work together, anybody can also join and contribute in the group. Through this, open source software projects encourage innovation and collaboration of community members. On the technical perspective, it is becoming more reliable and it has proved that open source software is increasingly competitive from its proprietary software counterpart(August et al. 2013). In addition, open source software is interoperable, customizable according to the needs and fulfills the software industry standards. From legal point of view, open source software licenses are copyright protected, they strictly ensure the users freedom to use, modify and distribute the programs.

3.2. Eprint

EPrints is a free and open source software package originally developed by researchers at the University of Southampton School of Electronics and Computer Science in 2000 (making it the oldest of the platforms in this report)(Andersson, S., Svensson 2013). It was designed specifically for archiving research papers, theses and teaching materials, though it can accept any content. As of July 2013, ROAR has recorded 500 implementations, making it the second most popular platform.

Like DSpace, EPrints follows a "turnkey" approach and some institutions have reported that the installation process is fairly straightforward (Beazley, 2010). The administrative backend provides access to configuration options. EPrints' Bazaar Store is an interesting concept, aiming to allow repository managers to install extensions with a single click. EPrints is capable of using a controlled vocabulary and authority lists, which can help ensure high metadata quality. It provides native support for Dublin Core with the possibility of exporting to a number of formats (e.g., METS, MODS and DIDL). Qualified Dublin Core and MARC are not supported(Castagné 2013).

3.3. Dspace

DSpace was jointly developed by Massachusetts Institute of Technology and Hewlett-packard in 2002, is an open source software which can be freely downloaded (Castagné 2013)(Andersson, S., Svensson 2013). DSpace is the software of choice for academic, non-profit, and commercial organizations building open digital repositories. It is free and customizable to fit the needs of any organization(Castagné 2013). DSpace preserves and enables easy and open access to all types of digital contents including text, images, moving images, mpegs and datasets. With an ever-growing community of developers committed to continuously expanding and improving the software, each DSpace installation benefits from the next (Biradar & Banateppanavar 2013).

3.4. Greenstone Digital Library Software

Greenstone is an open source digital library building and distributing software. It is a popular software particularly, the software's ability to redistribute digital collections on self-installing DVD/CD-ROMs made it more popular in developing countries such as Ethiopia with very limited Internet bandwidth (Ian H et al., 2000).

Greenstone is designed and developed using Java, making it ideal for many platforms including Linux and Windows. The current version, Greenstone3, is redesigned to improve the dynamic nature of the Greenstone toolkit and decrease the potential overhead incurred by collection developers. Moreover, it is distributed and can thus be spread across different servers. Furthermore, the new architecture is module based, using independent agent modules that communicate using single message calls.

Greenstone uses XML to encode resource metadata records. XLinks are used to represent relationships between other documents. Using this strategy, resources and documents are retrievable through XML communication. Furthermore, indexing documents enables effective searching and browsing of resources. The software operates within an Apache Tomcat Servlet Engine. This projects compared E-print, Dspace and Greenstone Digital Library Software and selected Greenstone based on evaluation criteria.

4. Lessons

4.1. Enabling Conditions

It is learned that successful digitization project requires considering various aspect of the project during the planning and implementation phases. These essential aspects of the project can be summarized in to four major categories. These are availability of Technical Infrastructure, Physical Infrastructure, Organizational Infrastructure and the knowledge Infrastructure as demonstrated in the figure below

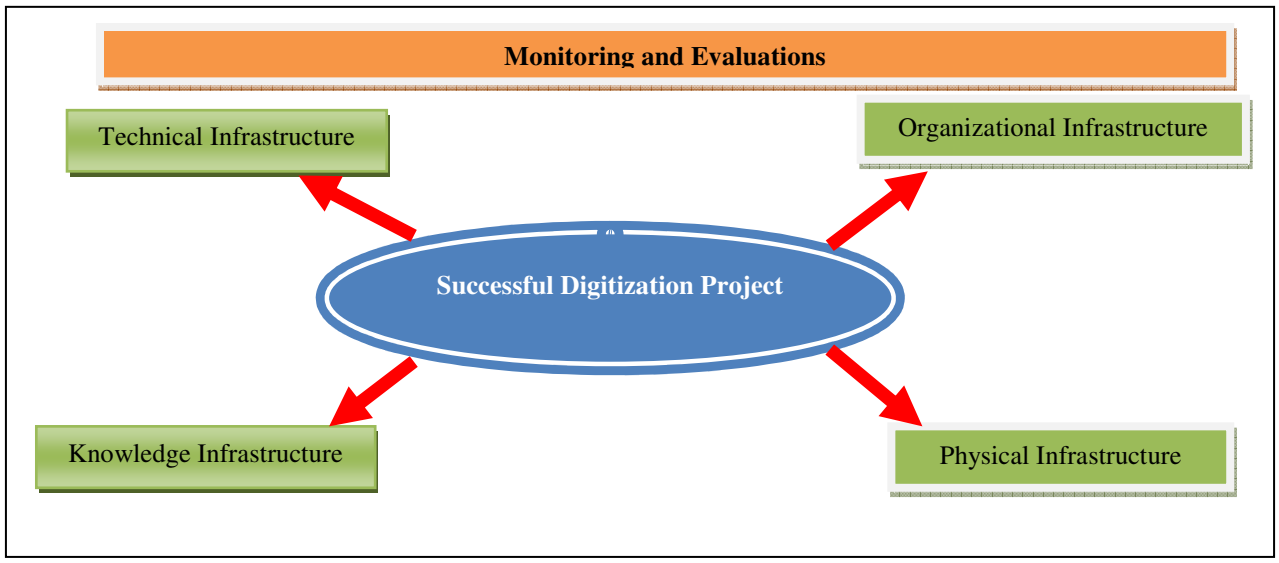

Figure 1: Library Digitization enabling environment framework

The Technical Infrastructures are the human resource particularly, technical skills required for digitalization software selection, installation, configuration, troubleshooting and training. This IT skill is very important in light of the technical difficulties in customizing and deploying most open source software's. Hence, prior to initiating such digitalization project one has to ensure the availability of technical infrastructure i.e. technical staff trained to do software selection, installation, configuration, troubleshooting and training considering the local context.

The Physical Infrastructure include the necessary Electric Power supply, Server computer, Desktops, Scanners, network infrastructure and physical space required to implement the project. The required size of Physical Infrastructure correlates with the size of the collection and the quantities of the physical infrastructure should be aligned with the time and budget of the project.

The organizational infrastructure encompasses the availability of rules and regulations, incentives, commitment of management, support from the University top management. For example, the adoption of institutional open access policy will allow making the local content of the university accessible to the public. This is similar with findings of (Biradar & Banateppanavar 2013).

The knowledge infrastructure concerns with the availability of content. It also deals with ensuring relevant documents are selected, the depth and width of the entire collection. Successful knowledge infrastructure incorporates contents that are relevant to the targeted users.

4.2 Open Access software Evaluation Criteria Framework

Evaluation of open source software is different from proprietary programs (Vimal 1998). There are dozens of OA software's that can be used for the digitization. Hence, there should be a set of important criteria to select the best digitization software. To this end, this project proposed a software selection framework as demonstrated in the table hereunder. It should be noted that prior to applying the evaluation criteria proposed (table 2), the existing technical, physical, and organizational and knowledge infrastructure were assessed. These helped to choose the correct software that suit our requirements. Following the results of the evaluation, Greenstone Digital Library software was chosen for undertaking the project.

Evaluation Criteria	Remarks	Mark/10	Priority (1 – 5)
Operating Systems and Hardware Configuration	e.g. Dspace is stable on Linux only not on windows e.g. We lack the technical staff to work on Linux	5	5
User friendliness and Documentation During Installation	It is complex than greenstone particularly, configuring tomcat	10	5
Larger User Community	e.g. Yes	10	4
Limitations	e.g. Dspace can do all the required modules	10	5
Input and Maintenance of Data			5
Indexing of stored Information			3
Retrieval of Stored Information			2
Output of Data			5
Ease of use			5
Tried and Tested Features			4
Good Documentation for users			1
Built-in Routines			2
Compatibility			4
Performance			5
Flexibility			4
Total = Σ Marks * Priority			

Table 2: Sample Software evaluation criteria using Dspace

4.3. Technical Lessons

Five key lessons during the Technical implementation of the project are outlined hereunder.

- Working with large size PDF collection in Greenstone takes very long time for building and indexing, building. Particularly, if the work is done from a single Server the building process becomes quite slow importing and building 5000 e-books took 51 hours. As remedy, we learnt that dividing the collection in to smaller sizes saves a lot of time. Accordingly, the 5000 collection were divided into five sub collection. Catalogers were assigned to enrich and build each sub collection (1000 E-books each) using the standard Metadata defined, it only took 33 hours to import and build the whole collection. Later, the sub collections were put together by simply copying from Greenstone/collect folder of the sub collections in to the main server Greenstone/collect folder.
- Working with PDF files in Greenstone requires special program to convert the PDF document in to text so that it can later be used for searching purpose. In this project, we used an extension called PDF-Box (trac.greenstone.org/browser/gs2-extensions/pdf-box/trunk/pdf-box-java.tar.gz) extracted and pasted it in to the greenstone/ext folder and restarted the server. It is also important to configure the Pdf plugin in the Greenstone Design tab and check on the PDF-Box extension.

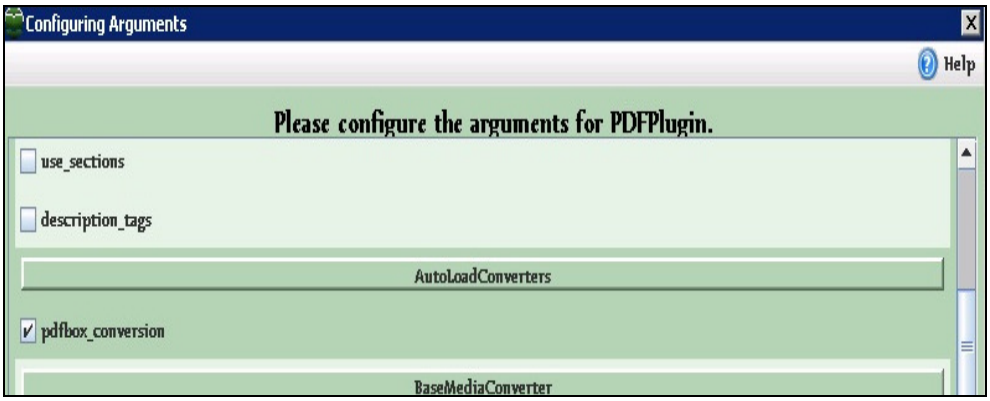

Figure 2: Enabling PDF-Box extension

- Many of the PDF files in our collections were failed to be processed or rejected during the building process. We learned that PDF security settings play a crucial role here. Hence, all security setting should be removed before adding new PDF documents in the collections. Nitro Professional software were used in this project.
- Suddenly, metadata information in the enrich panel could be empty. As a result, our catalogers were forced to redo the enriching process. In the process, we learned that this problem is a result of corrupted **metadata.xml** file in the collect/collection name/import folder. This can be fixed by opening metadata.xml with xml editor and make sure that the first

line begins with the code hereunder and fix any inconsistent data element in the metdata tags.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPEDirectoryMetadata SYSTEM "http://greenstone.org/dtd/DirectoryMetadata/1.0/DirectoryMetadata.dtd">
```

- The other key lesson learned is Enabling PDF full text search were very difficult. It requires special formatting of the collection. In the greenstone Format Tab, select search Vlist and paste the following code.

```
<td valign="top">[link][icon][/link]</td>
<td valign="top">{If}{[ex.FileFormat] eq 'PDF', <a
href=\"_httpcollection_/index/assoc/[archivedir]/[srclinkFile]#search=&quot;_query
terms_&quot;\">{Or}{[ex.thumbicon],[ex.srcicon]}</a>,
[ex.srclink]{Or}{[ex.thumbicon],[ex.srcicon]}[ex./srclink]}</td>
<td valign="top">[highlight]
{Or}{[dc.Title],[ex.Title],Untitled}
[/highlight]{If}{[ex.Source],<br><i>([ex.Source])</i>}</td>
```

5. Conclusion

Digitization is a complex process with many crucial dependencies between different stages over time. Particularly, building large size digital collections using greenstone digital library software requires dedicated team members, physical resources and more importantly, understanding the nature of the software and devise techniques to effectively utilize it. After the completion of this project, major lessons were learnt.

The main lessons identified in this project include minimizing importing and building time in large size collection, PDF to text conversion techniques, fixing metadata problems and enabling effective PDF full text searching. Future same or similar projects can learn from the experience of the Digitalization project in Haramaya University.

6. References

i. Andersson, S., Svensson, A., 2013. Repositories Recreated – The Finch Report versus DiVA in Sweden. , 33(2), pp.183–189.
ii. Arts, C. of A. and M., 1957. Annual Report,
iii. August, T., Shin, H. & Tunca, T.I., 2013. Licensing and competition for services in open source software. Information Systems Research, 24(4), pp.1068–1086.
iv. Biradar, B.S. & Banateppanavar, K., 2013. Steps for developing digital repository using DSpace: An experience of Kuvempu University, India. DESIDOC Journal of Library and Information Technology, 33(6), pp.474–479.
v. Castagné, M., 2013. Institutional repository software comparison: DSpace, EPrints, Digital Commons, Islandora and Hydra. University of British Columbia, (July).
vi. Hahn, T.B., 2006. Impacts of Mass Digitization Projects on Libraries and Information Policy. Available at: http://www.asis.org/Bulletin/Oct-06/hahn.html.
vii. Hasan, N., 2009. Issues and Challenges in Open Source Software Environment with Special Reference to India. , pp.266–271.
viii. Howard, R.M., 2007. Understanding “Internet plagiarism.” Computers and Composition, 24(1), pp.3–15.
ix. Mammo, Y., 2010. Haramaya University Library and Information Services: Looking back to look forward. The International Information and Library Review, 42, pp.14–26.
x. Myeza, J., 2010. A PRACTICAL GUIDE TO DIGITIZING A COLLECTION USING OPEN SOURCE SOFTWARE: A SOUTHERN AFRICAN PERSPECTIVE Joyce. , pp.210–222.
xi. Oak, M., Grade, A. & Pune, I., 2013. WORKING WITH D-SPACE : SETTING UP A DIGITAL INSTITUTIONAL. , 3(2), pp.1–9.
xii. Pankaja, N. & Raj, P.M., 2013. Proprietary software versus open source software for education. American Journal of Engineering Research, (07), pp.124–130. Available at: http://www.ajer.org/papers/v2(7)/O027124130.pdf.
xiii. Saint, 2004. Higher education in Ethiopia: the vision and its challenges. JHEA/RESA, 2(3), pp.83–113.
xiv. Singh, P.V. & Phelps, C., 2009. Determinants of Open Source Software License Choice : A Social Influence Perspective. Social Science Research Network, (2005), pp.1–39. Available at: http://ssrn.com/paper=1436153.
xv. Tramboo, S., 2012. A Study on the Open Source Digital Library Software ’ s : Special Reference to DSpace , EPrints and Greenstone. , 59(16), pp.1–9.
xvi. Vimal, V., 1998. Selection and Management of Open Source Software in Libraries .
xvii. VPSA, 2014. Haramaya University Vice President for Students Affair and Administration,