SHORT COMMUNICATIONS

A PRACTICAL GUIDE TO DIGITIZING A COLLECTION USING OPEN SOURCE SOFTWARE: A SOUTHERN AFRICAN PERSPECTIVE

Joyce Myeza
E mail: joyce.myeza@simmons.edu
Boston, Massachusetts (USA)

Received: 19 May 2009
Revised: 15 December 2009
Accepted: 13 January 2010

Abstract

The article provides an overview of the practical implementation of a digital library using open source software. Southern Africa has not fully embraced or incorporated open source software into their information management operations. This lack of adaptation is attributed to a number of reasons amongst which are lack of general awareness and the absence of appropriately trained librarians to take advantage of such technological sources. The article gives guidelines and recommendations on what to consider when planning to digitize a collection. The following issues will be looked at: digital rights management, institutional repositories, Metadata Encoding and Transmission Standard (METS is an XML Schema designed for the purpose of creating XML document instances that express the hierarchical structure of digital library objects) and its applications, the open archives initiatives, and open source software for digital libraries. The article also focuses on the practical steps in using open source software for digitizing. Open source software has been chosen because of its free availability.

Keywords: Digital libraries, Metadata, Open source software, Project management, Royal archives

Introduction

“Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities” (Schwartz 2006: 385). Digital libraries are the technological processes of improving access to and preserving data, information, images, and records for present and future generations. The libraries of the world are faced with problems such as space for keeping materials and decomposition of sources. The concept of digitization attempts to solve most of these problems. The advantage of digitizing is that of saving physical space, allowing remote access to information that was previously not reachable and making the presentation of the information more appealing to users. The opportunity to do the whole project using open source software allows for good hands-on experience for people who do not necessarily possess specialist information technology skills.

According to Chiware (2007) one of the challenges facing digital library projects in Africa has been the readiness of the university libraries in terms of skills and knowledge to implement the digital and electronic library services. There are many other challenges regarding funding, IT infrastructure, and Internet connectivity, lack of commitment from staff and management and the availability of African generated content to put into the digital collections. The creation and maintenance of digital libraries is not a once off event but a continuous process just like the maintenance of any other collection.

Open source software

Morgan (2008) defines open source software as the software that is distributed under one of a number of licensing arrangements that require that the software’s source code be made available and accessible as part of the package and permit the acquirer of the software to modify the code freely to fit their own needs provided that, if they distribute the software modifications they create, they do so under an open source licence. If these basic elements are met, there is no requirement that the resulting software be distributed at no cost or non-commercially.

Advantages of using open source software

Libraries should consider open source alternatives for many reasons, the primary one being that it eliminates dependence on a vendor to fix a bug or implement a feature. Even if a library does not have the in-house expertise to do the development themselves, because the underlying code is available, it may be possible to hire a consultant or another vendor to extend the product in needed ways. Most importantly, it gives libraries more control over their software in a way they have relinquished to their vendors. Another reason to consider an open source alternative for library systems is that the cost of acquiring and implementing open source solutions may be less than traditional proprietary software. There are still costs associated with open source software. Providing the hardware necessary to run the product and investment in the staff needed to support an open source system, while still costing the library money, may be much less than the total cost of running a proprietary software system (Wrosch 2007). The following are the available types of open source software.

Loose package

The loose package open source system is the one that allows the implementer to assemble the digital library from scratch, starting with an operating system, a web server where information reside, and a relational database which holds the records and programming language to create communication between the tables and clients. *Linux* is a computer operating system, which is freely distributed under the GNU General Public License. This gives any home, corporate, academic, and Government user the ability to modify the technology to his or her needs and to contribute to the ongoing development of the technology (Pitts 1999: 2). The advantage of using Linux is that of the availability of complete source code. *Apache* is an HTTP web server developed by a group of volunteers. Apache is free and can be obtained from its official website (http://httpd.apache.org). Several volunteers collaborated and wrote the source code for Apache (Appu 2002: 28-9).

According to Vaswani (2005: 11) *MySQL* is a multi-user relational database management system that is the de facto standard for database driven software applications, both on and off the web. It was designed for speed, stability and ease of use, and is freely available under the GNU General Public License. *PHP* is a programming language that makes it possible to incorporate sophisticated business logic into otherwise static websites. It provides support for different database systems. PHP source code is available free of charge on the web (Vaswani 2005: 7).

Complete package

DSpace is an open source software solution for accessing, managing, and preserving scholarly works in a digital archive. It was jointly developed by HP and the MIT Libraries in 2002. HP and the MIT Libraries began developing DSpace after MIT expressed the need for a robust software platform to digitally store its collections and valuable research data, which had previously existed only in hard copies. DSpace is a community-based open source platform capable of permanently storing data in a nonproprietary format so researchers can access its contents for decades to come. Because the archive is Internet-based, DSpace can be accessed from anywhere in the world via an Internet connection and federated with other archives (Robinson 2007). According to Barton (2005) DSpace consists of the following features:

License: DSpace is licensed under BSD distribution license.

Format: DSpace uses text, images, audio and video formats.

Metadata: DSpace Search and retrieval is based on qualified Dublin Core metadata.

Standard: DSpace comply with Open Archive Initiative and allows metadata in any format expressed in an XML Schema to be shared.

Platform: DSpace runs on Unix or Linux operating systems.

User interface: DSpace uses Web interface.

EPrints

Barton (2005) describes *EPrints* as free software which creates online archives. It was developed by the University of Southampton, UK and supports self-archiving of e-prints. It can be configured as an institutional repository. It has the following features:

License: EPrints is licensed under GNU General Public License.

Interface: EPrints uses Web interface.

Standards: EPrints complies with Open Archives Initiative standards.

Platform: EPrints runs on Unix or Linux operating systems.

Metadata: EPrints search and retrieval is based on metadata.

Greenstone

Witten (2008) defines *Greenstone* as a suite of software for building and distributing digital library collections. It provides a way of organizing information and publishing it on the Internet in the form of a fully searchable, metadata-driven collection. It has been used to create fully searchable and browsable collections of all kinds of documents, books, photographs, newspaper images, metadata such as library catalogues of MARC records, audio and video. The New Zealand Digital Library project began 13 years ago. It has been developed and distributed in cooperation with UNESCO and the Human Info NGO in Belgium.

License: It is distributed under the GNU General Public License

Platforms: Greenstone runs on all popular operating systems: all Windows versions, Linux, Mac and the iPod.

Interfaces: Greenstone has separate interactive interfaces for readers and librarians. End users access the digital library through the reader interface, which operates within a web browser.

Standards: Greenstone is strongly standards compliant. It incorporates a server that can serve any collection over the Open Archives Protocol for Metadata Harvesting (OAI-PMH), Z39.50 and SRW (Search or Retrieve via the Web). Collections can be exported to METS.

Metadata: The librarian interface includes flexible facilities for adding metadata to documents. Four predefined metadata sets are provided with the software.

Languages: One of Greenstone's unique strengths is its multilingual nature.

The current regional status pertaining to digitization

The exposure to information technology of Southern African libraries is gradually increasing owing to improved access to the internet. The increased exposure has brought about knowledge and awareness of information management tools and practices that are employed by counterparts in the developed world. Libraries in the region, both academic and public have been slow in adopting and or fully exploiting the new technological tools. There are many reasons for this. Coupled with the very high rate of development in technology where newer or updated versions of the tools are released on an almost annual basis means regional library institutions need to institute focused strategic interventions to bring themselves and their nations up to speed. If such a process of adaptation were to be sponsored by executive management and spearheaded by the relevant library schools, the current gap that exists between library services of Southern Africa and the developed world can be reduced to a small margin without requiring large financial inputs.

The current situation of digital libraries: South Africa

KwaZulu-Natal is a province that has the largest human population in South Africa and also ranks as the debatable third most developed in terms of infrastructure facilities and thus academic and public institutions. The province's institutions can thus be used as fair measure of the extent of digitization of the country. An investigation of the extent of digitization conducted in January 2009 showed that the province's highest library institution, the provincial library, is not digitized and there was no digitization project being undertaken at the time. There was also no evidence of digitization in the municipal libraries of the province, both in larger cities and the smaller rural towns. The situation with the province's academic libraries showed the University of KwaZulu Natal to be the only institution that had embarked on a digitization project of some sort. Barton (2005) defines Institutional repositories as a database with a set of services to capture, store, index, preserve and redistribute a university's scholarly research in digital formats. The progress on implementation of open source repositories by South African institutions nationally is growing though. Van Deventer and Pienaar (2008) reported the following institutions that have embarked on some digitisation activities using open source.

Rhodes University uses ePrints
University of Cape Town uses ePrints
University of Pretoria uses DSpace
Stellenbosch University uses DSpace
University of the Western Cape uses DSpace
Durban University of Technology uses DSpace

This shows there is progress in the adoption of open source institutional repository by South African Universities. However there has been a slow growth in the use of open source to create digital libraries. Steps that need to be followed to successfully create digital libraries are discussed in the following section.

Managing of a digitization project that is based on open source software

Precursor to project

When an institution is seeking to create a digital library it is important to motivate staff about the project as its impact would lead to significant change on the operational practices in the library and would also have an impact on customers. Staff motivation about the project should be extended to other strategic departments such as executive management, information technology and communications. Participation should be encouraged from the development of the concept in order to tap into under-utilized capacity and skills that already exist within the pool of employees. Information technology is an important department because it will assist with advice related to software and hardware, and this is usually the most difficult part of assembling a digital library. Another method that might work is to try to find a group of students who are interested to volunteer their skills in exchange for the experience.

The planning phase consists of identification of critical steps and the assignment of responsibilities to individuals who would carry out these steps.

1. Project Management
2. Environmental scanning
3. Budget
4. Digitization hardware procurement
5. Database and web
6. Metadata, OAI and METS
7. Selection of the content
8. Intellectual property
9. Evaluation and usability
10. Marketing
11. Documentation/Photography
12. Quality control

Project Management

The project manager oversees the working together of all the elements that define the digitization program. The tasks include the conception development, drafting of plan, communication, allocation of responsibilities and coordinating the execution of the building blocks of the program in a logical manner. For the project manager to perform the entire task successfully he or she needs to use a project management tool in order to track the progress of the activities. There are a number of open source project management tools available. *OpenProj* is a free, project management solution. OpenProj can be used as a substitute for Microsoft Project and other commercial project solutions. OpenProj is available on Linux, Unix, Mac or Windows (OPENPROJ 2009).

Environmental scanning

Environmental scanning entails researching and collecting information about digital library projects that were successfully implemented elsewhere. This task needs to be performed by an individual with research experience. The aim is to learn about challenges experienced, solutions sought and best practice.

Budgeting

The most important way of reducing costs is to motivate staff to do the project voluntarily and give them incentives like time off, free food etc. Another method that is popular in the United States is that of hiring volunteers to do the project. The incentive for them is experience. Starting a digital library with little or no budget still requires tracking labour costs through the entire process and being able to present the estimate of what the digital library would have cost. Budgeting necessitates preparing a detailed documentation of how much funds will be needed to start the project. It looks at costs for equipment, labour, server, and future maintenance. It is important to document all the processes for future similar projects. The documentation can be used as a referral for writing proposals asking for funding.

Digitization hardware procurement

Digitization requires knowledge of the size of the content that needs to be digitized. This assists in determining the quality and size of the scanner that will be needed to do the project. During this process it is important to consider the storage of the scanner as well as the future projects that might be pursued. If the project is anticipated to be of short duration with no future possibilities of expansion it is advisable to hire or rent the scanner. The next step is to create clear instructions for everyone who will be using the scanner; if possible it is advisable to provide training.

Database and web

Web design entails designing the page which will serve as an interface of the digital library. Something to note here is that some of the digital libraries software come with built-in web interface which makes the work manageable and easier. Database design entails deciding on the open source software to be used for the project. The trick here is that there are many open source software for digital libraries on the web. This task requires a thorough look at the organization’s information technology capabilities for now and future, for example, the server space that is available for the project. It is advisable that this task be done in consultation with the information technology department of the institution. It is important to study the term of use for the software chosen to avoid breach of agreement. Computer hardware entails the number of computers that will be needed to conduct the project; this depends on a number of people that will be needed to create metadata records.

Metadata, OAI and METS

“Metadata is a structured data that facilitate discovery, identification, selection, management and use of information. There are different types of metadata designed to enable different types of interactions with resources. A framework of guidance for building good digital collections categorizes metadata as descriptive, administrative, or structural with further subdivisions of these categories” (Cole and Foulonneau 2007).

The descriptive metadata person decides how the objects of the digital library would be described, what information would be captured about each object and what kinds of subject headings they would have. This person needs to look at the most widely used metadata or classification and cataloguing scheme around the country in order to create some uniformity. The concept metadata is not yet heavily used in Southern Africa. It is very important for this person to work on what is currently used. The preservation entails setting up the standards for archival preservation of images and documents, and develops forms to capture necessary preservation information. Preservation metadata is information that supports and documents activities related to digital preservation.

According to Cole and Foulonneau (2007) the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is designed to enable interoperability between digital libraries and to facilitate the more efficient dissemination of information. OAI-PMH works with structured data, specifically with data expressed using XML (Extensible Markup Language).The intended scope of OAI-PMH was descriptive metadata. The purpose of OAI-PMH is to define a standard way to move metadata from point A to point B within the virtual information space of the World Wide Web. OAI service providers focused on basic cross-repository search-and-discovery services that are portals and gateways to make cataloging and related metadata about scholarly collections more visible to internet users. METS (Metadata Encoding and Transmission Standard) is one of many metadata formats used with OAI-PMH. METS is intended for encoding descriptive, administrative, and structural metadata regarding objects within a digital library. METS is designed to encompass all metadata necessary to use and preserve objects. Because it supports transmission of structural, descriptive and administrative metadata, it can be used in creating submission, archival, or dissemination information packages as defined in the reference model for an open archival information system (OAIS). The scope of METS as a metadata format consists of the following sections: METS header, descriptive metadata, administrative metadata, file section, structural map, structural links, and behaviour section.

Selection of the content

The selection involves the process of selecting the content appropriate to be digitized. Custodians of digital collections must decide: what aspects of the digital objects’ creation and use environment are important enough to warrant capture, documentation, and preservation over time; given limited resources and available techno-logy, which of those aspects of context can reasonably be preserved and how to carry out the preservation of the contextual information (Lee 2007).

Intellectual property

The intellectual property is about the copyright issues that surround the content of the digital library. The intellectual property person is responsible for applying for permission where it is needed, and writing the term of use for the digital library. Digital rights management pertains to text, images, data, and other media in digital libraries as much as in any other setting. Managing digital rights necessitates coping with licensing agreements, payment systems, secure transactions, user authentication, and usage tracking (Barton 2005).

Evaluation and usability

The evaluation and usability person develops testing instruments for the website of the digital library. The person's responsibility is to find people to test the workability of the website and be able to send their comments. The information gathered by the evaluation and usability person about the changes that are needed is then fed back to the web design person. Nielsen (2008) states that usability is a quality attribute that assesses how easy user interfaces are to use. The word usability also refers to the methods for improving ease-of-use during the design process. If a website is difficult to use, fails to clearly state what a company offers and what users can do on the site, people leave.

Marketing

The marketing person is responsible for promotion of the digital library to the targeted community. She/he needs to identify who the potential users will be and brand the digital library, and develop the mission statement.

Documentation/Photography

The photographer is responsible for documentation of the whole digital library creation process, and to preserve the project. The documentation will serve as a reference for future generations.

Quality control

The quality control entails editing content and checking that everything works as intended and that content is clear and accurate.

Conclusion

It is paramount to acknowledge the existence of open source resources and to go further and investigate how such resources can be used to the benefit of non-profit making institutions. The responsibility for driving such a development falls squarely on the shoulders of library management and library academics who are tasked with adapting national skills to match the changes in the global environment. The discussion in this paper has shown that a digitization project can be accomplished by many institutions with minimal external expertise or financial outlay.

References

Appu, A. 2002. *Administering and securing the Apache Server*. Cincinnati, Ohio: Premier Press.

Barton, M. R. and Waters, M. M. 2005. Creating an institutional repository: LEADIRS workbook. Learning about digital institutional repositories. [Online]. Available WWW: http://www.dspace.org/implement/leadirs.pdf (Accessed 29 December 2008).

Chiware, E. 2007. Training librarians for the digital age in African university libraries. IT and research in African university libraries: present and future trends. Paper read at the 73th PRE-IFLA Satellite Meeting in Durban, South Africa, August 2007. [Online]. Available WWW: http://www.ifla.org/IV/ifla73/papers/Sat1-Chiware-en.pdf (Accessed 02 January 2009).

Cole, T. W. and Foulonneau, M. 2007. *Using the open archives initiative protocol for metadata harvesting*. Westport, CT: Libraries Unlimited.

Lee, C. A. 2007. Taking context seriously: a framework for contextual information in digital collections. University of North Carolina, School of Information and Library Science. Technical Report 2007(04): 1-37. [Online]. Available WWW: http://sils.unc.edu/research/publications/reports/TR_2007_04.pdf (Accessed 12 January 2009).

Morgan, E. L. 2008. Introduction: open source software in libraries. *Bulletin of the American Society for Information Science and Technology* 35(2): 8-9.

Nielsen, J. [2008]. Usability 101: introduction to usability. [Online]. Available WWW: http://www.useit.com/alertbox/20030825.html (Accessed 12 January 2009).

OPENPROJ . 2009. 2nd most popular software after Microsoft Project: Gui. *Program News* 20(3): 3-6

Pitts, D. 1999. Red Hat Linux 6 unleashed. [Indianapolis, Ind.]: SAMS.

Schwartz, C. 2006. Digital libraries: an overview. *Journal of Academic Librarianship* 26(6): 385-394

Van Deventer, M. J. and Pienaar, H. 2008. South African repositories: bridging knowledge divides. *Ariadne* 55, April 2008. [Online]. Available WWW: (http://www.ariadne.ac.uk/issue55/vandeventer-pienaar/) (Accessed 30 December 2008).

Vaswani, V. 2005. *How to do everything with PHP and MySQL*. New York: McGraw-Hill Companies.

Witten, H. 2008. The development and usage of the Greenstone digital library software. *Bulletin of the American Society for Information Science and Technology* 35(2): 31-38.

Wrosch, J. 2007. Open source software options for any library. *MLA Forum* 5(3). [Online]. Available WWW: http://www.mlaforum.org/volumeV/issue3/article3.html (Accessed 30 December 2008).