PERSPECTIVES ON...

• Digital Preservation in Open-Source Digital Library Software

by Devika P. Madalli, Sunita Barve and Saiful Amin

Available online 10 March 2012

Digital archives and digital library projects are being initiated all over the world for materials of different formats and domains. To organize, store, and retrieve digital content, many libraries as well as archiving centers are using either proprietary or open-source software. While it is accepted that print media can survive for centuries with some physical preservation techniques, digital media requires continuous processes to keep it compliant with current technology. It is not only necessary to organize digital content but also important to preserve it to ensure accessibility, sustainability, and retrieval across time. This paper presents an analytical study along with observations regarding digital preservation support available in existing open-source digital library software (OSS-DL) based on test beds created for that purpose.

Devika P. Madalli,
Documentation Research Training Centre, Indian Statistical Institute, 8th Mile Mysore Road, RV College PO, Bangalore 560 059, India
<devika@drtc.isibang.ac.in>
Sunita Barve,
National Centre for Radio Astrophysics, PO Box 3, Pune Univ. Campus, Pune 411007, India
<sunitab@ncra.tifr.res.in>
Saiful Amin,
Documentation Research Training Centre, Indian Statistical Institute, 8th Mile Mysore Road, RV College PO, Bangalore 560 059, India
<saiful@drtc.isibang.ac.in>.

Introduction

Information is increasingly produced in digital form such as text in pdf, doc or odt music in mp3, wav video in mpeg formats, and vast amounts of digital content are made available to users. Digital information is growing and exploding at a rapid rate; it is also available in heterogeneous forms, adding to its complexity. Hardware and software on which digital information is created are continuously changing. This presents a significant challenge in preserving digital resources and making them accessible for future use.

Often, one of the aims of digital repositories that are available either via the Internet or on an intranet is to preserve the intellectual output irrespective of format and application used to create resources. There are also new challenges, particularly in the digital environment. When digital documents are added to a digital repository, it is necessary to ascertain that the software and tools lend support to long-term preservation of the digital content.

This paper presents evaluation criteria that may be applied to assess digital preservation capabilities while taking stock of the digital preservation support available in OSS-DL. Evaluation criteria from a digital preservation point of view are defined here based on a study undertaken for the purpose. The evaluation against the important criteria is executed and reported here by installing selected OSS-DL in a test bed environment.

Open-Source Software

There are quite a number of open-source software programs available for building digital libraries, institutional repositories, digital archives or digital repositories. Open-source software is available for free under open-source license terms and conditions where the source code of the software is also available. The open-source code can be altered for further development, customization, and redistribution.

Since 1997, open-source software (OSS) has claimed a substantial market share of the computer industry, and a large number of OSS is available on the Internet. The world's largest OSS development web site is SourceForge.net. The SourceForge repository hosts more than 326,882 projects and has more than 46 million registered users.[1]

Libraries have started making use of OSS for various library applications; one of the most prominent uses is for building "digital libraries". In digital preservation, OSS plays a vital role as the back-end technology libraries adapt are mainly based on open standards, which is an important criteria in digital preservation support. Open standards by not being proprietary allow libraries in being transient to migrations. For creating digital libraries, OSS tools are increasingly considered as an alternative to commercial digital library systems due to dissatisfaction with commercial software, mainly because of a lack of functionality, prohibitive costs, inadequate support, etc.[2] OSS-DL, with free access and a high level of functionality, has been used by a large community all over the world.

The software programs discussed in the following sections were shortlisted by the authors during evaluation procedure, on the basis of their functionality, especially for managing digital collections.

Digital Preservation

Digital preservation or digital archiving includes processes that ensure the longevity of electronic documents. It applies to documents that are both "born digital" and stored online, and to the products of analog-to-digital conversion, if long-term access is intended.[3] Digital information can easily be lost due to a variety of problems, such as technical failures, inability to access physical storage media on which digital information is stored, or loss of software that interprets the stored information.[4]

Evaluation Criteria

Digital preservation is a very challenging area, and recently much of the digital library software has started to focus on this issue. There are not many holistic studies in the area of digital preservation support available in OSS-DL. However, there is some research that has defined criteria for trusted[5] and trustworthy repositories[6] from a digital preservation point of view.

The criteria defined here are selected from the above studies as well as other resources that talk about digital preservation aspects of OSS-DL. Following are some of the criteria that are identified from a digital preservation point of view presented in categorized list:

a. *File features: size, format etc*
- Does the repository software support any digital preservation strategy, and if yes, does it explicitly support any particular preservation strategy, such as bit-level preservation, format migration, or restrictions on submission of file formats in the repository software?
- Does the system preserves files' original identities, such as its name, size, and created date?
- How does the software manage compound objects (where multiple file formats of the same object are linked together)?
- Can new file formats be added or removed?
- Does the software have the ability to handle a variety of file formats and does it also support file format versioning?
- Does repository software use any format registries? If yes, which format registry is used (such as PRONOM or GDFR or DCC) to represent format information?
- What standards does the repository software use to describe file formats and does the software record representation information? Does it use Internet MIME types?
- Does the system support automatic format registration? For unknown formats, does the system send any message to the submitter requesting additional information about describing the format of the document?

b: *Integrity check*
- Does the system have any data integrity check for digital documents that are added into the repository?
- Does the software have quality control measures to ensure integrity and persistent identification of every document that is added into the repository?
- Does the repository software preserve pre-existing persistent identifiers for submitted objects?

c. *Metadata*
- Does the software support preservation of metadata for every document that is added into the repository such as[7]:
  - provenance — documenting the history of the object.
  - authenticity — validating that the digital object is in fact what it should be, and has not been altered.
  - preservation activity — documenting the actions taken to preserve the digital object.
  - technical environment — describing the technical requirements, such as hardware and software, needed to render and use the digital objects.
- Where does the repository software store the actual digital files and the metadata?
- How does the repository software verify that archival objects and metadata are correct? Are there automated checks of the metadata, such as to verify that a date entered into a field really is a date string?

d. *Licensing, rights and authentication*
- Does it have a mechanism to keep licensing conditions for individual images/objects in the repository?
- Does the software support tracking and managing copyrights and restrictions on use as required by contract or license?
- Can the repository software automatically validate checksums on a periodic basis?
- Are the checksums cryptographically signed to prevent tampering?
- Does the repository software provide audit logs of all events that have occurred in the life cycle of a package? What events are logged?
- If the repository ingests digital content with unclear ownership/rights, does software have policies addressing liability and challenges to those rights?
- Does the system have any mechanism in determining when objects in digital archives should migrate to new hardware and software?
- Can the software support scheduled events such that a human can be notified on a preset schedule to manually check for format obsolescence?

Data Collection and Analysis

Test Bed Setup

We have set up a test bed environment where we have installed major OSS-DL. All of the selected software are available under open source license terms and conditions, especially CDS-Invenio, DSpace, EPrints, FEDORA, Greenstone, DoKS, and MyCoRe. These software programs are used specially for creating digital archives/digital libraries/institutional repositories. All of these software programs are installed on a Debian Lenny Operating System with 2 GB RAM and 1 TB disk space. Latest versions of each software are selected and installed on the test bed server. A small collection of all document types, such as text, audio, video, data set files, etc., is uploaded for verifying different digital preservation features supported by the software.

Initially, nine OSS-DL were selected for the present study, including:

- CDS-Invenio (Switzerland)
- DoKS (Belgium)
- DSpace (USA)
- EPrints (UK)
- FEDORA (USA)
- Greenstone (New Zealand)
- MyCoRe (Germany)
- OPUS (Germany)
- SciX (Slovenia).

The DoKS software and SciX software were excluded from further evaluation due to non-availability of new versions as well as limited support for digital document management. Similarly, OPUS was excluded from further evaluation due to non-availability of its documentation in English. The following versions of software were installed on a test bed server:

- CDS-Invenio — Version 0.99.3
- DSpace — Version 1.7.0

- EPrints — Version 3.2.4
- FEDORA — Version 3.4.2
- Greenstone — Version 2.83
- MyCoRe — Version 2.0.2.

After installing and adding sample data, the installations were tested against the criteria listed above.

Metadata Support

Ideally, each digital document that gets added into the repository should have a detailed description; the software should have the ability to add administrative, structural, and bibliographic details to every digital object. Preservation metadata is not a different category but covers: technical details of the format, structure, and use of the digital content; history of the all actions performed on the resource including changes and decisions; authenticity information such as technical features or custody history; and the responsibility and rights of the document.[8] Administrative and technical metadata are important from a long-term preservation point of view. There are a variety of metadata formats available today for different types of documents, including Dublin Core, MODS, MARC, METS, LOM, ETDMS, MPEG21/DIDL, etc., hence it is necessary to know what metadata standard a OSS-DL supports. The following metadata support is available in the 6 software programs selected:

Software	Metadata formats
CDS-Invenio	Dublic Core, MARC 21
DSpace, MyCoRe EPrints	Dublin Core
FEDORA	Dublin Core, METS, MPEG-21, DIDL, IEEE, LOM, MARC, FOXML, ATOM
Greenstone	Dublin Core, New Zealand Government Locator Service Metadata Standard, RFC 1807 Metadata Element Set, Development Library Subset Example Metadata, Greenstone Metadata Set, Australian Government Locator Service Metadata

Persistent Identification

For stable long-term management of digital collections, persistent identifiers are required. A unique identifier is assigned to every digital document that is added to the digital repository software. This helps to establish a unique, widely supported identifier for digital documents on the Internet. For scholars to confidently cite a digital object, they must be assured that the object will be accessible via the citation for many years in future.[9] The following are different persistent identification features supported by each selected software:

CDS-Invenio: Its own identification number
DSpace: Handle.net
EPrints: Generates URI (Uniform Resource Identifier, a unique locator/address for a resource) for every document; also allows users to add other persistent identification numbers for every document
FEDORA: URI, as well as other identifier schemes found in PRONOM and the Global
Digital Format Registry (GDRF)
Greenstone: No Persistent Identification Scheme supported yet
MyCoRe: Uniform Resource Name.

Checksum and Versioning Support

A *checksum* is a computed value that allows one to check the validity of a digital resource. Typically, checksums are used in data transmission contexts to detect if the data has been transmitted successfully over the Internet.[10]

CDS-Invenio, DSpace, EPrints, and MyCoRe support MD5 checksums. Fedora supports a variety of checksum mechanisms, such as MD5, SHA-1, SHA-256, SHA-384, and SHA-512. No support for any checksum is provided with Greenstone.

CDS-Invenio, DSpace, FEDORA, MyCoRe, and EPrints support adding different versions of documents. Greenstone does not provide support for adding different versions of documents.

Automatic Format Recognition

Digital repositories can be configured to recognize certain prevalent formats. In this section we present the capabilities of the different softwares compared for this particular feature:

- CDS-Invenio: Has the ability to accept documents in all desired formats. The system administrator can limit the formats of submitted documents. This allows the repository to define a policy according to which it accepts specific formats of digital objects that it can manage from a technical point of view.
- DSpace: Provides support for as many file formats as possible, but the proprietary nature of many file types are not identified while uploading and they are treated as “other” formats.
- EPrints: Has some file formats listed, which are identified as soon as file is uploaded. The rest of the files are marked as “other” when uploaded into the repository.
- FEDORA: All MIME type file formats supported.
- Greenstone: When any digital document is uploaded, GSDL tries to identify the format of the file and suitable plugin required for opening the file, but not all file formats are recognized by Greenstone.
- MyCoRe: Uploads any file format but no preservation support.

Audit Logs

- CDS-Invenio: Software maintains search, indexing, and Apache logs. All the logs are stored in Invenio installation part.
- DSpace: Detailed log is supported in DSpace with its own log as well as Tomcat log.
- EPrints: Software does not have any log area of its own. Since it is running on Apache Web Server, it records some information in Apache (access/error log area).
- FEDORA: Software keeps detailed log in its area, such as client and server logs, and Tomcat logs are maintained.
- Greenstone: Maintains access and error logs.
- MyCoRe: Maintains log as a .txt file, but no detailed logs are maintained.

Details of the files

- CDS-Invenio: replaces files' original name with its own name.
- EPrints, Greenstone: Preserve files' own identity when uploaded into the software.
- DSpace, FEDORA: Change the file name to its own internal structure.
- MyCoRe: Changes file names with date, time, and internal number.

Actual Data File Storage

When actual files are stored, they are located in a separate directory.

- CDS-Invenio: Actual files are stored in “data” directory and metadata is stored in “mysql” tables.
- DSpace: Actual files are stored in “assetstore” folder and metadata is stored in “postgres” database.
- EPrints: The metadata is stored in “mysql” and actual files are stored in “disk0” directory of EPrints.

- FEDORA: Actual files are stored in "data/datastream" folder and metadata is stored in "mysql" database.
- Greenstone: All files and metadata are stored in "import" folder of installation. Metadata is stored as a metadata.xml file in GSDL.
- MyCoRe: Metadata and actual files are stored in "data" folder of docportal. MyCoRe uses "hsql" (Hyperstructured Query Language) database.

Conclusions

There is much yet to be known and studied when it comes to the preservation of digital information. It can be said that the state of development in digital preservation is still in a very early and experimental stage. There are no established models for preserving multimedia-related works, online dialogues, etc.[11] It is necessary to convert items from proprietary formats into open formats and open standards, so they can then be uploaded into a digital archive for future storage, retrieval, and preservation. Libraries will have to deal with digital materials in the future, so it is necessary that these software programs have proper digital preservation support with more user-friendly interfaces as well as with proper submission guidelines, as mentioned in OAIS (Open Archival Information System) Reference Manual.

Similarly, OSS for DLs have yet to come out with proper digital preservation support. To a large extent FEDORA supports more features that are essential from a digital preservation point of view, but it lacks a user-friendly interface; hence, there are not many installations of FEDORA. DSpace and EPrints are now used heavily all over the world to build digital repositories/institutional repositories. To some extent, both of these software programs support digital preservation. There are a large number of repositories available with DSpace. In India, many institutes have taken steps to build digital archives using DSpace.

It is necessary to convert items from proprietary formats into open formats and open standards, so they can then be uploaded into a digital archive for future storage, retrieval, and preservation. Libraries will have to deal with digital materials in the future, so it is necessary that these software programs have proper digital preservation support with more user-friendly interfaces as well as with proper submission guidelines, as mentioned in OAIS Reference Manual.

Notes and References

1. "SourceForge — Download, Develop and Publish Free Open Source Software," *Source Forge*, http://sourceforge.net/. [1st February 2011].
2. Marshall Breeding, "An update on Open Source ILS," *Information Today* 19, no. 9 (2002): 42–43. http://www.onlineinc.com/it/oct02/breeding.htm [1st February 2011].
3. A. Bullock, "Preservation of Digital Information: Issues and Current Status," *National Library of Canada*: 1999.
4. Andrew Waugh et al., "Preserving digital information forever," in *Digital Libraries* (ACM), 175–184.
5. Trusted Digital Repositories: Attributes and Responsibilities An RLG-OCLC Report (Mountain View, CA: Research Libraries Group, 2002), 70 p. http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf. [10th January 2011].
6. Trustworthy Repositories Audit & Certification: Criteria and Checklist, Version 1.0 (USA: OCLC, February 2007) 94 p. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf. [1st February 2011].
7. Patricia Galloway, "Preservation of digital objects," *Annual Review of Information Science and Technology (ARIST)* 38 (2004): 549–590.
8. Ingeborg Verheul, "Networking for digital preservation: current practice in 15 national libraries," in *IFLA Publication 119* (K. G. Saur, 2006), 271 p., http://www.ifla.org/VI/7/pub/IFLAPublication-No119.pdf. [1st February 2011].
9. Ronald Jantz, "Digital preservation: architecture and technology for trusted digital repositories," *D-Lib Magazine* 11, no. (6) (June 2005), http://www.dlib.org/dlib/june05/jantz/06jantz.html. [1st February 2011].
10. "Checksum Algorithm", http://www.flounder.com/checksum.htm. [10th March 2011].
11. Margaret Hedstrom, *Digital preservation: a time bomb for digital libraries," Computers and the Humanities 31 (1998): 189–202.*