Course 1: Digital Libraries – Introduction

Unit 4: Digital libraries functional components – software

Learning objectives

- To acquaint with software requirements of D.L.s
- To understand the available digital library software both proprietary and open source

1.4.1. INTRODUCTION

The digital libraries are increasing in number and they need software that cost-effectively manages virtually unlimited amounts of digital and electronic content. To create a digital library the following software are required

- Digital library software; OCR, image editing software.

They differ from integrated library management software (ILMS) as their purpose is library housekeeping and information retrieval while DL software is to create digital collections for full text access, preserve different formats from rare manuscripts to born digital for future conservation. However the ILMS also facilitate management of digital collections. Ex. SOUL (INFLIBNET); LS- Digital (LIBSYS)

The DL software depends on the type of digital library that is planned, whether it is large DL or a medium sized; whether it is a digital library or an institutional repository; whether it is a collection of only digital images or documents or a combination. There is software developed specifically for digital libraries that are based on local server network or using latest cloud computing technologies.

1.4.2. D.L. SOFTWARE

Software is probably the most important and central aspect of the technology infrastructure and digital library is not an exception. Digital library software works with the web server (and the operating system) in providing various digital library functionality including creation, organization and maintenance, indexing, search and retrieval.
Several software options are available today for creation and provision of access to digital library collections. A clear understanding of the requirements and features of the collection we plan to build is very important to assess and select the option that best suits our need. A good understanding of features expected of digital library software will also help us in making informed decision.

The features of digital library software include: support for different document types and formats, metadata support, online/batch content updating, indexing and storage, search and retrieval (metadata, full-text), multi-lingual support, interoperability support, access and usage management, collection administration, and support for standards compliance like Dublin Core, Unicode and XML.

Many of the digital library software are available as proprietary and as ‘open source’ that is a free software license and source code is available freely for customization and updating.

Examples of commercial / proprietary digital library software:

- EnCompass of Endeavor Information Systems (http://www.endinfosys.com)
- DigiTool of Ex Libris (http://www.exlibris.co.il, or http://www.exlibris-usa.com) Visual MIS (Multimedia and Imaging Solutions) of VTLS (http://www.vtls.com/)
- TEAMS of Artesia Technologies (http://www.artesia.com/)
- Insight of Luna Imaging (http://www.luna-imaging.com/)
- MuseSearch of MuseGlobal, Inc. (http://www.museglobal.com/)
- EOS.WEeb.

Examples of open source digital library software: Software available via open source license, complies with latest version of OAI PMH:

- Archimede- Canadian software solution for institutional repositories (http://www.bibl.ulaval.ca/archimede)
- Arno - Academic Research in the Netherlands Online, Tilburg University, The Netherlands(http://www.uba.uva.nl/arno)
- DSpace - MIT Libraries and the HP Labs, USA (http://www.dspace.org)
Examples of D.L. Software

There are time tested open source DL software, but Greenstone Digital Library System (GSDL) and Dspace happen to be the most popular in India. All of these are available for Linux operating system platform, a freely available popular operating system. Most of these packages also support the OAI interoperability protocol, an attractive proposition to share digital library collection with other collections on the Internet.

GSDL: [http://greenstone.org](http://greenstone.org)

Greenstone Digital Library software is produced by the New Zealand Digital Library Project at the University of Waikato. This project is a research programme aiming to develop the underlying technology for digital libraries and make it available publicly so that others can use it to create their own collections. The main architects of the software are Roger McNab and Stefan Boddie. Greenstone is software for building and distributing the digital library collection. It organizes the information digitally for publishing the internet or on CD-ROM. This software is developed and distributed in cooperation with UNESCO and the Human Info NGO. It is open source software available from [http://greenstone.org](http://greenstone.org) under the term of the GNU General Public License.

List of few organizations using GSDL in India is:

- Indian Institute of Management, Kozhikode ([http://www.iimk.ac.in](http://www.iimk.ac.in))
-Indian Institute of Science Publication Database
- Indian Institute of Technology, Bombay etc
- Baba Farid University of Health Sciences (BFUHS) ([http://library.babafariduniv.com](http://library.babafariduniv.com))
DSpace: [http://www.dspace.org](http://www.dspace.org)

DSpace is an open-source software package, which provides the tools for management of digital assets, and is commonly used as the basis for an institutional repository. It is also intended as a platform for digital preservation activities. Since its release in 2002, it has been a product of the HP-MIT Alliance, and is shared under a BSD licence. This software accepts all forms of digital materials including text, images, video, and audio files. There are more than 60 institutions that have adopted DSpace to maintain their digital collections. Few examples are:

- Central Institute of Medicinal and Aromatic Plants [http://kr.cimap.res.in/index.jsp](http://kr.cimap.res.in/index.jsp)
- Cochin University of Science & Technology [http://dspace.cusat.ac.in/dspace/](http://dspace.cusat.ac.in/dspace/)
- Indian Institute of Science, Bangalore [http://etd.ncsi.iisc.ernet.in/](http://etd.ncsi.iisc.ernet.in/)
- Indian Statistical Institute, Bangalore [https://drtc.isibang.ac.in/](https://drtc.isibang.ac.in/)

FEDORA: [www.fedora-commons.org](http://www.fedora-commons.org)

FEDORA or Flexible Extensible Digital Object Repository Architecture is a modular architecture built on the principle that interoperability and extensibility is best achieved by the integration of data, interfaces and mechanism as clearly defined modules. It is a Digital Asset Management (DAM) architecture, upon which many types of digital library, institutional repositories, digital archives, and digital libraries systems might build. Fedora is the underlying architecture for a digital repository and is not a complete management, indexing, discovery, delivery application. Fedora is developed jointly by Cornell University Information Science and the University of Virginia in 1997 by Sandra Payette, Carl Lagoze and Naomi Dushay. Since then, several modifications have been made to the architecture and in late 2005 version 2.1 were released. The current version is 3.2. It began as a DARPA and NSF-funded research project of Sandra Payette and Carl Lagoze at Cornell University’s Digital Library Research Group in 1997, where the first reference implementation and a COBRA based technical implementation were built.
The FEDORA Project is currently supported by generous grants from Andrew W. Mellon Foundation.

E-Print: [http://www.eprints.org](http://www.eprints.org)

E-Print software was developed as part of digital library project at University of Southampton, UK. It is available free under the term GNU (General Public License). It runs under Linux and creates online archive libraries in electronic prints. The default configuration creates a research paper archive, but could be modified and used for other purposes. The software is easy to install. The installer script automated the installation process. The document can be stored in any format and each individual E-Print i.e., research paper, can be stored in more than one format. The archive can use any metadata schema and it is to be decided what metadata fields ex. authors, title and journal to be kept for each e-print. The mandatory fields can be decided on the basis of the format in which it is stored.

List of institutional repositories using EPrints in India:

- National Centre for Catalysis Research (IIT): Catalysis Database [http://eprints.iisc.ernet.in/](http://eprints.iisc.ernet.in/)
- ePrints@IISc, National Centre for Science Information (NCSI), Indian Institute of Science, Bengaluru [http://eprints.iisc.ernet.in/](http://eprints.iisc.ernet.in/)
- Eprint@DU, University of Delhi [http://eprints.du.ac.in/](http://eprints.du.ac.in/)
- ePrints@IIMK, Indian Institute of Management Kozhikode Scholarship Repository [http://eprints.iimk.ac.in/](http://eprints.iimk.ac.in/)
- Indian Institute of Information Technology [http://eprints.iiita.ac.in/](http://eprints.iiita.ac.in/)
- OneWorld South Asia Open Archive Initiative [http://open.ekduniya.net/](http://open.ekduniya.net/)
- OpenMED@NIC [http://openmed.nic.in/](http://openmed.nic.in/)

Analysis of the use of above software reveals that GSDL is mostly used for Open Archive; the installation of GSDL on Windows Operating system is comparatively easy and one can also copy the entire DL developed on GSDL on to a CD / DVD for distribution.

Dspace is used to built institutional repositories, long term preservation of digital assets, image repositories, audio video repositories, Government Records reposts, networking, subjects
repositories, museum culture and it was mostly used for e-learning resources; Dspace provides a very detailed workflow so that any one can submit the document and can be scrutinized at various stages, before it is finally uploaded on the system. It uses a fuzzy logic based Lucene search engine.

FEDORA is mostly used for digital collections, education, science/research, institutional repositories, open access and finally in preserving and archiving. EPrints are used for open access research literature, scientific data, theses, reports and multimedia.

1.4.3. BASIC FEATURES OF D.L. SOFTWARE

Digital library software has to reflect the functions of the DL. The first function is to capture the material followed by organizing the same using metadata. Organization of DL is standards-based and supported by Unicode, metadata based on qualified Dublin Core, XML, and OAI-PMH. Therefore the DL software has to include all the following features:

- Capture and describe digital material using a workflow;
- Provide interface for online submission of research material (Intranet);
- Provide access to this material over the web (metadata and/or full text);
- Preserve digital material over long period of time;
- Expose metadata through OAI-PMH protocol with unqualified Dublin Core and other metadata standards.

Key requirements for digital library software: The key requirements for digital library software, according to K.T. Anuradha (2006) include:

- Document type (book, journal article, lecture..)
- Document formats (text, PDF, Word, PS, ...)
- Content acquisition (online and offline)
  - Metadata description, content tagging
  - Content uploading
- Indexing and retrieval
  - Structured/ full text indexing
  - Automatic metadata extraction
- Searching facility
• Boolean, Natural language; Phrase searching, Proximity, Exact match, Truncation

• Storage
  • Data compression; Efficient storage for metadata; Efficient location of metadata and documents

• Access and delivery
  • Results display: Structured search, browse, hierarchical browsing
  • On the web, E-mail, CD-ROM distribution

• Scaling up – for large collections
• Multilingual support
• Access management and security
• Usage monitoring and reporting
• Standards compliance
  • XML, Dublin Core, Unicode
• Interoperation
  • OAI, Z39.50 compliance, MARC

Criteria for selection of DL Software: Selection of software is the key activity in building a digital library. The software ought to support the activities of the designed library. Therefore while choosing DL software, basic features have to be evaluated. Repository Support Project in its software survey (2010) has defined the following criteria for the DL software:

1. License cost – free or commercial
2. Support – free, not-for profit paid, paid/ commercial
3. Update cost – free or with license
4. Supported item types – documents, images, audio, video, learning objects etc
5. Format conversion – PDF, XML or any standard format
6. Search logic – field specific, Boolean, other sorting options
7. Browse view options – Academic unit, subject, title, year, total collection etc
8. Default subject organization – DDC, UDC, Library of Congress Classification or any classification
9. Syndication – RSS, Atom
10. User validation – Registration required, LDAP authentication, Athens authentication
11. Web 2.0 features – bookmarks, rating, reviews, share, tagging etc
12. Statistical reporting – downloads or full records
14. Databases – Oracle, Postgre SQL, My SQL, Cloud storage
15. Script language – JAVA, PERL, PHP, AKAX etc
17. Administrative functions – import, export, workflow, report generation
18. Service support – training, consultancy, site visit etc

1.4.4. Optical Character Reader (OCR) software: OCR software is required to scan the printed or written documents and convert the scanned image into computer readable format that are editable, searchable and reused. “A system that provides a full alphanumeric recognition of printed or handwritten characters at electronic speed by simply scanning the documents that called OCR” (http://www.unescap.org/stat/pop-it/pop-guide/capture_ch06.pdf) Intelligent Character Recognition (ICR) is the module of OCR that has the ability to turn images of handwritten or printed characters into ASCII data. Sometimes OCR is known as ICR. Both are recognition engines used with imaging. OCR technologies recognize all types of characters / information and allow scanning, indexing and written in optical media. Therefore it provides a solution to document capture, and create an electronic document; indexing the documents. A major advantage of OCR is the converted documents take less space than the original file.

Most scanners come with an Optical Character Recognition, or OCR, program of some sort to make PDFs searchable. However there is some free OCR software available if one want have quick conversion.

Examples:

ABBY FineReader OCR software helps individuals turn scans of paper documents, PDF files, and digital photographs into searchable and editable formats.
**FreeOCR** is a free Optical Character Recognition Software for Windows and supports scanning from most Twain scanners and can also open most scanned PDF's, popular image file formats.

**RICOH Innovations** has created a Beta applications, **online Document Conversion tool**. The document conversion widget provides free OCR to convert the images into editable and searchable pdf, MsWord, HTML and text documents, providing capabilities such as pdf to doc conversion.

**Process of OCRing:** The process generally consists of three stages: scan the document, recognize it and then save in a convenient format (DOC, RTF, XLS, PDF, HTML, TXT etc.) or export data directly to one of Office applications such as Microsoft Word, Excel or Adobe Acrobat. The entire process of data conversion from original paper document, image or PDF takes less than a minute, and the final recognized document looks just like the original. A powerful OCR software allows to save a lot of time and effort when creating, processing and repurposing various documents.

**Functions of OCR:** OCR allows users

- To quickly automate data capture from forms,
- Eliminate keystrokes to reduce data entry costs
- Maintain the high level of accuracy required in applications.

In library context, OCR technologies are used

- Archival of printed documents in electronic format ex. TKDL
- to make electronic images of printed documents searchable, ex. Vidyanidhi

**1.4.5. SUMMARY**

Software is the important component to create and maintain a digital collection. There are commercial and open source software available. The open source digital library software like DSpace, EPrints, GSDL etc. are time tested and more popular India with technical support from LIS professional groups. The choice of software should take care of functionality issues like server platform support, Web supported, main program modules, source code availability etc. Therefore the LIS professionals need basic knowledge of the digital library software and be able
to compare their features. This unit provides basic understanding of the digital library software while Course 5 dealt with application aspects of DL software with focus on GSDL and DSpace.