International Journal on Emerging Technologies **5**(1): 82-86(2014)

ISSN No. (Online) : 2249-3255

Comparison of Open Source Digital Library Software with Special reference to Greenstone, DSpace and EPrints

Dr. Sadhna Saxena*, Dr. Surendra Singh** and Dr. V.A. Mulchandani***

**Technical Officer, M.P. Council of Science & Technology, Bhopal, (MP)*
***Prof. & Head, Department of Library & Information Science,*
Sandipani P.G. College, Ujjain, (MP)
****Sr. Technical Officer, M.P. Council of Science & Technology, Bhopal, (MP)*

(Corresponding author: Dr. Sadhna Saxena)

(Received 05 February, 2014 Accepted 19 April, 2014)

ABSTRACT: In the past years a great number of digital library and digital repository systems have been developed by individual organizations -mostly Universities- and given to the public as open-source software. The advantage of having many choices becomes a great headache when selecting a Digital Library (DL) system for a specific organization. To make the decision easier, we compared three such systems that are publicly available using an open source license, are compliant with Open Archives Initiative Protocol for Metadata Harvesting (OAIPMH) and already have a number of installations worldwide. Using these basic restrictions we selected for comparison the following Three broadly used DL systems i.e. DSpace, Greenstone, EPrints.

Each of these systems was been thoroughly studied based on basic characteristics and system features described in the following phases. The latest versions of those systems were examined. The DL systems are compared based on stated characteristics and the level of support on each of them. In phase 2, the characteristics needed by a modern DL system are discussed. In phase 3, the three DL systems are compared based on each of the DL characteristics and the results are summarized in a score table. Each system has its advantages and drawbacks, as stated in the above comparison, categorized by basic DL system characteristics and features. That comparison can only be used as a guideline by an organization in order to decide if one of these DL systems is suitable to host its digital collections.

Keywords: *Digital Library Software, Digital Library Architecture, DSpace, GreenStone, EPrints*

I. INTRODUCTION

Over a period of time, many proponents have forwarded various definitions of digital libraries and still more continue to emerge each day. According to Waters [1] the partner institutions in the Digital Library Federation (DLF) realized in the course of developing their program that they needed a common understanding of what digital libraries are if they were to achieve the goal of effectively federating them. So they crafted the following definition, with the understanding that it might well undergo revision as they worked together [2].

"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".

Arms [3] defined Digital library as: "A managed collection of information with associated services, where the information is stored in digital formats and is accessible over a network".

A digital library may allow either online or offline access to the elements it organizes and houses and may include multimedia as well as multilingual data. Digital library is an evolving area of research, development and application. Digital libraries is useful to :-

(i) Preserve the valuable documents, rare and special collections of libraries, archives and museums.
(ii) Provide faster access to the holding of libraries world wide through automated catalogues.
(iii) Help to locate both physical and digitized versions of scholarly articles and books through single interface.
(iv) Search optimization, simultaneous searches of the Internet make possible, preparing commercial databases and library collections.
(v) Offering online learning environment.
(vi) Making short the chain from author to user.
(vii) Save preparation/ conservation cost, space and money.
(viii) Digital technology affords multiple, simultaneous user from a single original which are not possible for materials stored in any other forms.

In the last years a great number of digital library and digital repository systems have been developed by individual organizations -mostly Universities- and given to the public as open-source software. The advantage of having many choices becomes a great headache when selecting a Digital Library (DL) system for a specific organization. To make the decision easier, we compared three such systems that are publicly available using an open source license, are compliant with Open Archives Initiative Protocol for Metadata Harvesting (OAIPMH) [4] and already have a number of installations worldwide. Using these basic restrictions we selected for comparison the following Three broadly used DL systems:

- DSpace [5], developed by the MIT Libraries and Hewlett-Packard Labs (BSD open source license)
- Greenstone [6], produced by the University of Waikato (GNU General Public License)
- EPrints [7], developed by the University of Southampton

Each of these systems was been thoroughly studied based on basic characteristics and system features described in the following phases. The latest versions of those systems were examined. The DL systems are compared based on stated characteristics and the level of support on each of them. In phase 2, the characteristics needed by a modern DL system are discussed. In phase 3, the three DL systems are compared based on each of the DL characteristics and the results are summarized in a score table.

Finally, in phase 4, the results of this comparison are commented and cases for which, each of these systems is suitable, are proposed.

II. DL SYSTEMS CHARACTERISTICS

The basic characteristics and features that expect from a modern integrated DL software are:

1. *Object model*. The internal structure of the digital object [8] (entity that integrates metadata and digital content) in the DL system. Existence of unique identifiers for the digital object and every part of it is also important to ensure preservation and easy access.

2. *Collections and relations support*. Collection description metadata, definition of sub-collections and templates that describe the format of the digital objects or the presentation of the collection. Definition of relations between objects of the same or different types.

3. *Metadata and digital content storage*. The storage capabilities are stated, along with the preservation issues. It is important for the DL system to ensure standard as long as user defined metadata sets and multiple formats of digital content.

4. *Search and browse*. The mechanisms used for indexing and searching of the metadata. It is important for the DL system to support indexing not only for a restricted metadata set, but also for selected metadata fields.

5. *Object management*. Methods and user interfaces provided from the DL system to manipulate (insert, update and delete) metadata and digital content.

6. *User interfaces*. Provided user interfaces for end-user access on the DL, its collections and the digital objects.

7. *Access control*. Support for users and groups, authentication and authorization methods. Level of restriction for access and update (DL, collection, digital object and content).

8. *Multiple languages support*. Multiple languages should be supported in the user interface, in the metadata fields and in the digital content. The character encoding is of great importance in order for the DL systems to be fully multilingual.

9. *Interoperability features*. Standards that the DL systems support in order to ensure interoperability with other systems. Export of the digital objects in open standard formats is also important.

10. *Level of customization*. Customization of the DL system in collection level, the format of the digital objects and the services provided. The quality and methods provided by the application programming interfaces (APIs) of the DL systems.

III. DL SYSTEMS COMPARISON

In the following, the five open access DL systems are compared based on the characteristics identified in the previous section. The level of support of each characteristic and specific considerations for each DL system are discussed.

Object model: *Greenstone*: Basic entity in Greenstone is *document,* which is expressed in XML format. Documents are linked with one or more resources that represent the digital content of the object. Each document contains a unique document identifier but there is no support for persistent identifiers of the resources.

DSpace: The basic entity in DSpace is *item*, which contains both metadata and digital content. Qualified Dublin Core (DC) [9] metadata fields are stored in the item, while other metadata sets and digital content are defined as bit streams and categorized as bundles of the item. The internal structure of an item is expressed by structural metadata, which define the relationships between the constituent parts of an item.

DSpace uses globally unique identifiers for items based on CNRI Handle System. Persistent identifiers are also used for the bit streams of every item.

EPrints: Basic entity in EPrints is the *data object*, which is a record containing metadata. One or more documents (files) can be linked with the data object. Each data object has a unique identifier

Collections and Relations Support: *Greenstone*: A collection in Greenstone defines a set of characteristics that describe its functionality. These characteristics are: indexing, searching and browsing capabilities, file formats, conversion Plugging and entry points for the digital content import. There are also some characteristics for the presentation of the collection. The representation of hierarchical structure in text documents is supported for chapters, sections and paragraphs. The definition of specific sections in text document is implemented through special XML tags.

XLinks in a document can be used to relate it with other documents or resources.

DSpace: Supports collections of items and communities that hold one or more collections. An item belongs to one or more collections, but has only one owner collection. It is feasible to define default values for the metadata fields in a collection. The descriptive metadata defined for a collection are the title and description.

There is no support of relations between different items.

EPrints: There is no consideration of collections in EPrints. Data objects are grouped depending on specific fields (subject, year, title, etc). There is no definition of relations between documents, except using URLs in specific metadata fields.

Metadata and Digital Content Storage: *Greenstone*: Both documents and resources are stored on file system. Metadata are user defined and are stored in documents using an internal XML format.

DSpace: DSpace stores qualified DC metadata in a relational database (PostgreSQL or Oracle). Other metadata sets and digital content are represented as bit streams and are stored on file system. Each bit stream is associated with a specific bit stream format. A support level is defined for every bit stream format, indicating the level of preservation for the specified file format.

EPrints: Metadata fields in EPrints are user-defined. The data object, containing metadata, is stored in a MySQL database and the documents (digital content) are stored on file system.

Search and Browse

Greenstone: Indexing is offered for the text documents and specific metadata fields. Searching capabilities provided for defined sections in a document (Title, chapter, paragraph) or in whole document. Stemming and case sensitive searching is also available. Managing Gigabytes (MG) open-source applications is used to support indexing and searching. Browsing catalogs can be defined for specific fields using hierarchical structure.

DSpace: Provides indexing for the basic metadata set (qualified DC) by default, using the relational database.

Indexing of other defined metadata sets is also provided using Jakarta Lucene API. Lucene supports fielded search, stemming and stop words removal. Searching can be constrained in a collection or community. Also, browsing is offered by default on title, author and date fields.

EPrints: Indexing is supported for every metadata field, using the MySQL database. Full text indexing is supported for selected fields. Combined fielded search and free text search are provided to the end-user. Browsing is provided using specified fields (e.g. title, author, subject).

Object Management

Greenstone: New collections and the contained documents are built using the Greenstone Librarian Interface or the command line building program.

DSpace: Items in DSpace are created using the web submission user interface or the batch item importer, which ingests XML metadata documents and the constituent content files. In both cases a workflow process may initiate depending on the collection configuration. The workflow can be configured to contain from one to three

steps where different users or groups may intervene to the item submission. Collections and communities are created using the web user interface.

EPrints: A default web user interface is provided for the creation and editing of objects. Authority records can be used helping the completion of specific fields (e.g. authors, title). Objects can also be imported from text files using multiple formats (METS, DC, MODS, BibTeX, EndNote).

User Interfaces

Greenstone: The default web user interface provides browsing and searching into collections, navigating into hierarchical objects (like books) using table of contents. Presentation of documents or search results may differ depending on specified XSLTs.

DSpace: A default web user interface is provided in order for the end-user to browse a collection, view the qualified DC metadata of an item and navigate to its bit streams. Navigation into an item is supported through the structural metadata that may determine the ordering of complex content (like book pages or web pages).

A searching interface is provided by default that allows the user to search using keywords.

EPrints: The web user interface provides browsing by selected metadata fields (usually subject, title or date). Browsing can be hierarchical for subject fields. Searching environment allows user to restrict the search query using multiple fields and select values from lists.

Access control

Greenstone: A user in Greenstone belongs to one of two predefined user groups: an administrator or a collection builder. The first user group has the right to create and delete users, while the second builds and updates collections. End-users have access to all the collections and the documents.

DSpace: It supports users (e-people) and groups that hold different rights. Authentication is provided through user passwords, X509 certificates or LDAP. Access control rights are kept for each item and define the actions that a user is able to perform.

These actions are: read/write the bit streams of an item, add/remove the bundles of an item, read/write an item, add/remove an item in a collection. Rights are based in a default-deny policy.

EPrints: Registered users in EPrints are able to create and edit objects. Users are logged in using their username and password pair.

Multiple Languages Support

All the DL systems use Unicode character encoding, so the support of different languages can be supported. Every system can use multiple languages in the metadata fields and digital content.

EPrints provide an XML attribute on metadata fields to define the language used for the field value. Greenstone provides ready to use multilingual interfaces already translated in many languages.

Interoperability Features

All the DL systems support OAI-PMH in order to share the metadata of the DL with other repositories. Greenstone support Z39.50 protocol for answering queries on specific metadata sets. DSpace is able to export digital objects as METS XML files. Both systems also use persistence URLs to access the digital content providing a unified access mechanism to external services. DSpace also supports OpenURL protocol providing links for every item page. EPrints exports data objects in METS [10] and MPEG-21 Digital Item Declaration Language (DIDL) format.

Level of Customization

Greenstone: It provides customization for the presentation of a collection based on XSLTs and agents that control specific actions of the DL. Greenstone architecture provides (i) a back end that contains the collections and the documents as long as services to manage them and (ii) a web based front end that is responsible for the presentation of collections, documents and their searching environment.

DSpace: Although DSpace has a flexible object model is not so open in constructing very different objects with independent metadata sets because of its database oriented architecture. The user interface is fixed and provides only minor presentation interventions. Another disadvantage is the full support of only specific file formats as digital content.

EPrints: The data objects in EPrints contain user defined metadata. Plug-ins can be written in order to export the data objects in different text formats. A Core API in Perl is provided for developers who prefer to access basic DL functionality. Based on the above analysis, the three DL systems were graded for each of the characteristics. The minimum score is 1 and the maximum is 5.

Characteristics	**DSpace**	**Greenstone**	**EPrints**
Object Model	4	3	2
Collection support and relations	4	5	1
Metadata and digital content storage	4	3	3
Search and browse	4	4	4
Object management	4	2	4
User interfaces	4	4	4
Access control	5	2	2
Multiple languages support	3	4	4
Interoperability features	5	4	5
Level of customization	3	4	3

IV. CONCLUSION AND SUGGESTIONS

It is difficult to propose one specific DL system as the most suitable for all cases. Each system has its advantages and drawbacks, as stated in the above comparison, categorized by basic DL system characteristics and features. That comparison can only be used as a guideline by an organization in order to decide if one of these DL systems is suitable to host its digital collections.

Usually the needs for each organization vary depending on the number of collections, the types of objects, the nature of the material, the frequency of update, the distribution of content and the time limits for the development of a DL. In the next paragraphs, guidelines for the selection of a DL system are provided depending on different organization needs.

1. Consider a case where an institution or university needs a digital repository for research papers and dissertations produced by students and stuff. In that case, the most appropriate DL system is DSpace, since it by default represents communities (e.g. university departments) and collections (e.g. papers and dissertations), while workflow management supported is important for item submission by individuals.
2. Consider a case where an organization needs one digital collection to publish its digital content in a simple form, in strict time limits. In addition, the organization prefers to integrate the web interfaces of the DL with a portal like website. In that case the most appropriate DL systems is EPrints, since they separate the concerns of presentation and storage, are not bind to specific metadata standards and provide simple web interfaces for the submission and presentation of documents and metadata.
3. Consider a case where an organization wants to electronically publish books in an easy to use customizable DL system. In that case the most appropriate DL system is Greenstone, since it is easy to represent books in a hierarchical manner, using table of contents, while the full text of chapters can be searchable.

REFERENCES

[1]. Waters, Donald J. (1998). "What are digital libraries?" *CLIR Issues*. No. 4 http://www.clir.org/PUBS/issues/issues04.html#dlf

[2]. Wikipedia. (April 2006). "Digital Library" In *Wikipedia, the Free Encyclopedia*. http://en.wikipedia.org/wiki/Digital_libraries

[3]. Arms, W. (2000). Digital libraries. *Cambridge, MA: MIT Press*, pp.2.

[4]. C. Lagoze and H. Van de Sompel. The Open Archives Initiative: Building a low-barrier interoperability framework. In Proceedings of the Joint Conference on Digital Libraries (JCDL' 01), 2001.

[5]. DSpace Federation. Available at http://www.DSpace.org/

[6]. Greenstone Digital Library Software. Available at http://www.greenstone.org/

[7]. EPrints for Digital Repositories. Available at http://www.eprints.org/

[8]. R. Kahn and R. Wilensky. A Framework for Distributed Digital Object Services.

[9]. DCMI Metadata Terms. Dublin Core Metadata Initiative. Available at http://www.dublincore.org/documents/dcmi-terms/

[10]. METS: An Overview & Tutorial. Library of Congress. Available at http://www.loc.gov/standards/mets/METSOverview.v2.html