CHAPTER 3

Digital Repositories and Metadata

- Digital Library History
- What is a Digital Library?
- What is an Institutional/Digital repository?
- Definition Digital/Institutional Repository
- Institutional Repositories and Digital libraries
- Evolution of Digital Repositories (DR)
- The need for Digital Repositories
- Digital Library Tools
- Components of Digital Library System
- Open Source Digital Library Software
- Metadata in Digital Repositories
- Interoperability in Digital Repositories
- Conclusion
- References
Digital Repositories and Metadata

Digital Libraries evolved in the mid and late nineties and have become very important sources of information. They are distinctively more advantageous than just information on the web as digital libraries collect, organize and serve information in defined domains and through well proven knowledge organization systems and planned information services. In this chapter we present a background of digital libraries/repositories and trace their evolution. Presented also are important tools, projects and excerpts from popular works with an aim to bring into focus the relation of metadata and digital repositories and issues that arise in the context of the present work.

3.1 Digital Library History

The term "digital library" is the most recent in a long series of names for a concept that has been written about nearly as long as the development of the first computer: a computerized "library" that would supplement, adds functionality, and even replaces traditional libraries (Harter, 1996).

The domain of digital libraries though still evolving has really matured over the last 15 years. The roots of the present day digital libraries may be traced to the information retrieval systems of the 1960s and the hypertext systems of the 1980s. Digital libraries have evolved from the techniques and principles developed by early information retrieval researchers. Automatic indexing and search systems were pioneered in the 1960s; and today's digital libraries build on the solid foundations of more than four decades of research in information retrieval. However, the digital libraries as we know them today have been conceived and developed only since the 1990s (Fox and Urs, 2002).

Vannevar Bush wrote about the "memex," which is often cited as having stimulated much of the early application of computers to information retrieval. Although the memex was a mechanical device based on microfilm technology, it anticipated the idea of hypertext. Library automation began in the early 1950s with punched card applications to library technical services operations. Licklider coined the phrase "library of the future" to refer to his vision of a fully computer-based library, and ten years later, F.W. Lancaster wrote of the soon-to-come "paperless library." About the same time Ted Nelson invented and named hypertext and hyperspace. In the recent past, terms such as "electronic library," "virtual library," "library without walls," "bionic library," and others have been used. Karen Drabenstott has produced an excellent analytical survey in this regard (Harter, 1996).
Chapter 3: Digital Repositories and Metadata

The relatively recent use of the term "digital library" resulted from the Digital Libraries Initiative funded by the National Science Foundation, the Advanced Research Projects Agency, and the National Aeronautics and Space Administration in the United States. Impelled by the sudden explosive growth of the Internet and the development of graphical web browsers, in 1994 these agencies granted 24.4 million dollars to six U.S. universities for digital library research (Harter, 1996). The term was quickly adopted by computer scientists, librarians, and others. Thus, while the term "digital library" is new, work in bringing digitized information resources to libraries has a history spanning several decades.

As the information landslide now known as the 'web' grew, selectively libraries began choosing sites to classify or catalog. For better control and dissemination of the information on the web, digital library initiatives began sprouting up. The early digital libraries first began as a way of cataloging information that was already available via the web. As information technologies developed and matured, digital libraries soon had the capabilities to add their own content to the web. No longer were libraries simply cataloging what existed; digital libraries quickly turned into places of information storage and retrieval (Graham et al., 2005).

However, as Murray points out, problems quickly arose as to how to best store and access this information. Additionally, concerns grew as to what type of material was to be collected, included, and shared in the digital repositories (Graham et al., 2005).

Digital libraries have evolved rapidly over the past decade and are now as varied as physical libraries. Although digital libraries have been driven mainly by developments in technology, progress has also been made in addressing the intellectual and social issues involved in sharing knowledge in digital forms (Fox and Marchionini, 2001). Research libraries in several countries are leading development, including the USA (such as DSpace at MIT in collaboration with Cambridge University in the UK, the University of California’s eScholarship Repository, (see www.bepress.com/repositories.html), Canada (where CARL has developed a programme), and the UK (where the JISC Focus on Access to Institutional Resources (FAIR) Programme (www.jisc.ac.uk/index.cfm) has a number of projects under way).
3.2 What is a Digital Library?

There is very little agreement in the literature about what constitutes a digital library. Especially thoughtful treatments of this subject are by Miksa and Doty; and Levy and Marshall. One may insist on a relatively narrow definition -- based explicitly on the properties of the traditional print library -- or consider a continuum of broader possibilities. The most inclusive view takes a digital library to be, as its starting point, essentially what the Internet is today (Harter, 1996).

The variety of conceptions of what a Digital Library is, has had a substantive impact on attempts to define and bound the term 'Digital Library'. Since 2006 the term has been generally used to refer to systems that are heterogeneous in scope and provide diverse types of functionality. These systems include digital object and metadata repositories, reference-linking systems, archives, content administration systems (mainly developed by industry), and complex systems that integrate advanced digital library services (mainly developed in research environments). This "overloading" of the term 'Digital Library' results in digital library services and systems that do not deliver interoperability and reuse of content and technologies (Candela et al., 2007).

There seems to be a belief that a digital library is just about search ("can I find it?") and access ("can I get it?"). These functions are indeed essential (and remain challenging), but they are just part of an information environment. Traditional libraries are much more than well-organized warehouses of books, maps, serials, etc. In their full expression, they are places where people meet to access, share, and exchange knowledge. The resources they select and services they offer should reflect the character of the communities they serve (Lagoze et al., 2005).

As suggested by Borgman, digital libraries should match and indeed dramatically extend traditional libraries and they need to be much more than search engines. Like all types of libraries, they should feature a high degree of selection of resources that meet criteria relevant to their mission, and they should provide services, including search, that facilitates use of the resources by their target community. But, freed of the constraints of physical space and media, digital libraries can be more adaptive and reflective of the communities they serve. They should be collaborative, allowing users to contribute knowledge to the library, either actively through annotations, reviews, and the like, or passively through their patterns of resource use. In addition, they should
be contextual, expressing the expanding web of inter-relationships and layers of knowledge that extend among selected primary resources. In this manner, the core of the digital library should be an evolving information base (Lagoze et al., 2005).

Digital Library Federation (1998) has come up with the following working definition of the digital library: “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”.

What exactly constitutes a digital library has never been easy to pin down but we often come across the terms ‘Digital library’ (as explained above) and the term ‘Digital/Institutional repository’ in the professional literature. Although they serve the same purpose of storing and providing access to digital content, it may be advisable to examine if they are any different from each other.

3.3 What is an Institutional/Digital Repository?

As mentioned earlier, Digital Libraries/Repositories are far from well defined. Definitional agreement may only extend to notions of accessible collections of information.

The development of Institutional Repositories (IRs) has emerged as a new strategy that allows universities and other organizations to apply serious, systematic leverage to accelerate changes taking place in scholarship and scholarly communication, both moving beyond their historic relatively passive role of supporting established publishers in modernizing scholarly publishing through the licensing of digital content, and also scaling up beyond ad-hoc alliances, partnerships, and support arrangements with a few select faculty pioneers exploring more transformative new uses of the digital medium (Lynch, 2003).

In the words of Lynch, many technology trends and development efforts came together and contributed to the development of digital repositories. Online storage costs have dropped significantly; repositories are now affordable. Standards like the open archives metadata harvesting protocol are now in place;
progress has also been made on the standards for the underlying metadata itself (Lynch, 2003).

### 3.4 Definition Digital/Institutional Repository

An institutional repository is the collective intellectual output of an institution recorded in a form that can be preserved and exploited (Yeates, 2003). Repositories are the key to the ability of institutions to respond to future needs for more dynamic cross-boundary communications and services. Definitions of institutional repositories range from the broad and inclusive to the narrower (Westell, 2006).

According to Richard Johnson's definition: A digital archive of the intellectual product created by the faculty, research staff, and students of an institution and accessible to end-users both within and outside of the institution with few, if any, barriers to access (Johnson, 2002).

SPARC identifies four IR characteristics: "institutionally defined; scholarly; cumulative and perpetual; open and interoperable" (Crow, 2002).

Crow (2002), has made a substantial case for IRs in a position paper for SPARC (www.arl.org/sparc/), which defines them, as: "digital collections capturing and preserving the intellectual output of a single or multi-university community". Crow further states that: ...content may include pre-prints and other works-in-progress, peer-reviewed articles, monographs, enduring teaching materials, data sets and other ancillary research material, conference papers, electronic theses and dissertations, and gray literature.

According to Lynch (2003): A mature and fully realized institutional repository will contain the intellectual works of faculty and students – both research and teaching materials – and also documentation of the activities of the institution in the form of records of events and performance and of the ongoing intellectual life of the institution.

IRs go beyond discipline-based eprint servers. Many fairly standard content management functions are needed for IRs, including access control mechanisms, updating and information retrieval. Key aspects are facilities for easy submission of content by authors and academics, appropriate workflows for institutional purposes and sufficiently powerful, open standards-based dissemination functionality to attract contributors within a largely voluntary
implementation. The systems also require unusual levels of scalability and methods of ensuring the integrity of content so that users trust the repository for the long term (Yeates, 2003).

3.5 Institutional Repositories and Digital Libraries
If we look at the concept of institutional repositories within the broader context of digital libraries based on the definitions of both, 'Digital libraries' can mean many things, but they are libraries first and foremost, and are built upon the enduring principles of information management. Libraries are themselves repositories, and have always dealt in the management of repositories for their users. With libraries now routinely managing repositories of various types in digital format, what does it mean to qualify as a 'repository' with 'institutional'? (MacColl et al., 2006).

It is tempting to use the term 'digital library' about any collection of digital objects which have some means of navigation and retrieval, but approaching the question using Ranganathan’s Laws would suggest that a collection of items is not a digital library merely by virtue of the items being digital. Rather, a digital library first needs to have virtues of being a library first. An academic or research library is organized for use in the pursuit of human advancement. The fact that its contents – or a large proportion of them – are digital is merely an accident of history. Digital libraries are, therefore, much more than aggregations of documents on the World Wide Web. A digital library, a collection of information that is both digitized and organized, gives us powers we never had with traditional libraries (MacColl et al., 2006).

An institutional repository is a digital library of local digital content including "grey" literature. Grey literature "is any documentary material that is not commercially published and is typically composed of technical reports, working papers, business documents, and conference proceedings" (Smith, 2008).

When we talk about the 'institutional repository', we use ‘institution’ to refer to the educational or research establishment, which is the library’s parent body. Institutional repositories have emerged from universities, but are spreading into other types of educational organizations too, such as colleges and research institutes. However, research repositories until quite recently were based only around disciplines (MacColl et al., 2006).
Academic libraries have begun to see the need to develop institutional digital libraries alongside subject-based digital libraries. The institutional library needs a presence on the Web – a place to describe its print and web-based services, and to bring together the content it makes available to its users. Digital libraries, then, belong both to knowledge domains and to institutions, in the same way, as do repositories, which are constituent elements of each.

**Examples of Digital Libraries and Repositories**

<table>
<thead>
<tr>
<th>Digital libraries</th>
<th>Repositories</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disciplinary</strong></td>
<td></td>
</tr>
<tr>
<td>Alexandria Digital Library</td>
<td>arXiv</td>
</tr>
<tr>
<td><a href="http://www.alexandria.ucsb.edu">www.alexandria.ucsb.edu</a></td>
<td><a href="http://www.arxiv.org">www.arxiv.org</a></td>
</tr>
<tr>
<td>Perseus Project</td>
<td>PubMed Central</td>
</tr>
<tr>
<td>Digital Scriptorium</td>
<td>EconPapers</td>
</tr>
<tr>
<td><a href="http://sunsite.berkeley.edu/Scriptorium/">sunsite.berkeley.edu/Scriptorium/</a></td>
<td><a href="http://econpapers.repec.org/">econpapers.repec.org/</a></td>
</tr>
<tr>
<td>Center for Electronic Texts in the Humanities</td>
<td>CogPrints</td>
</tr>
<tr>
<td><a href="http://www.ceth.rutgers.edu/">www.ceth.rutgers.edu/</a></td>
<td><a href="http://cogprints.org">cogprints.org</a></td>
</tr>
<tr>
<td><strong>Institutional</strong></td>
<td></td>
</tr>
<tr>
<td>California Digital Library</td>
<td>Edinburgh Research Archive</td>
</tr>
<tr>
<td>Illinois Digital Academic Library</td>
<td>DSpace at MIT</td>
</tr>
<tr>
<td><a href="http://www.idal.illinois.edu/">www.idal.illinois.edu/</a></td>
<td><a href="http://DSpace.mit.edu/index.jsp">DSpace.mit.edu/index.jsp</a></td>
</tr>
<tr>
<td>LSE Research Articles Online</td>
<td></td>
</tr>
<tr>
<td><a href="http://eprints.lse.ac.uk/">eprints.lse.ac.uk/</a></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: The institutional repository in the digital library (MacColl et al., 2006)

Table 3.1 breaks down both digital libraries and repositories by institution and discipline. The libraries, on the left, depend more and more on the repositories, on the right, to provide them with the selections of collections they present as libraries, whether institutional or disciplinary.
3.6 Evolution of Digital Repositories (DR)

A DR provides access to the content and materials produced by members of that institution, typically a university or other educational establishment. The impetus for the DRs was boosted by the Open Archives Initiative (OAI) in 1999. OAI introduced the Protocol for Metadata Harvesting (OAI-PMH) that enables compliant sites to be interoperable, thus making institutional, rather than only disciplinary repositories visible and viable. OAI was aimed initially at eprint archives but was soon applied to other digital library content (Hitchcock et al., 2007).

Rather than the number and size of repository initiatives, the availability of software and its interoperability are key indicators of the maturity of the field. In the late 1990s, repositories were often homegrown and built from scratch. With the dawn of the 21st century, several initiatives started to share their software source code openly (Aschenbrenner et al., 2007).

Aschenbrenner et al. (2007), argue that the majority of repository communities have coalesced around a trinity of nodes (as prominently represented in the Open Repositories conference series): these nodes are EPrints (University of Southampton), DSpace (Massachusetts Institute of Technology), and Fedora (Cornell University and the University of Virginia). Overall, the availability of repository software stabilized extremely quickly. However, the repository universe is not limited to this trinity, and the current equilibrium may change at any time as novel repository-like systems emerge in related fields such as e-Infrastructure (e.g., iRODS, gCube), Semantic web platforms (e.g., Talis Triple Store, Tupelo), commercial companies (e.g., Microsoft, IBM), content-type specific repositories (e.g., Flickr, YouTube, SlideShare), and the so-called Web 2.0.

Digital repositories have evolved rapidly over the last few years. By June 2009, the OpenDOAR directory – a portal for voluntary repository registration – listed more than 1,400 operational repositories covering a range of contexts and tasks. The "hype curve" metaphor used below, which characterizes the adoption cycle of new technologies, shows what may be in store for the future of repositories (Aschenbrenner et al., 2007).
After the steep rise of repositories as dissemination platforms in the early part of this century, the lack of user adoption has posed a severe challenge to the repository community. This made repository evolution dip towards the "trough of disillusionment" (see Figure 3.1). That is because the benefits of repositories were not made sufficiently clear to the scholarly user. To increase the value of the repository to the user, some repositories are focusing more on features such as: preservation of the user's intellectual assets (read: trusted digital repositories), scientific collaboration through reuse (publications as well as primary data), and embedding repositories into the user's scientific workflows and scholarly workbench (Aschenbrenner et al., 2007).

3.7 The need for Digital Repositories
Scholarly communications are being restructured for the digital environment. Already much discussion about the future of scholarly publishing has resulted in widespread experiments in open archiving. These have been supported by development of the OAI Protocol for Metadata Harvesting.
Knowledge, or intellectual capital, produced by scholars and their parent institutions (universities) tends to end up being disseminated by commercial publishers, who are required to undertake the work of identifying and selecting appropriate material for capturing, preservation and publication. Output from individual universities is dispersed to publishers and then to library collections. These are increasingly limited in what they can afford to purchase. The Internet enables the principal functions of scholarly communication to be unbundled, giving rise to new ways of sharing knowledge and new opportunities for institutions to use their intellectual capital as a more effective indicator of academic quality (Yeates, 2003). The concept of digital/institutional repository allows the institutions to record their collective intellectual output in a form that can be preserved and exploited (Madalli, 2003).

3.8 Digital Library Tools
Digital libraries have evolved from being simply online collections of articles or resources to complex systems that offer a range of services. Building digital libraries involves the integration of complex systems, including collections of documents with varied structure, media, and content (with diverse types such as standard metadata elements like "title," and distributions of interrelationships such as direct references). Add to that mix, hardware and software components interoperating across different data structures and processing algorithms and multiple people, communities, as well as institutions with different goals, policies, and cultures (Madalli, 2003).

The challenges of integration and interoperation are faced in corporate, government, and academic settings. Solutions involve building upon the experience of traditional communities of practice to cooperate and leverage information technology (Fox and Marchionini, 2001). One of the main challenges is the expectation of information professionals to suddenly learn and respond with relevant services in the onslaught of digital resources and online services. There is not much consideration or time allowed for the learning process. Another challenge is getting content into digital repositories. While the technology itself is relatively easily available and fairly easy to handle, content creation and collection is quite an arduous and expensive effort. Another
important issue is rights management and selective authorization of the people related to digital repositories. These are the main tasks any DL initiative has to first deliberate upon, sort out and plan for with unambiguous policy support.

Once the policy decisions are made regarding the content, authors, etc, other factors such as choice of the type of system to be adapted, technological and other criteria should be drawn up. The expected features of digital library software can be enlisted as follows:

- Low cost, (including all hardware and software components)
- Technically simple (to install and manage)
- Robust
- Scalable
- Open and inter-operable
- Modular
- User Friendly
- Multi-user (including both searching and maintenance)
- Multimedia digital object enabled; and
- Platform independent (including both client and server components)

3.9 Components of Digital Library System

Planning and implementation of digital libraries poses new challenges and involves policy making regarding the members, content, content management, governance, maintenance and the technical know-how. DLs are data intensive and complex and hence require robust software to manage the data and handle user interactions. Machine centric design of digital library software has given way to user centric design. Digital libraries comprise of components that handle repositories building, maintenance, searching and retrieval and distributed information services in a networked environment. The components of digital libraries include the following:

- Resources/Collection
  - Types of resources
  - Formats
  - Multilingual Collection

- User interface
  - Memberships/login
  - Submission
  - Search/save
  - Alert system
Chapter 3: Digital Repositories and Metadata

- Metadata Standard and Interface
- Indexing
- Search/Browse/Display facility
- Interoperability Standards
- Import/Export modules
- Persistent Identifiers
- Authentication and Rights/Access Management
- Network elements
3.10 Open Source Digital Library Software
Fortunately digital libraries have generated a lot of interest and the open source development community has offered a few packaged systems to develop and maintain digital repositories (Madalli, 2003).
Open source software has been a nebulous reference to any software that is free, and is often confused with freeware and shareware. The Open Source Initiative (OSI; www.opensource.org) has therefore become a certification body for open source software under a commonly agreed-upon definition for "open source". Highlights of the OSI's definition of open source include: free distribution and redistribution of software and source code; licenses that allow distribution of modifications and derived works and non-discrimination against persons, groups or fields of endeavor (Goh et al., 2006).

The availability of the source code in open source software allows users to modify and make improvements to it, and such contributions could come from a diverse talent pool of programmers. Further, because source code is accessible and modifiable, contributions also lead to improvements in the functionality of the software. In addition, updates can be obtained at low or even no cost and there are no royalties or license fees. There is less likelihood of being dependent on a single software provider or being trapped into long-term software support contracts, which restricts flexibility in implementation.

Some initiatives include Andrew Mellon Foundation funded Flexible Extensible Digital Object and Repository Architecture (Fedora) Project at the University of Virginia (www.fedora.info/), Caltech Collection of Open Digital Archives (CODA – http://library.caltech.edu/digital/). Given that one key aim of DRs is to make management easier for users, it is likely that the application service provider (ASP) model will be relevant. It speeds up institutional development, allows experimentation and drives innovation while fostering integration with commercial services. The ebrary Institutional Repository Pilot Program (www.ebrary.com/libraries/ir.jsp), for example, provides a hosted solution for ebrary customers, offering critical researcher-led benefits. However, for such a model to work, institutions would need to trust their service providers to maintain service persistently, or at least to be able to transfer the repository elsewhere if they cease to do so, and it is therefore likely that interoperable repositories are needed before such options are widely adopted. Such interoperability would also strengthen the long-term security of repositories generally, beyond any federation using a single technical platform.

Some of the popular DL packages such as Greenstone Digital Library Software from University of Waikato, New Zealand, DSpace Digital Library Suite from HP-MIT libraries collaboration, Fedora (Flexible Extensible Digital Object and
Chapter 3: Digital Repositories and Metadata

Repository Architecture) from Cornell and Virginia University joint project and EPrints from University of Southampton are being discussed below:

3.10.1. Greenstone Digital Library Software (GSDL-
http://www.greenstone.org)

3.10.1.1 Software distribution and development

Greenstone is issued under the terms of the GNU General Public License. It originated in 1996, and the current production version (Greenstone2) was designed in 2000 although it is continually being extended. A complete redesign and reimplementation, Greenstone3, has been described and released. Greenstone3 allows documents to be dynamically added to collections; provides more flexible ways to dynamically configure the run-time system by adding new services; lowers the overheads incurred by collection developers when accessing this flexibility to organize and present their content; and modularizes the internal structure. The design is based on widely accepted standards that were unavailable when Greenstone2 was designed (Witten and Bainbridge, 2005; Madalli and Seth, 2005)

3.10.1.2. Features of Greenstone Digital Library Software:

- **Ease of use:** The Greenstone software is intended to help people design and build collections quickly and easily.

- **Customizable:** The facilities that a collection provides and the user interface for searching and browsing, are highly customizable at many different levels. Users can easily specify what document formats will be included (e.g. HTML, Word, PDF, PostScript, PowerPoint, Excel); where already-available metadata (if any) comes from (e.g. XML files, OAI archives, Latex bibliographies); what searchable indexes will be provided (e.g. full text, perhaps differentiated by language, and certain metadata such as titles); and what browsing structures will be available (e.g. list of authors, titles, classification hierarchy).

- **Browsing:** Browsing capabilities vary from collection to collection depending on the metadata available and the facilities that the collection designer wishes to provide. Greenstone includes predefined browsing structures based on certain kinds of metadata. Any textual metadata can be presented as an alphabetically sorted list, which can optionally be tabbed into alphabetic ranges that are chosen automatically to include a reasonable number of documents in each range. Date metadata can be presented in a list
that allows selection by year and month. Metadata with hierarchical structure, such as library classifications, can be presented as a tree whose nodes open to reveal the data beneath. In this case the user must provide an auxiliary file giving labels for intermediate nodes of the hierarchy (e.g. subject headings corresponding to each classification number).

- **Searching:** Searching the full text of all documents in the collection is a basic facility, included by default in all collections. Collection designers can determine whether searching should be on a paragraph, section, and/or whole-document level (this affects the scope of matches to a given query). They can also ask for full-text indexes to be built on metadata items (e.g. titles, authors). They can split the collection into sub-collections that can each be searched individually, or use language metadata to restrict searches by language.

- **Formatting presentations:** Formatting statements are used to control the presentation to the user of each “screen” that the system generates. They determine how target documents are displayed—whether they are preceded by title, for example, or indented. They control the search results page, where they determine what metadata is presented as a “snippet” that stands for matching documents, whether it should be preceded by an appropriate document icon, whether there should be a hyperlink and what is its target. In collections that include different versions of a document (e.g. Word and HTML extracted from it), link to both versions can be presented in the search results list so that users can determine which one to read.

- **Multilingual:** Greenstone is multilingual. Currently interfaces in Arabic, Chinese, Czech, Dutch, French, Galician, German, Hebrew, Indonesian, Italian, Japanese, Kazakh, Maori, Portuguese, Russian, Spanish, Thai, Turkish and English languages are available (Madalli and Seth, 2005).

3.10.2. *Eprints* (http://software.eprints.org)

Current version: GNU EPrints 3.1.3

GNU EPrints is generic archive software developed by the University of Southampton. It is intended to create a highly configurable web-based archive. GNU EPrints primary goal is to be set up as an open archive for research papers, and also other objects such as images, research data, audio archives - anything that can be stored digitally, with some changes to the configuration. The EPrints software has been developed under GNU/Linux. It is intended to
work on any GNU system. It may well work on other UNIX systems too. Other systems that EPrints is running on include Solaris and Mac (OS/X) (Madalli and Seth, 2005)

3.10.2.1. Features of GNU EPrints

- Standard installation via "configure"
- Very configurable and adaptable. It is possible to add new tools and scripts using the modules provided.
- Can store documents in any format that archive administrator wishes. Each individual research paper (or eprint) can be stored in more than one document format.
- The archive can use any metadata schema; the administrator decides what metadata fields to be held about each eprint. This is decided in three or four stages:
  - Decide a maximal set of metadata fields that should be stored (for example, "authors", "title", "journal", "journal volume", etc.)
  - Decide what types of eprint should be stored (for example, refereed journal article, thesis, technical report, unpublished preprint)
  - For each eprint type, decide which metadata fields should be stored for Eprints of that type, and which of those fields are mandatory.
  - Decide how these metadata fields should be projected into the Open Archives world.
- GNU EPrints can be placed in a configurable, extendable subject hierarchy. This hierarchy can be used to view and search the archive.
- Submission of all eprints is via a simple but extremely powerful WWW-based interface. Papers can be uploaded just as files, in a compressed bundled file (such as a .zip file), or automatically mirrored from an existing website by specifying a URL.
- Data integrity checks are performed automatically without the need for administrator intervention. Some "stub" routines allow individual sites to add their own integrity checks if they desire. These checks can greatly reduce the workload of the site administrator.
- Authors can have associated metadata. Again, the site administrator can decide what metadata fields will be stored with each author's record.
• Users can subscribe either as authors or readers, via a web form or an automatically processed e-mail account.
• Submitted papers (if administrators desire) go through a moderation process. Submitted papers are placed in a buffer, where they can be approved by a moderator, rejected outright, or returned to the author for amendment.
• Moderation process is also performed using a WWW-based interface. In fact, once the site is up and running, it will require very little maintenance that cannot be performed using the WWW-based maintenance interface.

3.10.3. Fedora Digital Library (http://www.fedora.info/index.shtml)
Fedora was developed as a research project at Cornell University, and successfully implemented at University of Virginia as a prototype system to provide management and access to a diverse set of digital collections. The Fedora project funded by the Andrew W. Mellon Foundation to build an open-source digital object repository management system based on the Flexible Extensible Digital Object and Repository Architecture (FEDORA).

Fedora is a general-purpose digital object repository system that can be used in whole or part to support: institutional repositories, digital libraries, content management, digital asset management, scholarly publishing, and digital preservation. The Fedora repository system is open source software licensed under the Mozilla Public License. The documentation states the following features:

3.10.3.1 Features and Modules of Fedora (Fedora Project, 2005)
• Web Services: The interface to the Fedora repository system consists of three open APIs that are exposed as web services: Management API known as API-M, Access API known as API-A, and Access-Lite API known as API-A-Lite.

• Flexible Digital Object Model: The Fedora digital object model provides the flexibility to create many different design patterns for digital objects. The Fedora digital object acts as a container for Datastreams (bitstreams of content or metadata) and Disseminators (linkages to services for transformation of content or computation).
Chapter 3: Digital Repositories and Metadata

- **Datastreams**: A Datastream is the component of a Fedora digital object that represents content and metadata. The content may physically reside inside the repository, or it may reside outside the repository and be pointed to (by-reference). The Fedora repository system supports content of any MIME type.

- **Default Disseminator**: A Disseminator associates a set of behaviours with a digital object. These behaviours expose a set of service methods that are available on the digital object. The Default Disseminator is a built-in disseminator that is available on every object. It provides a generic set of service methods for viewing the contents of a digital object and for reflecting on other services that may be available for the object.

- **Extensible Disseminators**: Fedora provides the ability to associate custom disseminators with digital objects to expose a set of behaviours beyond those offered by the Default Disseminator. Extensible Disseminators provide a means of associating service methods with digital objects to transform content, run computations on content, or present custom renderings of content. Service methods that are associated with an object can support user-supplied parameters that are handled at run time.

- **Content Versioning**: The Fedora content versioning system is enabled with the release of Fedora 1.2. Now, any modifications made to a Datastream or Disseminator through the Fedora management interface (API-M) will automatically result in a new version of that Datastream or Disseminator being created by Fedora. The Fedora repository maintains all versions of all Datastreams and Disseminators, thereby creating a history of how objects change over time. Additionally, Fedora maintains an audit trail record of the nature of the object change events (e.g., who, what, when, why).

- **XML Ingest and Export**: Digital objects can be submitted to a Fedora repository as XML-encoded files that conform to an extension of the Metadata Encoding and Transmission Standard (METS) schema. The schema for the extended version of METS used by Fedora can be found at http://www.fedora.info/definitions/1/0/mets-fedora-ext.xsd. Digital Objects can also be exported from the repository in this format.

- **Digital Object Storage**: The Fedora repository stores digital objects in XML. A Fedora repository also maintains a cache of its digital objects in a
relational database. The database stores the most current view of every digital object. The database cache is used to enhance run-time performance for requests made on the Fedora Access web service (API-A).

- **Access Control and Authentication**: Release 1.2 includes a simple form of access control to provide access restrictions based on IP address. IP range restriction is supported in both the Management and Access APIs. In addition, the Management API is protected by HTTP Basic Authentication.

- **Searching**: Fedora automatically indexes the primary Dublin Core record for each digital object, as well as selected Fedora-specific metadata fields. The Fedora repository system provides a search interface as part of the Fedora Access API (API-A).

- **OAI Metadata Harvesting Provider**: Every Fedora digital object has a primary Dublin Core record that conforms to the schema at: http://www.openarchives.org/OAI/2.0/oai_dc.xsd. This metadata is accessible using the OAI Protocol for Metadata Harvesting, v2.0.

- **Fedora Administrator Client**: The Fedora repository system includes a Java Swing client for managing the repository. Fedora Administrator can be used to search/browse the repository, create new digital object, modify existing digital objects, ingest objects into the repository and export objects, and create/ingest batches of objects.

- **Migration Utility**: A migration utility is provided to perform mass export and mass ingest of objects. At the core, the migration utility is built upon two newly enhanced command-line functions: fedora-export and fedora-ingest. Used together, these two command line functions can support a variety of scenarios involving moving or copying objects between repositories. The utility is general-purpose in that it can be used to copy or move objects among repositories for many reasons including upgrading from previous releases of Fedora. Client access to the migration comes in two forms: (I) run fedora-export and fedora-ingest at the command line or (II) use the Fedora Administrator client to invoke these functions.

- **Batch Utility**: The Fedora repository system includes a Batch Utility as part of the Fedora Administrator client that enables the mass creation and loading of data objects.
3.10.4. DSpace Digital Library (http://www.DSpace.org)

Latest version: DSpace 1.5.2

DSpace was initially developed by Hewlett-Packard and MIT in collaboration with the objective to create a package that could provide an institutional repository, which addressed the problem of digital preservation as a central theme. A number of individuals from institutions using DSpace have taken on the role of developers, and a community of interested parties has evolved who have started to feed code back into the core. DSpace future development and implementation is part of the open-source development model.

The application itself provides ways of capturing, storing, indexing, preserving and disseminating digital objects. DSpace repositories may include, but are not limited to, research papers, conference papers, book chapters, datasets, learning objects and, of course, E-Theses.

DSpace facilitates the following:

- Captures
  - Digital research material in any format directly from creators
- Describes
  - Descriptive, technical, rights metadata
  - Persistent identifiers
- Distributes
  - Searches metadata
  - Delivers via Web, with necessary access control
- Preserves
  - Large-scale, stable, managed long-term storage

DSpace has incorporated features required for sophisticated and integrated operations of digital libraries. Accordingly its functions include comprehensively all activities form collection planning building and web based information services.

3.10.4.1 DSpace features

- **Data Model**: This is the information model according to organizational structure. Each DSpace site is divided into communities; these typically correspond to a laboratory, research center or department. Typically in an academic institution it is by the various departments or faculties of study. As of DSpace version 1.2, these communities can be organized into a hierarchy. Communities contain collections and each collection further
comprises *items*. *Items* are the basic archival elements. Each *item* is owned by one collection though it may be mapped to others. *Items* are further divided into bitstreams that are computer files.

- **Metadata**: Metadata in DSpace is of three types:
  - **Descriptive**: DSpace uses qualified Dublin Core metadata elements and provides worksheets in the submission process. It allows the DL manager to select the subset of elements as per requirements or to add more elements.
  - **Administrative**: This includes preservation metadata, provenance and authorization policy data.
  - **Structural**: This includes information about how to present an item, or bit streams within an item, to an end-user, and the relationships between constituent parts of the item.

- Ingest Process and Workflow and reviewing: In the open access repositories to ensure quality of the content a strict review process should be put in place so that material in the open access will gain the same credibility as that of “ranked journals”. DSpace allows defining a workflow for the reviewing process. The workflow is setup for each collection. Reviewers can be assigned for reviewing the contents as well as for correcting the metadata of the items.

- Search and Browse: Browsing facility in DSpace allows the users to browse by communities and collections (hierarchies are indicated in the community homepage). Users may also browse by the authors and titles (alphabetical index) and by date across collections or within a particular collection. DSpace incorporates Jakarta’s Lucene search engine. In addition to normal search features such as Boolean, wildcard searches, it has advanced features of searching such as the Levenshtein Distance algorithm for fuzzy logic and term boosting (assigning weights to words in a query).

- OAI Support: DSpace is OAI version 2 compliant and allows harvesting the records in XML carrying the DC elements.

- OpenURL Support: OpenURL support enables opening an item from a referred link to it. DSpace will display an OpenURL link on every item page, automatically using the Dublin Core metadata.
• Import and Export: Records can be imported into and exported from DSpace repositories. DSpace includes batch tools to import and export items in a simple directory structure, where the Dublin Core metadata is stored in an XML file.

3.11 Metadata in Digital Repositories
A metadata schema consists of a set of elements designed for a specific purpose, such as describing a particular type of information resource. As defined in the report of the American Library Association Committee on Cataloging: Description and Access (CC: DA) Task Force on Metadata: "A metadata schema provides a formal structure designed to identify the knowledge structure of a given discipline and to link that structure to the information of the discipline through the creation of an information system that will assist the identification, discovery, and use of information within that discipline" (Chan and Zeng, 2006a).

In the literature, the words "schema", "scheme", and "element set" have been used interchangeably to refer to metadata standards. In practice, the word "schema" usually refers to an entire entity including the semantic and content components (which are usually regarded as an "element set") as well as the encoding of the elements with a markup language such as SGML (Standard Generalized Markup Language) and XML (Extensible Markup Language). A metadata element set has two basic components:

- Semantics – definitions of the meanings of the elements and their refinements.
- Content – declarations or instructions of what and how values should be assigned to the elements.

For each element defined, a metadata standard usually provides content rules for how content should be included (for example, how to identify the main title), representation rules for content (for example, capitalization rules or standards for representing time), and allowable content values (for example, whether values must be taken from a specified controlled vocabulary or can be author-supplied, derived from text, or added by metadata creators working without a controlled term list.)

It has been observed that many metadata standards provided an element set without considering the encoding format in their preliminary versions. For example, Dublin Core (DC), VRA (Visual Resource Association) Core
Categories, Categories for the Description of Works of Art (CDWA), and the Learning Object Metadata (LOM- http://ltsc.ieee.org/wg12/index.html) were all published and accepted in terms of their semantics and content long before the specific encoding methods for their data models were published. On the other hand, a few other metadata standards, like the Encoded Archival Description (EAD -http://lcweb.loc.gov/ead) Document Type Definition (DTD), provided an encoded element set from the beginning. The EAD DTD, a standard for encoding archival finding aids currently using XML, was published a decade ago with an SGML DTD.

3.11.1 Metadata in DSpace

DSpace uses a qualified Dublin Core metadata standard for describing items intellectually (specifically, the Libraries Working Group Application Profile) DSpace supports a hierarchical form of metadata that can be attached at the document level. By default, only three fields are required: title and submission date (if item published earlier), all other fields are optional. There are additional fields for document abstracts, keywords, technical metadata and rights metadata, among others. This metadata is displayed in the item record in DSpace, and is indexed for browsing and searching the system (within a collection, across collections, or across Communities). For the Dissemination Information Packages (DIPs) of the OAIS framework, the system currently exports metadata and digital material in a custom XML schema while working with the METS community to develop the necessary extension schemas for the technical and rights metadata about arbitrary digital formats (Smith et al., 2003).

DSpace supports content/metadata import feature. If data has to be imported from other DSpace repository then it requires collection Id, eperson, source file etc. in XML format. In case the data is available in Dublin Core format, DSpace supports import of metadata/digital objects. The system supports importing of bulk metadata/contents of digital objects but it does not import automatic metadata for the digital objects that are added into the repository. The software supports content/metadata export of the collections or items, which are available in the repository (Prasad and Madalli, 2003).

DSpace also supports exporting data into METS format. The contents of the METS are in mets.xml file format. Generally data is exported in XML file format from DSpace. DSpace supports to add descriptive metadata for all digital objects that are added into the repository. DSpace by default supports the following metadata fields such as: Author, Title, Other Title, Date of Issue,
Chapter 3: Digital Repositories and Metadata

Publisher, Citation, Series/Report No., Identifier, Type, Language, Subject, Keywords, Abstract, Sponsors, Description fields.

DSpace allows addition of descriptive metadata (such as author, title, subject, publisher, copyright, year of publication etc. as well as administrative metadata (such as technical aspects of digital documents, source of information (when and how it was created, rights management, what is the file size, what file format digital document is in, what is needed to view digital document etc.) for each digital object that gets added into the repository. It also supports preservation metadata, provenance and authorization policy data. Most of this data is held within DSpace's relational database schema. Provenance metadata is stored in Dublin Core records. Structural metadata (such as description, owner, data type, date deposited, version number, data of last revision, how compound objects are put together, how pages are ordered to form chapters etc.) is also provided in DSpace.

The software supports UTF-8 (Unicode character set) for metadata entry and allows addition/editing of metadata fields. Help is available in the form of help messages given in each metadata field while entering data. DSpace has input forms in xml files, which can be used to customize metadata fields as per the end user’s requirements. DSpace also has default metadata entry templates and all the metadata formats are documented. It can be customized for the creation of different interface for metadata entry, while submitting documents. DSpace allows the verification of all the metadata, and digital object before moving them into the repository. However, there are no automated checks of the metadata, such as verifying that a date entered into a field is really a date string available in the current version. DSpace also allows for metadata extensibility and complexity and supports metadata versioning.

DSpace supports controlled vocabulary, which at present is not quite compliant with W3C standards like OWL. It is OAI-PMH (Protocol for Metadata Harvesting) compliant. Since every item that is added into the DSpace repository has one Dublin Core descriptive metadata record, basic metadata interoperability across all the items in the repository is possible. The system supports automated content acquisition, harvesting and automatic metadata generation. Metadata is stored in PostgreSQL database and actual files are stored in “assetstore” directory where DSpace is installed.
3.11.2 Metadata in Eprints

The archive can use any metadata schema; the administrator decides what metadata fields to be held about each Eprint. This is decided in three or four stages:

- Decide a maximal set of metadata fields that should be stored (for example, "authors", "title", "journal", "journal volume", etc.)
- Decide what types of eprint should be stored (for example, refereed journal article, thesis, technical report, unpublished preprint)
- For each eprint type, decide which metadata fields should be stored for eprints of that type, and which of those fields are mandatory.
- Decide how these metadata fields should be projected into the Open Archives world.

3.11.3 Metadata in GSDL

Many popular document and metadata standards are incorporated into Greenstone. It can deal with documents in HTML, Word, PDF, PostScript, PowerPoint, and Excel formats (amongst others); images in TIFF, GIF, PNG, and JPEG formats (amongst others); and metadata in Dublin Core, MARC, CDS/ISIS, and ProCite formats (amongst others). It can deal with multimedia formats such as MP3, MIDI, and QuickTime. Greenstone's type of approach to document and metadata standards creates many demands for conversion facilities. For example, users can change metadata elements from one metadata scheme to another by making choices interactively as they drag documents from one collection to another, or in other circumstances they can either use a default mapping to convert, for example, MARC records to Dublin Core, or define their own crosswalk file (Witten and Bainbridge, 2005).

Emerging digital library standards are also supported. Greenstone incorporates web-mirroring software, so that whole sites can be downloaded using the HTTP protocol, to a pre-specified depth, and ingested into a collection. Metadata (and documents, if appropriately referenced) from an Open Archives Initiative (OAI) server can easily be ingested too, and any Greenstone collection can be served over the OAI protocol for metadata harvesting (OAI-PMH). Greenstone collections can be exported into the METS Metadata Encoding and Transmission Standard, and METS collections can be imported into Greenstone. The particular form that Greenstone uses has been submitted to the METS Board as a proposed METS Profile (Witten and Bainbridge, 2005).
In Greenstone, one or more metadata sets are associated with each collection. There are a few pre-prepared sets, of which Dublin Core is one. Modifications to existing sets and new ones can be defined using an auxiliary Greenstone application called GEMS (Greenstone Editor for Metadata Sets). One important set is the extracted metadata set, which contains information extracted automatically from the documents themselves (e.g. HTML Title tags, meta tags, or built-in Word author and title metadata). This is always present behind the scenes, though it may be hidden from the user. The system keeps metadata sets distinct using namespaces. For example, documents can have both a Dublin Core Title (dc.Title) and an extracted Title (ex.Title); they do not necessarily have the same value. Behind the scenes, metadata in documents, and metadata sets themselves, are represented in XML (Witten and Bainbridge, 2005).

Metadata in Greenstone can be a simple text string (e.g. title, author, and publisher). Or it can be hierarchically structured, as with hierarchical classification values, in which case new values can be placed in the classification tree (Witten and Bainbridge, 2005). In addition, it is multivalued: each element can have more than one value. This is used, for example, for multiple authors. The Librarian interface allows existing metadata values to be reused where appropriate, encouraging consistency in metadata assignment by eliminating the need to retype duplicate values. In Greenstone each item of metadata is flat, but metadata can be attached to individual sections within a document.

### 3.12 Interoperability in Digital Repositories

Recent decades have witnessed a proliferation of metadata schemas for description of digital resources. Each has been designed based on the requirements of the particular user community, intended users, type of resources, depth of description, etc. Problems arise when building a large digital library or repository with participants using different description methods or metadata records prepared according to diverse schemas. The diversity of standards for description of resources of many different types poses particular challenges to the users as well as for those who are responsible for managing these resources. In Roy Tennant’s words, “Users should be able to discover through one search what digital objects are freely available from a variety of collections, rather than having to search each collection individually” (Tennant, 2001). Users should not have to know or understand the methods used to describe and represent the contents of the digital collection; but in reality, they are experiencing difficulties in retrieval. How to enable the sharing and exchange of data and facilitate, for the user, a “one-stop” seamless search, also
referred to as "federated search," presents considerable challenges. To achieve that, the different metadata schemas must be made interoperable to enable conversion and exchange of data. Thus, the purpose of interoperability is to facilitate the exchange and sharing of data prepared according to different metadata schemas and to enable cross-domain searching. In recent literature, a great deal has been written about interoperability between and among different metadata schemas (Chan, 2005).

### 3.12.1 Definition of Interoperability

There have been many attempts at defining the concept of interoperability. A few examples are given below:

"Interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality" (NISO, 2004).

"Interoperability is the ability of two or more systems or components to exchange information and use the exchanged information without special effort on either system" (CC:DA, 2000).

"Interoperability: The compatibility of two or more systems such that they can exchange information and data and can use the exchanged information and data without any special manipulation" (Taylor 2004).

It is becoming generally acceptable in the information community that interoperability is one of the most important principles in metadata implementation. Other basic metadata principles include simplicity, modularity, reusability, and extensibility (Duval et al., 2002; Zeng et al., 2003). These principles inform about metadata database design as well as other system-dependent developments. From the very beginning of a metadata project, the principles that enable user-centered and interoperable services should be foremost in design and implementation.

### 3.12.2 Metadata Interoperability Projects at Different Levels

In recent years, numerous projects have been undertaken by the many players and stakeholders in the information community to achieve interoperability among different metadata schemas and their applications. Ideally, a uniform standard approach would ensure maximum interoperability among resource collections. If all participants of a consortium or repository were required to use
the same schema, such as the MARC (Machine-Readable Cataloging) format or the Dublin Core (DC), a high level of consistency would be maintained. This, of course, has been the approach in the library community for over a century and is the ultimate solution to the interoperability problem (Chan and Zeng, 2006a).

However, although it is a conceptually simple solution, it is not always feasible or practical, particularly in heterogeneous environments serving different user communities where components or participating collections contain different types of resources already described by a variety of specialized schemas. The uniform standardization method is only viable at the beginning or early stages of building a digital library or repository, before different schemas are adopted by the participants. Examples include the MARC standards used in union catalogs of library collections and the Dublin Core-based Electronic Theses and Dissertations Metadata Standard (ETD-MS – http://www.ndltd.org/standards/metadata/current.html) used by members of the Networked Digital Library of Theses and Dissertations (NDLTD) (Chan and Zeng, 2006a).

Different Metadata sets are evolved often in the same domain or for similar purposes. When users of such services need to search across the collections of two or more such repositories, the need for interoperability arises. Hence the uniform standard approach is not applicable; therefore other mechanisms of achieving interoperability must be adopted. From a methodological point of view, implementing interoperability may be considered at different levels: schema level, record level, and repository level (Chan and Zeng, 2006b):

a) Schema level – Efforts are focused on the elements of the schemas, being independent of any applications. The results usually appear as derived element sets or encoded schemas, crosswalks, application profiles, and element registries.

b) Record level – Efforts are intended to integrate the metadata records through the mapping of the elements according to the semantic meanings of these elements. Common results include converted records and new records resulting from combining values of existing records.

c) Repository level – With harvested or integrated records from varying sources, efforts at this level focus on mapping value strings associated with particular elements (e.g., terms associated with subject or format elements). The results enable cross-collection searching.
It should be noted that the models discussed in this article are not always mutually exclusive. Sometimes, within a particular project, more than one method may be used.

3.12.2.1. Crosswalks
A crosswalk is "a mapping of the elements, semantics, and syntax from one metadata scheme to those of another" (NISO, 2004). Currently, crosswalks are most commonly used method to enable interoperability between and among metadata schemas. This method begins with independent metadata schemas. Attempts are made to map or create crosswalks between equivalent or comparable metadata terms (elements and refinements). Sometimes other terms such as "field", "label", "tag", etc. are used to refer to "element". The mechanism used in crosswalks is usually a chart or table that represents the semantic mapping of data elements in one data standard (source) to those in another standard (target) based on the similarity of function or meaning of the elements (Chan and Zeng, 2006a).

In the present work, a study of Qualified Dublin Core metadata standard and MODS is undertaken with the objective of developing crosswalks to establish interoperability between repositories using either of these. The study is presented in chapter 4 and implementation of the same in DSpace repository is demonstrated in chapter 5.

3.13 References


Chapter 3: Digital Repositories and Metadata


Chapter 3: Digital Repositories and Metadata


