Literature Synthesis

Miles Robinson
Computer Science Department
University of Cape Town
E-mail: miles.robbo@gmail.com

Abstract:

Digital library systems are required to provide more functionality than basic access and storage. Some of the core functionality of an effective digital library system was identified as Organisation; Preservation; and Accessibility. In order for data to be well organised within a digital library, good metadata is essential. It is also important that the metadata is OAI-PMH compliant as this facilitates interoperability. Good organisation within a digital library forms the backbone of the additional services provided by a digital library system. Preservation ensures that data is kept available indefinitely; while Accessibility deals with issues raised by copyrights, and constructing a portable system.

In this paper, DSpace, Fedora, OpenDLib and Greenstone were analysed as digital library systems. Each of them implemented the core functionality outlined, with some excelling in certain areas. Experiments on the Bleek and Lloyd collection found that a much more portable solution can be produced. However, it should be noted that no standard method of evaluation currently exists, which may lead to inaccurate results when comparing systems.

1. Introduction

This literature synthesis seeks to identify what architectures and toolkits are being used in digital libraries. The core functionality of digital libraries will also be investigated, as well as how the architectures and toolkits of digital libraries assist in the promotion of this functionality.

Research into digital libraries has grown dramatically since the 1990s [2]. However, a number of resources have indicated that there is still confusion surrounding the term “digital library” [2, 3, 12, 15]. This confusion is due to the absence of a standard definition of a digital library. As a result, no clear method of evaluating digital libraries exists. Furthermore, this lack of clear, consistent evaluation has hindered the research and development of effective digital libraries [17].

It has been established that at their most basic level, digital libraries provide storage and access to various digital objects. These digital objects can be any form of digital content (image, text, video or audio) [12]. Digital libraries are growing in popularity and are being used by various organisations throughout the world. This has, in turn, meant that for a digital library to be considered trustworthy and effective, it should support some core functionality beyond the basic storage and access capabilities. The core functionality that should be supported by a digital library will be addressed in the next section. Following that, four of the available digital library systems will be examined. Critical comparisons between the architectures will then be made and, finally, conclusions will be drawn based on the findings.

2. Core Functionality of Digital Libraries

Some of the core functionality that should be supported by a digital library beyond the basic storage and access capability include: organisation; preservation; and accessibility [3]. The inclusion of this core functionality helps ensure that a digital library is trustworthy and effective. Additionally, the core functionality provides the backbone upon which many of the services and further functions of digital libraries are built [3]. Each of the above mentioned core functions will now be individually examined:

2.1 Organisation

The organisation of data in a digital library is important as it forms the backbone of a variety of user services (for example, searching), as well as internal services (for example, preservation). The naming of digital objects and the metadata attached to each object need to be addressed in order to facilitate accurate data organisation.

The name attached to a digital object serves as a unique identifier to that object. To ensure that this identifier is unique, a permanent, collective system for naming objects needs to be established. This has various implications, most notably that the name cannot be associated with a particular location [3]. This ensures that a document is not lost if the location of a resource is changed. Various schemes have been proposed to solve the issue of persistent naming. Some of these schemes include PURLS – Persistent URLs; URN – Uniform Resource Name; DOI – Digital Object Identifier System [3]. Although these schemes exist, issues surrounding the migration of digital objects to newer technologies may require a specific institution to govern a system of unique names [3].

Metadata is the additional data associated with a digital object that is used to describe its content and attributes. A digital object's name (see above) is an example of metadata associated with a digital object [3].The metadata system being used should provide the framework to describe any digital object being stored in a particular digital library. Metadata associated with a digital object will be split into descriptive and administrative metadata [7]. Descriptive metadata should follow the Dublin Core schema. Dublin Core is the most widely accepted schema, and provides a means of describing any type of resource. In addition, the Dublin Core can be qualified to facilitate higher precision [7]. Administrative metadata is further split into technical metadata, rights metadata, and preservation metadata. Each of these will typically follow a well accepted metadata schema. The schema used to store technical metadata (describes technical characteristics of a digital object – for example, format) is determined by the type of digital object (i.e. image, text, video, and audio objects will follow different schemas) [7]; the METS schema for Rights Declaration can be used for rights metadata; and the PREMIS (PREservation Metadata Implementation Strategies) schema is available for preservation metadata [7].

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) can further benefit the organisation of a digital library [19]. OAI-PMH provides a model that ensures metadata follows a well-defined format to allow third parties to easily harvest it. This will allow the third parties to build services (for example, metadata based searches) that add value to the metadata [19]. OAI-PMH also provides the benefit of interoperability between digital libraries. Interoperability allows for the interaction of existing services, as well as the potential to create new services for particular infrastructures. In addition contributions to a larger, general system are possible through interoperability [20].

2.2 Preservation

Data preservation is an important aspect of a digital library, as they may contain precious data [23]. Preservation involves ensuring that a digital object remains available on the digital library forever. There are four important aspects that need to be addressed regarding preservation.

Integrity is the first aspect of preservation. Integrity involves ensuring important that digital objects in the digital library are not modified accidentally or purposefully (unauthorised modification). Integrity is important with regard to preservation as it ensures that a digital object is not incorrectly modified during its lifetime [10]. Several techniques can be implemented in order to ensure the integrity. Examples of these techniques include: digital signatures; audit trails; and checksum [10, 14].

Digital signatures are a hash of a byte stream that is converting it into a digest. The digest is then periodically recreated and compared to the original. If there are any discrepancies they will be reported, and the original digital object will be restored [10].

Audit trails make certain that the life-cycle of a digital object is recorded and maintained. Each time a digital object is modified by an authorized user, the metadata of that object will be updated to indicate that there is a new version. [10].

Checksums are an addition of the basic components of a message. Like digital signatures, the value is recalculated and compared to the original in order to determine any inconsistencies [14].

The second aspect is the structural design surrounding the digital object when it is being stored. This refers to both the metadata associated with the object as well as the format that the object is stored as.

LOCKSS (Lots of Copies Keep Stuff Safe) is a peer-to-peer preservation system [21]. It can be used for preservation as two of its primary functions are preserving and auditing. In addition, LOCKSS does not require much technical administration [21]. OAI plugins can also be used with LOCKSS. This provides further motivation to store metadata in an OAI format, as it will assist in both interoperability and preservation [21].

Technologies may go out of date in the future, thus rendering the format’s associated with them obsolete. To solve this problem, data could be migrated to a different format [9].

However, this is not a perfect solution as data is easily lost in the migration process. Research is being done into this area, but the act of migration is still fairly experimental [10].

The third aspect is the preservation of the storage medium. This refers to the actual hardware that the digital objects are being stored on. The rapid development in computer hardware means that a particular storage medium is likely to go out of date within a few years [3, 9]. This means that digital objects will have to be regularly transferred from older storage media to newer ones. While undergoing the transfer process there is the risk that data is lost as the newer technology no longer has the ability to read the information (for example CD-ROM drives cannot read floppy disks). This will require digital libraries to meticulously ensure that data is not lost when it is being transferred from storage medium to storage medium [10].

The ongoing maintenance of a digital library is the final aspect of preservation. Should a digital library go unmaintained, there is a high likelihood of data being lost or corrupted. In order to keep digital objects preserved, it is important that regular maintenance is performed. This is particularly relevant when data is to be migrated, or the system is being upgraded [1].

2.3 Accessibility

Accessibility deals with two key issues, namely access control to information; and availability of information.

Access control involves restricting access to digital objects. Due to the nature of digital objects, they can be easily copied and remotely accessed. As a result, paper-based copyright is ineffective in the digital environment [3]. Because digital libraries do not own the copyright on the material they are storing, certain mechanisms need to be developed in order to manage the copyright on digital objects [3]. To achieve this, middleware can be used to grant access to authorised users only. This mechanism will allow the digital library to have a greater deal of control over who accesses what object [8].

Digital libraries should be readily and economically available to the public [15]. This means that an environment should be established in which users can easily access content in the digital library [15]. Content should also be available at all times. In order to ensure that digital libraries are easily accessible, they must be portable. This means that it should be able to work in as many environments as possible. The portability of database systems is generally low, as they require installation and they are not platform independent. In comparison XML systems require no installation and there are many tools available to interact with XML – making it more platform independent. Thus XML systems are more portable. In addition XML formats assist with preservation as it is human-readable, whereas databases are usually stored in binary formats. [23]

3. Systems being used

Four current digital library systems will now be analysed. The systems that are to be examined are: DSpace; Fedora; OpenDlib; and Greenstone.

3.1 DSpace

DSpace is an open source system that functions as a digital library. It provides a means for storing and managing research and educational materials. The focus of DSpace is simplicity, meaning that it provides all the necessary functions for a digital library, but is implemented in a manner that should provide an easy experience for the user [13].

DSpace manages the organisation of digital objects through the use of metadata. DSpace holds three types of metadata about each digital object that it stores, specifically: descriptive metadata; administrative metadata; and structural metadata [5, 14]. DSpace's metadata does comply with OAI-PMH. This allows for greater interoperability between architectures [14].

DSpace requires administrator duties. In order to ensure that the digital objects within DSpace are preserved, administrators will be required to "vacuum" the database, as well as perform backups. These duties are performed using regular SQL commands [5]. In addition the administrator will be required to maintain the integrity of the digital objects by invoking a checksum [5].

Accessibility has been addressed by DSpace. Access to digital objects stored on DSpace is controlled through the access control lists. Each user will have an access control list associated to them, and any item that they cannot interact with (Read, Write, Execute) will not be available to them [5, 14].

DSpace provides further functionality to its users. Some of the additional functions include: plugin support; authentication; supervision and collaboration; search and browse; import and export; and statistics [5].

3.2 Fedora

Fedora (Flexible Extensible Digital Object Repository Architecture) provides an extensible architecture that allows for the storage, management, and access of digital objects. Like DSpace, Fedora is open source repository software that allows users to manage their resources effectively [22].

Fedora manages the organisation of the digital objects through the use of metadata. Each digital object stored by Fedora has a Dublin Core record associated with it. Additionally, each record conforms to OAI-PMH, allowing for easy data harvesting. Further metadata may be associated with each digital object. This secondary metadata does not necessarily follow the Dublin Core schema [22].

Each digital object is stored in XML files in Fedora. However, a relational database is also used.

Preservation is supported in Fedora through RDF support and semantic triple store technology. RDF support facilitates the merging and migrating of data, even if the descriptive schemas differ [16]. This is a useful feature as it allows digital objects to be moved from one digital library to another without too much difficulty.
The triple store technology ensures that backups have been made, and have been stored in different locations. This is helpful as data can easily be restored in case of emergency [22].

Like DSpace, Fedora makes use of checksum to ensure the integrity of the digital objects. However, unlike DSpace, Fedora supports automatic checksum, meaning the administrator does not have to perform the check manually [22].

Fedora also supports accessibility by restricting certain users to certain digital objects. User can customise the access control surrounding their digital object [22]. This will prevent unauthorised users from accessing restricted content.

In addition Fedora offers further functionality. The following services are some of the additional functions: basic search; versioning; XML based ingest and export; administration interface; repository rebuilder utility; and Web-based administration [22].

3.3 OpenDLib

OpenDLib is a repository service used to store and disseminate digital documents. OpenDLib makes use of an HTTP-based protocol, and it aims to be application independent by being generic with respect to type and contents [18].

OpenDLib facilitates multiple metadata formats. OpenDLib makes use of XML to store metadata, as well as to provide a template to map new metadata to. The OAI-PMH protocol can be facilitated in the repository [18]. However, it is up to the user to customise these attributes.

Access control is also determined by the user. Each object stored within OpenDLib will have its own access policy restricting which users have access to it. Administrators are required to handle users, as well as to facilitate preservation and integrity [18].

3.4 Greenstone

Greenstone digital library software is an open source system that allows for the construction of information collections. Greenstone facilitates easy maintenance, and is easily extendible using plugins [24].

Documents are stored in a simple HTML-like format known as GML (Greenstone Markup Language) [24]. When documents are added to the digital library, they are automatically converted into this format. Metadata associated with the digital object is included in GML. The metadata is also OAI-PMH compliant, facilitating easy interoperability [24].

Administration functionality is included. This allows users to protect documents so that they can only be accessed by registered users. Greenstone is developed to be accessed over the World Wide Web, therefore boasting high portability [24].

4. Critical Comparison

To evaluate the effectiveness of the above digital library systems, a critical comparison between them will be performed. This will highlight general strengths and weaknesses of digital library architecture.

Metadata in all of the above systems is OAI-PMH compliant. This implies that all the above digital library systems allow for interoperability and easy access. This also suggests that all the systems have the ability to support greater preservation, as LOCKSS can be used.

Preservation is supported more comprehensively in Fedora, as it makes use of XML files (which are human-readable), as well as triple store (which allows for easier migration).

Integrity is reliant on the administrator in DSpace and OpenDLib. However, Fedora supports automatic checksum. This is much more convenient

Accessibility is controllable in similar manners for all the systems. Each has a manner of restricting access to certain digital objects.

The portability of Fedora, Greenstone and OpenDLib is greater than that of DSpace. DSpace relies on databases, whereas the others are more Web-based. However, it has been expressed that an even more portable manner can be used. This is demonstrated on the Bleek and Lloyd collection, where a system requiring no installation was devised [23].

Although some usefulness can be derived from comparing the digital library systems, there is no specific manner in which they can be accurately evaluated. Although there are certain procedures that can be followed when evaluating a digital library, there is a lack of a consistent standard that should be followed [17]. In addition the user’s experience and opinion must also be considered, as this plays a large role in the adoption of a particular digital library. The usability of the digital library is a vital aspect that is often overlooked [11].

5. Conclusion

The core functionality that should be included in order to ensure a digital library is effective is as follows: Organisation; Preservation; and Accessibility.

Modern digital libraries systems do implement this functionality to some degree, while providing the user with a wide range of additional features, as well as providing numerous other benefits (such as portability).

However, it was expressed that there is difficultly in accurately measuring the effectiveness of a digital library despite knowing the core functionality. This is because no standard evaluation system exists, and the usability of the digital library is often neglected. It was also illustrated that faster systems (such as the one devised for the Bleek and Lloyd collection) may be better.

6. References

[1] Ackerman, M.S. and Fielding R.T, Collection Maintenance in the Digital Library. *Proceedings of Digital Libraries*. 1995.

[2] Borgman, C. What are digital libraries? Competing visions. *Information Processing and Management. 35*. 227 – 243. 1999.

[3] Cleveland, G., Digital Libraries: Definitions, Issues and Challenges.*UDT Occasional Paper. 8*. 1998.

[4] Dublin Core. 2011. Retrieved 25 April 2011, from Dublin Core Metadata Initiative: http://dublincore.org/

[5] DSpace Documentation. 2011. Retrieved 25 April 2011, from DSpace: http://www.dspace.org/1_7_0Documentation/

[6] Fedora Commons. 2011. Retrieved 25 April 2011, from Fedora Commons: http://fedora-commons.org/

[7] Gartner, R. Metadata for digital libraries: state of the art and future directions. *JISC Technology & Standards Watch*, 2008

[8] Gourley, D. An Architecture for the Evolving Digital Library. *Washington Research Library Consortium*. N.D.

[9] Hedstorm, M. Digital preservation: a time bomb for Digital Libraries. *Computers and the Humanities*. *31*(3). 189 – 202.1997.

[10] Jantz, R. and Giarlo, M.J. Digital Preservation: Architecture and Technology for Trusted Digital Repositories. *D-Lib Magazine.11*(6). 2005.

[11] Kling, R. and Elliot, M., Digital Library Design for Usability. 1994.

[12] Lagoze, C., Krafft, D.B., Payette, S. and Jesuroga, S. What Is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL. *D-Lib Magazine*. *11*(11). 2005.

[13] Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., Tansley R. and Walker, J.H. DSpace: An Open Source Dynamic Digital Repository. *D-Lib Magazine*. *9* (1). 2003.

[14] Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G. and Smith, M. The DSpace Institutional Digital Repository System: Current Functionality. N.D.

[15] Waters, D. What are digital libraries? *CLIR Issues*. *4*. 1998.

[16] W3C RDF: 2011. Retrieved 25 April 2011, from W3C Semantic Web: http://www.w3.org/RDF/

[17] Zuccala, A., Oppenheim, C. and Dhiensa, R. Managing and evaluating digital repositories. *Information Research*.*13*(1). 2008.

[18] Castelli, D. and Pagano, P. A Flexible Repository System: The OpenDLib Solution. *Elpub2002 Proceedings*. 2002.

[19] Lagoze, C. and Van de Sompel, H. The Making of the Open Archives Initiative Protocol for Metadata Harvesting. *Hi Tech*, *20*(2), 2003

[20] Payette, S., Blanchi, C., Lagoze, C and Overly, E.A. Interoperability for Digital Objects and Repositories, *D-Lib Magazine*. *5* (5). 1999.

[21] Santhanagopalan, K., Fox, E.A. and McMillan, G. A Prototype for Preservation and Harvesting of International ETDs using LOCKSS and OAI-PMH. *9th ETD Conference*. 2006.

[22] Staples, T., Wayland, R. and Payette, S. The Fedora Project: An Open-source Digital Object Repository Management System. *D-Lib Magazine*. *9* (4). 2003.

[23] Suleman, H. Digital Libraries Without Databases: The Bleek and Lloyd Collection. *Proceedings of Research and Advanced Technology for Digital Libraries*, *11th European Conference* (ECDL 2007), 392-403, 16-19 September 2007.

[24] Witten, I.H., McNab, R., Boddie, S. and Bainbridge, D. Greenstone: A Comprehensive Open-Source Digital Library Software System. 2000.