SimplyCT Literature Synthesis

Stuart Hammar
Computer Science Department
University of Cape Town
E-mail: shammar@cs.uct.ac.za

ABSTRACT

A digital library architecture that is lightweight, efficient and general has not yet been developed. Assessing current architectures provides an understanding into the key functions that such a system should possess. Five digital architectures were focused on to uncover their key features and flaws. After which, it was noticed that only CALJAX is lightweight, distributable and takes advantage of powerful client-side Web browser languages. CALJAX exposes the idea of being able to function without any additional software. The other architectures need preinstalled software to function. In addition, CALJAX uses the power of the humble Web browser for powerful client-side operations. Scalability does appear to be an issue for CALJAX, but its concept can be adapted to cater for this. This uncovers a technological paradigm that digital library architectures have not explored enough of and should explore more of.

1 INTRODUCTION

The 21st century has arrived and has brought with it the need for a sustainable digital storage solution. A perfect solution to storing mankind's heritage, in a digital form, does not exist. A universal, reconfigurable, scalable, easily searchable and preservable digital repository system is desirable. To develop such a system, the challenges, needs, successes and failures of current and previous solutions need to be uncovered.

Below, various digital library (DL) architectures and their components have been discussed, analysed and compared. The DL architectures which are discussed include the Flexible and Extensible Digital Object and Repository Architecture (Fedora) [11]; DSpace [13]; the National Science Digital Library (NSDL) architecture [9]; the Tufts Digital Library (TDL) architecture [7]; and CALJAX [4]. Due to the lack of standardisation of digital library architectures, the comparison and evaluation of systems is challenging.

The topic of digital library architectures is a broad one. Only a few specific areas will be discussed in this article. The focus is on challenges faced by digital library architectures.

2 CHALLENGES FACING DIGITAL LIBRARY ARCHITECTURES

Kuny and Cleveland [8] highlight that the Internet may change the fundamental concept of a library in the 21st century; they were right. The importance of a DL model – or more generally as a digital repository model – is essential in going in to the future and for disseminating information.

The creation of a successful DL system poses many technological challenges. Information of all formats – video, audio, image and text – needs to be stored in collections [8]. These collections need to be accessible and discoverable by multiple users simultaneously. They also need to be easily copied for preservation purposes.

The main challenges of digital library architectures include: data and metadata storage; search and discovery services; curatorship and access control; and preservation.

2.1 Data Storage

The way digital repository systems store their data can influence various choices on preservation, interoperability and dissemination. Three different storage techniques are discussed below.

DSpace offers two methods for storing digital content [1]. The first is in the file system on the server. The second is using the Storage Resource Broker (SRB). Both methods can be achieved using an API. SRB is suggested as an optional file storage system or to be used in conjunction with the server file storage system. The data objects are stored in and retrieved from a file system via the bitstream storage manager API [19]. The relationships between the data, the bitstream information and metadata are stored in a relational database on the server.

On the other hand, the CALJAX DL system is database-free and stores its data in a central repository [15]. The central repository contains a collection of digital objects and metadata. These are stored as files in hierarchical directories, where each file is associated with a metadata file. This allows digital collections to be easily distributed. To distribute the repository, the central repository's contents are simply copied on to a removable media device. This is a very simple and lightweight system in comparison to DSpace's heavyweight infrastructure.

The NSDL system is a Networked Digital Library (NDL) system. Its actual digital objects are stored on various servers. Using a metadata repository (discussed in *Section 2.2*), the metadata information is made accessible to its services [9]. The digital content is made accessible through HTTP or FTP linked via the object identifier in the metadata.

DSpace and CALJAX support almost every file type available [4,19]. However, only text-based file formats (e.g. PostScript, PDF, ASCII text, HTML and Word documents) are supported by NSDL [9].

2.2 Metadata Storage

Metadata is ultimately data about data. However, from an architectural perspective, there appears to be no significant difference between metadata and data (i.e. metadata is data in its own right). Librarians make use of metadata to catalogue printed information [3,5,8]. Additionally, they make use of a fixed vocabulary for describing the data in the collections they keep [5]. Conducting searches on indexed metadata information is more efficient than searching full-text documents. To aid with interoperability and to create standardisation, many digital toolkits and architectures (e.g. DSpace, Fedora and NSDL) make use of standard metadata formats such as Dublin Core (DC).

The Open Archives Initiative (OAI) has developed a Protocol for Metadata Harvesting (OAI-PMH) as a promising method for connecting data providers to service providers [17,19]. The OAI-PMH is a client-server protocol. Providers use the OAI-PMH to expose metadata in different ways, and the service providers use the protocol to gather or harvest the metadata. The providers can then process the data and add value to it in the form of services [12], such as a search or citation service. Shearer [12] highlights that using the OAI-PMH helps with the interoperability of repositories so that they can contribute to a larger global system. Following from the OAI's standard, Suleman [18] supposes that standardised components can then be designed for DLs. He goes on to state that these components could include search engines and browsing services – both of which are important for resource discovery and dissemination.

DSpace architectures maintain a DC metadata record for each data object that is stored. Three DC fields are made compulsory per item, namely the title, language and submission date [13]. Any other metadata fields are made optional. Each metadata record is stored in a relational database on the server [1]. DSpace supports the OAI-PMH as a data provider. Only the basic unqualified DC metadata set export is enabled by default; this is easy for DSpace since it stores DC metadata for each item. Therefore, the inclusion of the DC record simplifies the sharing of metadata and interoperability.

The NSDL manages its metadata differently to the previously mentioned architectures. This is because it is a networked digital library. The metadata that it gathers from its various independent collections are not all in the same format, consequently it supports eight standard formats of metadata [9]. The NSDL gathers metadata and stores it in a central metadata repository. To populate the metadata repository, the NSDL harvests metadata in five ways: via OAI; via FTP, e-mail or Web-upload; through direct entry; and by using a Web crawler. Constructing this metadata repository allows for a faster searching service than if all the metadata were searched on the distributed network [2].

There is no single standard for metadata or data in DL systems. Therefore, deciding which metadata to store and how to store it is a challenge and each DL system handles this differently. The OAI-PMH has provided a way for repositories to become interoperable and is a step towards a standardised component.

2.3 Search and Discovery Services

Information on the Internet exists in all formats [8]. Search engines provide a means for users to search the Internet for information. A search service is an integral part of digital library architectures. However, searching the full-text of each document is an inefficient and expensive process [4].

CALJAX is a prototype system, which provides users with a completely offline digital repository system built on AJAX [4]. CALJAX makes use of Java to pre-process and generate data to facilitate an indexed search. The indexed search is then performed using JavaScript and the results are displayed to the user via their Web browser. CALJAX removes the necessity for a heavyweight back-end to do processing and instead uses the power of the Web browser.

As discussed earlier, the NSDL makes use of a centralised metadata repository. The reason for this is that using a distributed style of searching – such as in the Dienst system – does not scale well [2]. As the number of independent servers grows, the services become less reliable and less responsive. Moreover, in the NSDL, it is not expected that each system understands the same query formats, making a distributed search difficult.

The search engines in both TDL and DSpace make use of the Lucene library [7,19]. TDL makes use of an indexing and search function to make its data available to users [7]. TDL uses methods to show the content of a digital object to the search engine. This is indexed and returned as DC metadata. Using this type of indexing allows full-text and advanced searching to be conducted, in addition to metadata searching.

Another important service that is widely adopted by many digital library systems is OAI-PMH. As mentioned previously, it greatly facilitates dissemination of metadata and aids in the development of a global DL architecture.

2.4 Creating Digital Libraries

The creation of digital libraries poses the question as to whether institutions should use standalone architectures (e.g. DSpace or Fedora) or whether networked DL architectures (e.g. NSDL) should be focused on.

DSpace and Fedora are free and publicly available digital library frameworks. Both DSpace and Fedora support multiple, if not all, digital file formats in their repositories. Fedora makes use of various APIs to allow applications to interact with its data repository [14,19]. It makes use of metadata and data objects, which are stored in XML and databases. However, Fedora does have software prerequisites and does not allow for distributable copies of its collections to be made. Furthermore, Fedora is only a framework and is not a complete system.

As discussed in *Section 2*.1, CALJAX is a very lightweight architecture for creating digital collections. CALJAX is a database-free DL system. The system is OS independent and can be run directly from a portable media device [15]. No preliminary software needs to be installed; the user only needs a Web browser to search and manage the digital collections. CALJAX also supports almost every file type so long as a metadata file is associated with it [4].

NDLs are meant for sharing resources among DLs of similar interests and content [16]. The NSDL has integrated many smaller digital libraries to provide information to people all over the world. Historically, the largest NDL on the Internet was the Networked Computer Science Technical Reports Library (NCSTRL) [10]. NCSTRL uses Dienst as its framework. Dienst is a system for organising a set of separate services running on networked servers to cooperate in providing the services of a digital library.

Of the systems discussed, CALJAX is the easiest to set up. It is OS independent and does not need any preinstalled software. It is a proof-of-concept system and its experimental results indicate that it is a feasible system [15]. However, when scalability and collection size are noted, DSpace, Fedora and the NDL architectures seem to become more feasible system choices.

2.5 Curatorship and Access Control

Kuny and Cleveland [8] point out that librarians are an authority and they, therefore, provide a trusted service. DL systems need to be carefully audited to ensure that the information and collections that they store are trusted. To guarantee that digital collections are managed correctly, an authority needs to have access to adding, removing, editing and monitoring functions. These curators or administrators will be able to edit the metadata of the digital information to ensure that it remains useful.

Document discovery and retrieval can be used anonymously in DSpace. However, for a user to use submission, subscription or administration features they must be authenticated.

Therefore, for a user to perform a task on an object or a collection, the user must have permission to perform that task [19]. DSpace's user interface is Web-based. End-users and curators each have a different interface.

DSpace and Fedora are set up in such a way that a user may be granted access rights to certain collections. Furthermore, their rights may be only to read information, or to edit and manage it too. In order to maintain a digital library, various users need to have different access rights to different collections in the repository.

Administration of NDLs is more complicated than that of standalone DL systems. In the case of NCSTRL, Dienst is used to provide a means for definition and management of distributed collections [10]. For NDLs, administrators need to ensure that their repository is kept valid and up to date to certify that the information is always accessible on the NDL.

There is no standardised manner for handling access control in digital libraries, and the architectures offer their own solutions to this. Yet, they all tend to have similarities in their access control models.

2.6 Data Preservation

DL architectures should be able to store large amounts of information that can be accessible many years into the future.

A key factor of DSpace is preservation. Therefore to ensure preservation of resources, DSpace associates each bitstream (the actual file) with a bitstream format [19]. Fundamentally, a bitstream format is a distinctive way to refer to a particular file format. Each file format, which a user submits, is captured in the DSpace repository. To facilitate preservation, each bitstream format has a support level, which indicates how likely it is that the content will be preserved in the future by the institution.

Similarly, Fedora makes use of DataStreams [11]. DataStreams preserve the internal format and encoding of the file type. However, Fedora stores the DataStreams inside a DigitalObject so that mixed forms of data can be treated in a uniform manner. The DigitalObjects are stored in a repository for later access.

Hitchcock et al. [6] claims that digital repository software is not sufficient to ensure preservation of data. It is suggested that repository support teams be employed for preservation management.

3 CONCLUSION

The ideas surrounding what digital libraries are have been discussed. The techniques used by current architectures have been noted and these can illustrate the way forward for new digital library architectures.

There are similarities in all of the architectures regarding user interfaces for curators and public users. The general trend is to make use of an HTML user interface. The architectures also make use of back-end systems, which do the searching and returning of data to the user. Only CALJAX makes use of the client's browser to perform back-end operations. Future digital libraries could potentially make use of powerful client-side scripting languages, such as JavaScript, to reduce the need for complex back-end systems.

More similarities arise with metadata management. Many of the architectures appear to have adopted the OAI-PMH for harvesting and disseminating metadata. This assists the systems with interoperability and harvesting metadata from independent systems. Making use of good metadata assists the architectures with dissemination, interoperability as well as for making resource discovery easy. Many of the architectures employ the Dublin Core metadata standard for each of their digital items.

The browse and search functions in the architectures use indexed metadata as the searchable information. From the handler information in the metadata, the actual digital objects can be retrieved for the user. The reason the architectures use metadata search – as opposed to full-text search – is to aid with scalability and to improve search performance.

A perfect general solution does not exist. Each of the architectures seems to answer a specific set of user needs. The systems that answer a general set of needs appear too complicated and heavyweight. The power of the client's Web browser is only being harvested by CALJAX. The amount of processing happening at the back-end of the system could be shifted on to the browser. Thus, the back-end could be simplified greatly and made lightweight. This presents an opportunity for a simpler, more robust, preservable, scalable and standardised solution to be created.

4 REFERENCES

[1] DSpace Documentation. 2010. http://www.dspace.org/1_7_0Documentation/.

[2] Arms, W., Hillmann, D., Lagoze, C., et al. A spectrum of interoperability: the site for science prototype for the NSDL. *D-Lib Magazine 8*, 1 (2002).

[3] Baldonado, M., Chang, C.-C.K., Gravano, L., and Paepcke, A. The Stanford Digital Library metadata architecture. *International Journal on Digital Libraries 1*, 2 (1997), 108-121.

[4] Bowes, M. CALJAX: An In-Browser Digital Repository System. 2009. http://people.cs.uct.ac.za/~kumoyo/mbowes/honsproj/resources/files/report.bwsmar002.pdf.

[5] Daniel, R., Lagoze, C., and Payette, S.D. A metadata architecture for digital libraries. *Proceedings of the IEEE International Forum on Research and Technology Advances in Digital Libraries - ADL '98*, IEEE Computer Society (1998), 276-288.

[6] Hitchcock, S., Brody, T., Hey, J., and Carr, L. Digital preservation service provider models for institutional repositories: towards distributed services. *DLib Magazine 13*, 5/6 (2007).

[7] Kumar, A., Saigal, R., Chavez, R., and Schwertner, N. Architecting an extensible digital repository. *Proceedings of the 2004 joint ACM/IEEE conference on Digital libraries - JCDL '04*, (2004), 2.

[8] Kuny, T. and Cleveland, G. The Digital Library: Myths and Challenges. *IFLA Journal 24*, 2 (1998), 107-113.

[9] Lagoze, C., Terrizzi, C., Hoehn, W., et al. Core services in the architecture of the national science digital library (NSDL). *Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries - JCDL '02*, ACM Press (2002), 201.

[10] Pandey, R. Digital Library Architecture. *DRTC Workshop on Digital Libraries: Theory and Practice*, DRTC (2003).

[11] Payette, S. and Lagoze, C. Flexible and extensible digital object and repository architecture (FEDORA). *Research and Advanced Technology for Digital Libraries 1513*, (1998), 517–517.

[12] Shearer, K. Institutional repositories: Towards the identification of critical success factors. *Canadian Journal of Information and Library Science 27*, 3 (2003), 89–108.

[13] Smith, M., Barton, M., Branschofsky, M., et al. DSpace: An Open Source Dynamic Digital Repository. *D-Lib Magazine 9*, 1 (2003).

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

[15] Suleman, H., Bowes, M., Hirst, M., and Subrun, S. Hybrid online-offline digital collections. *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on - SAICSIT '10*, ACM Press (2010), 421-425.

[16] Suleman, H., Fox, E.A., and Madalli, D. Design and Implementation of Networked Digital Libraries: Best Practices. *DRTC Workshop on Digital Libraries: Theory and Practice*, DRTC (2003).

[17] Suleman, H., Fox, E.A., Kelapure, R., Krowne, A., and Luo, M. Building digital libraries from simple building blocks. *Online Information Review 27*, 5 (2003), 301-310.

[18] Suleman, H. and Fox, E.A. Designing Protocols in Support of Digital Library Componentization. In M. Agosti and C. Thanos, eds., *Research and Advanced Technology for Digital Libraries*. Springer, Berlin, Heidelberg, 2002, 75–84.

[19] Tansley, R., Bass, M., Stuve, D., et al. The DSpace Institutional Digital Repository System: Current Functionality. *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries - JCDL '03*, IEEE Computer Society (2003), 87–97.