Digital Asset Management System

A detailed comparison of the DSpace and Fedora Open-source software packages including a review of some of the issues involved in developing a Digital Asset Management System for the National Library of Wales.

Author/Awdur: Nigel Callaghan

Dyddiad/Date: 30. January 2004

Technoleg Taliesin Cyf.
Hafan, Taliesin
Machynlleth
SY20 8JH, Cymru/Wales
+44 (0)1970 832573
nigel@technoleg-taliesin.com
http://www.technoleg-taliesin.com

Table of Contents

1.Executive Summary

This report presents a detailed comparison of the technical features of two open source digital library management products, DSpace and Fedora, and compares them against a requirements list drawn up by staff at NLW. It also looks at a number of related issues and identifies possible next steps in developing a Digital Asset Management System for NLW.

The report is based on a review of available documentation for the products: time-scales did not allow for hands-on evaluation.

Although superficially both products are intended as digital library management systems, they are clearly aimed at different markets: DSpace is focussed on electronic publication and storage for academic institutions, whilst Fedora aims to deliver a basic architecture for an object-oriented general digital object repository.

The report lists over 50 detailed requirements and project issues and then attempts to gauge to what extent each of the products meets the requirements, and what alternatives may be possible.

The report also notes several related issues, including the need for system upgrade over the long-term. A list of some current users of each of the packages is also included, together with a summary of their proposed usage.

The report does not make any recommendations as to which of the two products should be chosen (if either). It is intended to serve as a summary of the features, advantages and disadvantages of each package so that the appropriate Library staff can make an informed decision on the way ahead. What is clear however is that neither product provides a complete match for the requirements of NLW 'out of the box' and both would require considerable development effort (and/or the purchase of additional commercial software) to implement a full system covering the entire lifecycle of digital objects at NLW. It suggests that the decision should focus on a core subset of the most fundamental requirements, principally which package is most suited to handle storage and retrieval of the wide range of digital objects that the Library will need to conserve over the coming decades.

Finally the report suggests some possible next stages after a decision is made, namely a detailed hands-on evaluation of one or both packages, to be followed by (or in step with) more detailed planning for a full pilot project.

2.Introduction

This report has been prepared at the request of Avril Jones, Head of the Computer Section at the National Library of Wales.

The Library already has a number of projects involving digital materials and these will be growing considerably in the future. A need has therefore been identified for some form of central Digital Asset Management System that can handle the needs of the Library in this area. Preliminary work identified two likely software packages, both open source, namely DSpace (developed by MIT and Hewlett Packard) and Fedora (developed by The Universities of Virginia and Cornell). To assist in the decision-making process as to which solution to adopt (if either) it was decided to request Technoleg Taliesin to review the available information about the two products and to produce a report summarising and comparing the features of the two packages.

The work was carried out in late December 2003 and January 2004.

The research was conducted on the basis of documentation on the two systems that could be obtained from the Internet. Attempts were made to contact several users of the systems (particularly the Dspace@Cambridge project, and the National Library of Portugal) but no response was received. Notes from interviews by LLGC staff with the ThesesAlive project at Edinburgh and the Refugee Studies Centre project at Oxford were also made available. In fact a wide range of documentation was available on the Internet, covering technical documentation, presentations and user experience.

E-mail discussion lists are also available for both products providing support and guidance. The list archives were looked at and the live lists were monitored for several weeks.

A full list of documents referred to is in Appendix B, and these will be filed in the Computer Section at LLGC on completion of this report.

Minor disclaimer: a lot of information sources have been used in preparing this report, but the lack of response from existing users, and the lack of fully functioning systems to look at, means that it is possible that some minor details are incorrect, particularly when stating that a particular function is not available – I may just have missed the references to it. To be sure of a particular requirement match it will be necessary to get pilot systems up and running and try it.

N.B. Product names: It is possible to get quite confused when looking for references to both DSpace and Fedora: “dSpace” (note the small 'd') is a company that provides electronic control unit software, DSpace is also a small utility to report on disk free space. The Fedora™ project is currently involved in legal wrangles with Red Hat who have decided to use 'Fedora' as the name of a version of Linux.

3.Overview of available software

The prime purpose of this review is to examine in detail the two main products available in this area and to compare the features provided by and issues involved with using them. Although the focus is on the main products there are also additional research and commercial projects under way to extend the functionality of both products, and these will also be briefly discussed.

There are a number of commercial 'digital library' products available. Prior to commencing on the development of the Fedora system, the University of Virginia evaluated these and concluded that there were a number of shortcomings in them:

> • *Most products are narrowly focused on specific media formats that offer good solutions for managing and delivering video or images but lack adequate tools and support for structured (e.g., XML or SGML) electronic texts or the ability to intermingle media types.*
>
> • *Many products perform well at document management but offer no features for dealing with video or images.*
>
> • *None of the products we examined adequately addressed the need to track and manage the array of ancillary programs and scripts that play an essential role in the delivery of that digital content.*
>
> • *Many products fail to effectively deal with the complex interrelationships among digital content entities. As an example, consider an electronic text in the form of a five hundred-page book. The book consists of a single file containing all five hundred pages of text, marked up using XML. In addition to the XML file, there are also five hundred images that represent the scanned pages from the original hard copy edition of the book. There are also twenty-five audio files that provide a recording of the book's content read aloud. To the librarian, all of these digital media are digital manifestations of the intellectual object known as the "book" and all are closely related to one another.*
>
> • *Few of the products attended to the critical issue of interoperability, failing to provide an open interface to allow sharing services and content with systems from other vendors at other libraries." [9]*

A very useful report by the Open Society Institute [10] provides a comparison of features between DSpace, Fedora and five other packages. The other packages were:

ARNO (Academic Research in the Netherlands On-line). Designed "to provide a flexible tool for creating, managing, and exposing OAI-compliant archives and repositories"

CDSware (CERN Document Server Software). Designed "to handle very large repositories holding disparate types of materials, including multimedia content catalogs, museum object descriptions, confidential and public sets of documents etc.".

Eprints (University of Southampton). " software which creates on-line archives. The default configuration is a repository of the research output of an academic institution."

i-TOR (Tools and Technologies for Open Repositories): basically a content management system for developing web sites linked to repositories.

MyCoRe (University of Essen): a bundle of software tools to support digital libraries and archiving solutions.

It may be worth looking in more detail at the last two, as a possible way of complementing either DSpace or Fedora. There is also a commercial product “VITAL” from VTLS which is an add-on to Fedora. Despite repeated requests I have been unable to obtain any detailed information on this. Their website states:

> *“Based on FedoraTM, the open source software and architecture upon which VITAL is built, VITAL provides every feature—storing, indexing, cataloging, searching, retrieving—required to handle large text and image rich content collections. VITAL takes advantage of technology standards such as XML, TEI, EAD and Dublin Core to easily describe and index an assortment of electronic resources.*
>
> *VITAL builds on the FedoraTM repository architecture by providing VTLS developed work flow extensions, management utilities and enhanced searching capabilities. Using FedoraTM defined web services, VITAL provides a mechanism for your organization to create tools, enhance the functionality provided by VTLS, or leverage the open source community for future applications.”*

Both DSpace and Fedora are open source products aimed at the creation and management of digital libraries and digital assets (the distinction between a digital library and a collection of digital assets may be important). The emphasis of the two however is different, as is the list of features currently available in the product “out of the box”. In both cases considerable additional development work will be needed to adapt and extend the supplied software to fit the Digital Asset Management needs of NLW, although in the case of Fedora, VITAL may be an option to reduce this workload.

3.1.Aims of DSpace and Fedora

Before looking in detail at the features of the two products and the extent to which they meet the requirements of NLW, it is worth recording the primary aims of the two systems.

DSpace

DSpace appears to be aimed very much at the academic market, and is positioned as a tool for electronic self-publication and storage of intellectual output from universities etc., but with capabilities for handling a wider range of digital works.

> *“DSpace is a groundbreaking digital repository for the digital intellectual output of a university. It is designed to capture, store, index, preserve, and redistribute research material in digital format”* [1]
>
> *“DSpace is an open source software system that enables institutions to:*
>
> - *Capture and describe digital works using a custom workflow process*
> - *Distribute an institution's digital works over the web, so users can search and retrieve items in the collection*
> - *Preserve digital works over the long term*
>
> *To collect, distribute, and preserve research materials in increasingly complex digital formats is a time-consuming and expensive chore for individual faculty and their departments, labs, and centers to manage themselves. The DSpace system provides a way to manage these research materials and publications in a professionally maintained repository to give them greater visibility and accessibility over time.”* [2]

DSpace is intended to provide a fully-functioning but basic system:

> *“DSpace was built breadth-first: it supports every function that a research*

> *organization needs to run a production digital repository service, but as simply as possible. The project focus was on building a production quality system."* [4]
>
> *"DSpace followed the librarian's inclination to create a system that would be as easy as possible to implement and use, rather than push strictly in the direction of digital library research from which a more flexible system might have emerged. DSpace, therefore, was designed as an open source application that institutions and organizations could run with relatively few resources."* [5]

Fedora

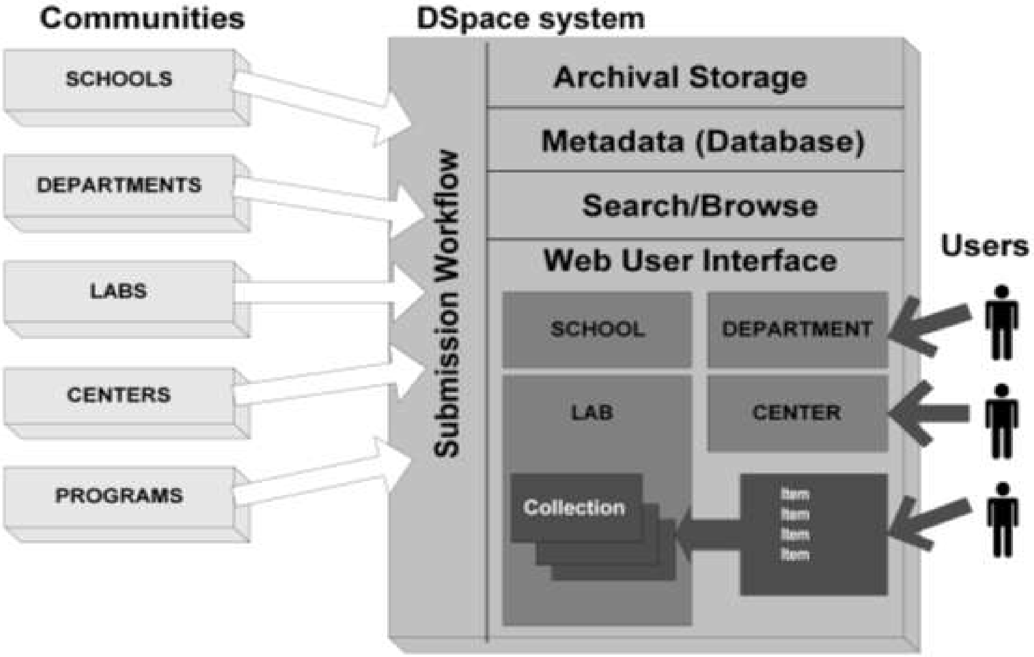
The Fedora project developed from academic research started in the late 1990's at Cornell and Virginia universities aimed at developing an object-oriented architecture for digital repositories.

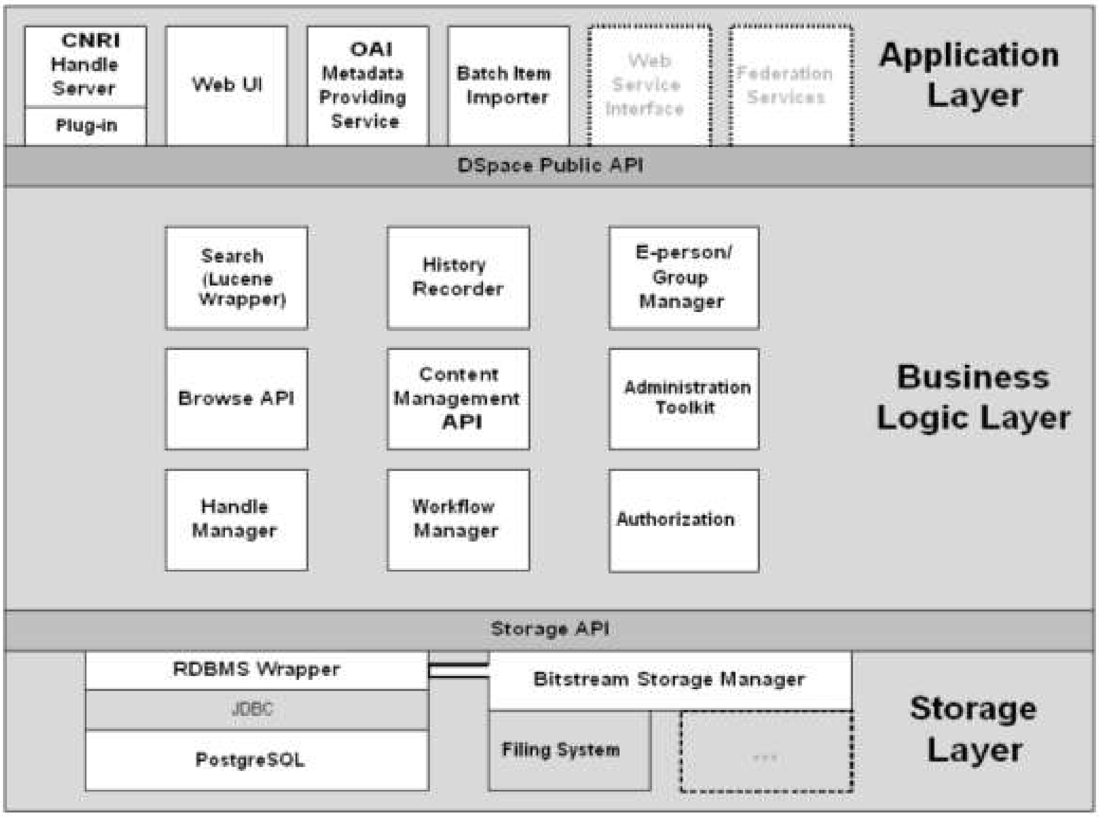
FEDORA stands for **F**lexible **E**xtensible **D**igital **O**bject and **R**epository **A**rchitecture, which summarises its purpose: at this stage it is primarily an architecture, and a set of 'plumbing' rather than being a complete digital library 'out of the box', although plans for future releases aim to move in that direction. Recent releases have already added to the functionality.

3.2.Main components of the systems

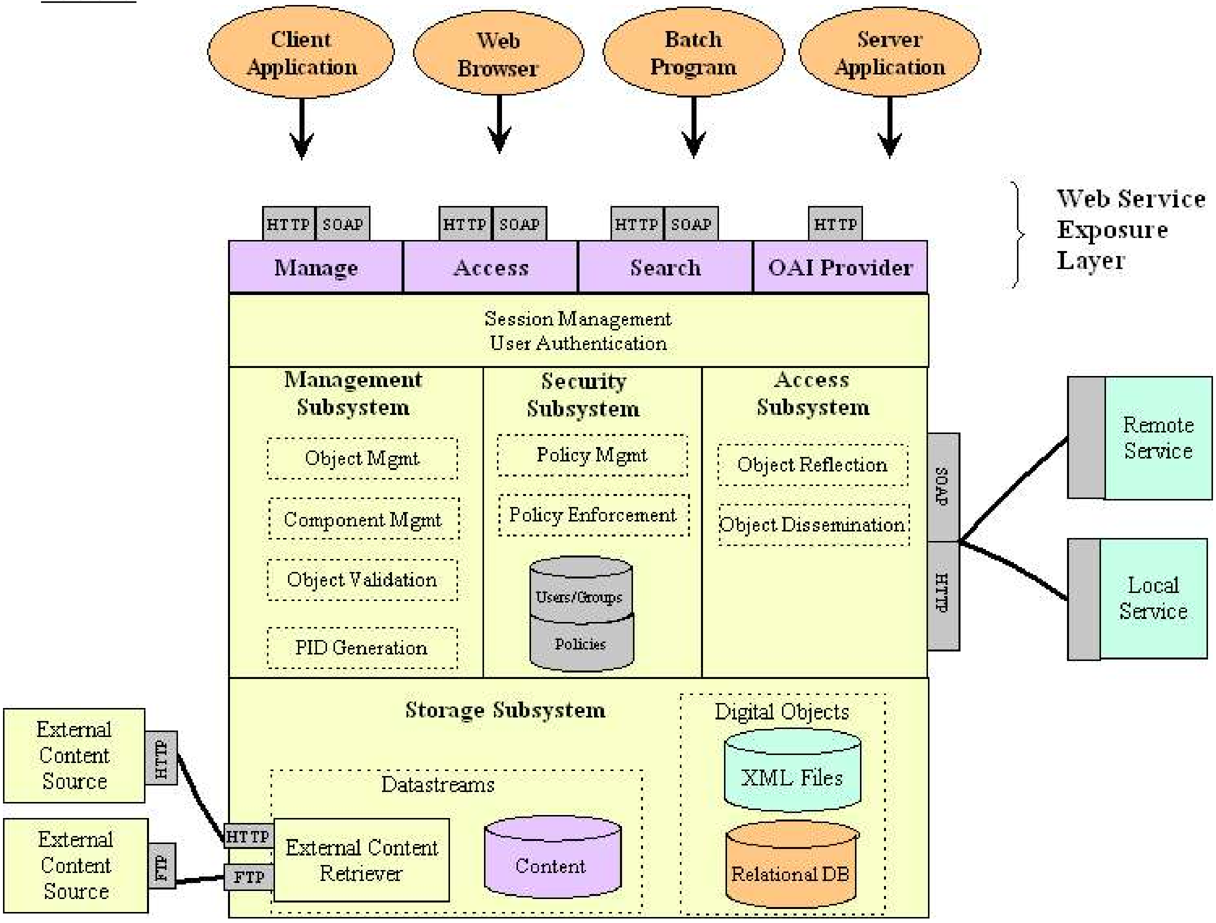
Diagrams are the easiest way to see the overall scope and shape of each of the two systems:

DSpace





Fedora


Client Application
Web Browser
Batch Program
Server Application
HTTP
SOAP
HTTP
SOAP
HTTP
SOAP
HTTP
Manage
Access
Search
OAI Provider
Web Service Exposure Layer
Session Management
User Authentication
Management Subsystem
Object Mgmt
Component Mgmt
Object Validation
PID Generation
Security Subsystem
Policy Mgmt
Policy Enforcement
Users/Groups
Policies
Access Subsystem
Object Reflection
Object Dissemination
SOAP
HTTP
Remote Service
Local Service
Storage Subsystem
Datastreams
External Content Retriever
Content
Digital Objects
XML Files
Relational DB
External Content Source
HTTP
External Content Source
FTP
HTTP
FTP

4.NLW System requirements

A detailed list of requirements has been produced [7] which have been categorised as 'essential' or 'desirable'. This is the basis of the detailed matching exercise in the next section.

As a starting point it is worth summarising the overall functional requirements of a complete Digital Asset Management system.

The simplest way of approaching a high-level requirements definition is to consider the lifecycle of a Digital Object in the system, and to ensure that each stage in the lifecycle is handled, and to add to that any further technical and administrative requirements.

For the purposes of this exercise (rather than as a rigorous definition), the digital object can be defined as the collection of digital items that represent the actual archive item – e.g. a collection of images, files or documents in different formats, - together with the metadata that describes the object and the internal relationships between parts of the object.

'System' in this case refers to the complete Digital Asset Management system.

The system must be able to perform the following tasks:

1. **Creation of Digital Object**
   - Ingest existing digital data items (both born-digital and items where a digital version has been generated elsewhere).
   - Creation of new digital items (digitisation of existing non-digital items)
   - Creation of additional versions of digital items (including 'standard' versions of proprietary formats, e.g. TEI version of a Word document, pdfs of various files, reduced JPEGs and thumbnails for TIFFs etc.)
   - Creation/capture of metadata for the Digital Object.
   - Combining of multiple items with the metadata to create a digital object.
2. **Storage of Digital Object**

   Safely store all components of a digital object over time.
3. **Enquiry on Digital Objects**

   Search and retrieve information about the objects in the repository.

   Report on the contents of the repository, including file reviews.
4. **Data extraction**

   Retrieve a copy of all or part of a digital object from the repository.
5. **Display digital object components**

   Display or otherwise process a copy of the object components in the repository.
6. **Modification of Digital Objects**

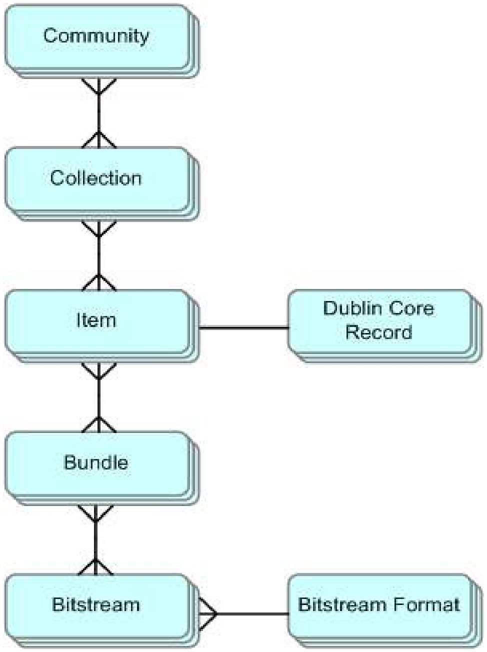
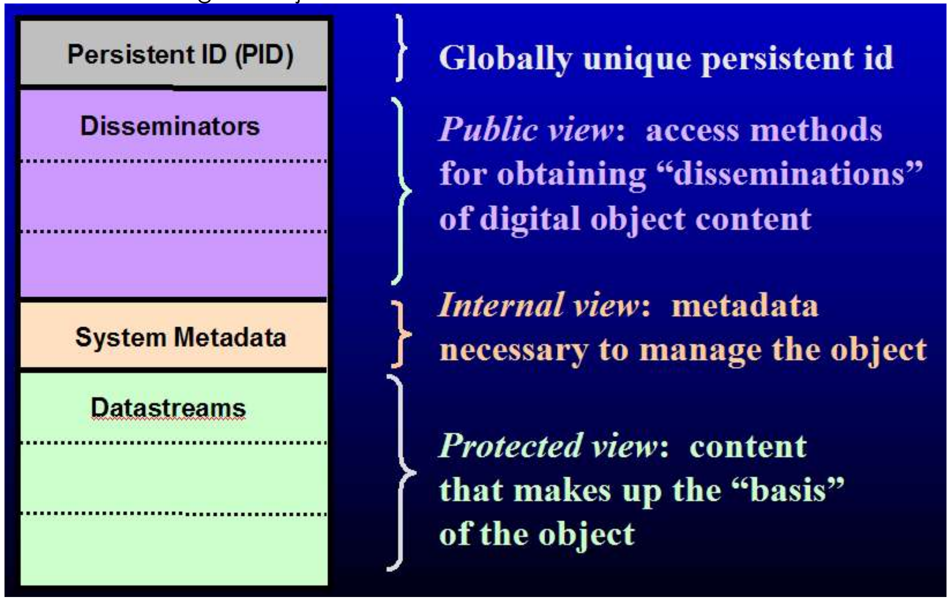
   Update (and delete) metadata and digital object components, including managing migration between formats over time.

In addition the system must perform various administrative functions, such as reporting.

5.Detailed requirements matching

5.1.Basic data model

DSpace and Fedora have different ways of modelling the underlying objects in the repository:

DSpace	***Fedora***
DSpace has a strictly hierarchical model (although objects can appear in multiple instances of a parent – this functionality is apparently not very stable).  Community Collection Item Dublin Core Record Bundle Bitstream Bitstream Format The way data is organized in DSpace is intended to reflect the structure of the organization using the DSpace system. Each DSpace site is divided into communities; these typically correspond to a laboratory, research centre or department. Communities contain collections, which are groupings of related content. Each collection is composed of items, which are the basic archival elements of the archive. Items are further subdivided into bundles of bitstreams. Bitstreams are, as the name suggests, streams of bits, usually ordinary computer files. Bitstreams that are somehow closely related, for example HTML files and images that compose a single HTML document, are organised into bundles. [3]	"The Fedora architecture is based on object models that by definition are templates for units of content, called data objects, which can include digital resources, metadata about the resources, and linkages to software tools and services that have been configured to deliver the content in desired ways."[8] Fedora has adopted a pure object-orientated data model: the repository is used to store Digital Objects.  Persistent ID (PID) Disseminators System Metadata Datastreams Globally unique persistent id Public view: access methods for obtaining "disseminations" of digital object content Internal view: metadata necessary to manage the object Protected view: content that makes up the "basis" of the object There does not appear to be a formal classification of community/collection etc. Documentation refers to 'collection' objects that describe other objects, which can themselves be collections, implying an open-ended arrangement. Individual objects can contain embedded datastreams, or references to datastreams elsewhere in the repository or external to the repository. The system metadata is solely that needed to manage the object: additional metadata describing content can be stored as one or more datastreams.

5.2.Detailed requirements

Requirement	***DSpace***	***Fedora***	***Other options***
ARCHITECTURE			
1. OAI-PMH compatible	Yes	Yes	
2. Sophisticated workflow that operates both inside and outside NLW network	Workflow limited to max 3 steps, and at Item level (not bundle/bitstream). Basically just submission/review/ authorisation. “The reason for this apparently arbitrary design is that is was the simplest case that covered the needs of the early adopter communities at MIT. The functionality of the workflow system will no doubt be extended in the future.” [3]	No. Administrators can set objects to 'Inactive' so they are not visible to the public. Improvements planned for future release.	Develop general workflow system to manage digitisation process prior to submission to repository.
3. XML compatible	Yes – internally, and on import/export.	Yes.	
4. Workflow from beginning to end for digitisation of NLW materials: • a) Manage tracking of materials through entire digitisation process (des) • b) Create files from archival files and manage them • c) Manage creation of files from archival files as required • d) Control who is allowed to do what with what and who did what with what • e) Control and manage everything which is being digitised including on-demand digitisation (des)	a) No b) No c) No d) In general there is a 'police state' model, i.e. no actions without an explicit policy, but this only applies to the processes within DSpace, not the general digitisation process. e) No	a) No b) No c) No d) No – so far as the digitisation process goes e) No Fedora allows for versioning so objects could be created and datastreams added during the digitisation process, rather than as a one-off. Improved workflow in future release.	Develop system to manage digitisation process, which links into main DAMS repository and stores information either in intermediate area or straight into the repository as it is created.
5. Integrated workflow processes: with other applications and other NLW departments (des)	No	No	See above

Requirement	*DSpace*	*Fedora*	*Other options*
6. Management of simple and complex data objects: • Multi-layered, e.g. a book is one physical object that may be digitised into 500 digital objects – need to maintain all layers and be able to identify the one physical object they represent. • Multi-dimensional – e.g. a website is made up of many linked pages – need to maintain navigational capability and manage external links.	Objects are described at 'Item' level, which comprises one or more 'Bundles' each of which contains one or more bitstream files. A bundle can be e.g. "a group of HTML and image bitstreams making up an HTML document". It is unclear how easy it is to navigate through a 'website' directly as the bitstreams are allocated new physical file names within the DSpace repository.	Digital objects can be of any complexity. It seems that they can be grouped into an indefinite number of layers of 'collections'. Improved support for 'kinship' relationships and collections due during 2004. Same issue as DSpace on websites, but as we also define 'behaviours' it would be possible to define a process that e.g. builds a temporary website on a server as part of the 'view website' method.	
7. Ability to implement lifecycle management, e.g. review files	No.	No. But more 'Administrative Reporting' scheduled for 2004.	Can develop add-ons.
8. Scalability e.g. capacity, number of users -- Capacity to include a very large number of digital assets -- Allow more than one person to work on different aspects of the same digital object at the same time -- Allow many users to gain access to the same data at the same time	No obvious limitations, although some suggestions about performance improvements that could be made. Bitstreams can be spread across multiple 'stores', but 'new' additions are added to a 'current' store. Manual intervention is needed to decide when to switch to the 'next' store. Could be an issue for large bitstreams, as I don't think we can just set up a 2TB raid array and forget about it. May be some usability issues if a 'collection' includes thousands of items. Database has no serious limitations on size. Multiple users can access object.	Has been tested with 10 million objects so far. No concurrency issues.	
9. Ability to accept multiple file formats	Yes. Stores information about the precise format with each bitstream. Formats are either "supported" (basic e.g. XML, TIFF, PDF), "known" (common proprietary e.g. Word), and "unsupported" (odd stuff e.g. program binaries, raw bitstream still kept). Can store multiple formats of the same bitstream e.g. .doc and PDF.	Yes.	

Requirement	*DSpace*	*Fedora*	*Other options*
10.Harvest automatically from the data files themselves as much of the technical metadata as possible	No	No.	Can develop add-ons to pre-process files.
11.Ability to point to sources of information which are outside the DAMS, so that there is no need to maintain all the metadata on one system nor to maintain all the data on the same system	No?	Yes	
12.Manage rights	Yes. All users defined with groups/role which has rights. Can also define special groups based on e.g. IP Address. Admin interface has management tools to manage rights at collection and item level.	Uses package called POeT to secure administrative access to the repository. Phase II release will extend policies to object level.	
13.Manage orders for digital files	No.	No.	Can develop as add-on.
14.Manage the distribution/publication of the products of the digitisation programme as website content and/or as other media	Depends on what is meant by 'manage'. By storing the digital products in the repository then whatever options exist for retrieving and viewing objects will be available.	As DSpace, but wider options for defining 'methods' for retrieving data.	
15.Manage the distribution/publication of the products of the digitisation programme in response to the enquiries of specific users	Ditto. Includes a 'subscription' option whereby users can be notified by e-mail when items are added to specified collections.	Ditto.	
STANDARDS			
16.Create and maintain meaningful and relevant metadata effectively in order to enable the integration of all of the Library's digital assets: administrative metadata; structural metadata; discovery metadata; and, especially, presentational metadata	Currently limited to qualified DC set, but other metadata can be stored as bitstream. Unicode supported. Full list in Appendix A.	Uses basic system metadata plus DC set, but can store other metadata as datastream. Unicode supported.	

Requirement	***DSpace***	***Fedora***	***Other options***
17.Support and maintain the Library's current metadata standards • *XML compatible* • *METS* • *MODS* • *MARC21* • *Dublin Core* • *EAD* • *TEI*	See above.	See above. METS is used to encode the XML for the digital objects.	
18.Be adaptable to future standards development	Probably.	Probably.	
<u>ACCESS AND SECURITY</u>			
19.Provide means of access to the products of the digitisation programme	Yes.	Yes.	
20.Control who is allowed to do what with what in the digitisation programme (internal)	Limited to control over adding/retrieving/ processing items in the repository, rather than during the actual creation of the digital items.	Uses package called POeT to secure administrative access to the repository. Phase II release will extend policies to object level.	
21.Control permissions to use files, i.e. who is allowed to do what with what in the digitisation programme (external)	Yes	Can restrict access to objects to administrators only. See above for phase II extensions	
22.Create and manage persistent identifiers for digital objects in order to ensure ongoing accessibility	Yes: uses CNRI Handles.	Generates (globally) unique PID internally at present, but can use existing one when ingesting if required. System is due for review to look at external generation with multiple repositories.	
23.Manage the long-term preservation of the products of the digitisation programme	Should be OK for long-term preservation of the files, (but see section 6.3), but limited in not storing the methods used to access the data over time.	Should be OK for long-term storage, but see section 6.3	

Requirement	*DSpace*	*Fedora*	*Other options*
24.Extranet access for depositors	Yes	Possible	
25.Ability to create group profiles	Yes	Unclear.	
26.Authentication -- Must be compatible with NLW network authentication system -- External depositors will need extranet authentication (des)	Not directly but instructions exist on how to develop links to external authentication modules. HKUST have linked to LDAP server.	Unclear.	
INTEROPERABILITY			
27.Ability to work/integrate with other systems in the Library which manage its collections	API's provided so possible to develop processes to link multiple systems.	API's provided so possible to develop processes to link multiple systems, both web-based and GUI client.	
28.Compatible with other systems – within NLW, cataloguing & indexing systems if separate, email system for workflow, multiple browser support	General e-mail facilities included. Current browsers seem OK. Also works on Netscape 4.7, WebTV: Lynx is a bit messy on layout, but works.	See above.	
29.Ability to launch other software applications from within the system e.g. Word, Excel, PowerPoint, Acrobat (???)	No.	Not directly, but stores details of behaviours and how to invoke software to display an object.	
30.Compatibility with pdf rendering software	Yes. Output is rendered via the browser, so if the bitstream is a PDF and the browser supports pdf then it will work.	Yes. Includes software (Apache FOP renderer) to render output as a PDF.	
31.Import/export metadata facility	Yes – also export with the bitstream as well.	Yes	
32.Import/export files facility (including bulk) – ability to set file size limits, schedule bulk imports/exports	Yes – can export collection or item. Bulk scheduling can be handled with a cron job.	Yes. Unclear about size limits or scheduling.	
PRESENTATION			
33.Web-based presentation	Yes	Yes	

Requirement	*DSpace*	*Fedora*	*Other options*
34.GUI interface	Only via the browser. No GUI client.	Some GUI clients already(Admin) – can also develop other GUI client applications that communicate with the repository (via SOAP Web services)	
35.Minimum keystrokes for main functions	Need to run pilot to check.	Need to run pilot to check.	Run usability tests in pilot, modify as necessary.
36.Fast page loading speed – varying capability to download based on telecomms available (important for external users and depositors)	Basic system is compact with no un-necessary graphics.	Only minimal web front-end provided 'out of the box' so design of application up to NLW.	
37.e-envoy compliance	Seems to comply with highest level at least, i.e. para 5.3 (e-GIF version 5). May need to clarify how to comply with content management metadata requirements.	Seems to comply with highest level at least, i.e. para 5.3 (e-GIF version 5).	
38.Option to present content and metadata in Welsh or English	Not clear if there is supported “lang=xx” option on metadata to differentiate welsh and English versions of the same element. Will need to modify page templates to present bi-lingual interface. University of MSH-Alpes seem to have translated the basic templates into French.	Admin GUI client in English, but written in Java so can presumably be translated. Ditto web pages, but design is up to NLW. Same issue of metadata as DSpace.	
39.Ability to search sets of data as well as across all data	Can search individual communities, collections or all of DSpace.	Unclear – depends on the definition of 'sets' – improved 'collections' support is due in the next release.	
40.Free text, Boolean, and authority file/metadata searching	Limited to standard set of metadata fields: Keyword, author, title, subject, abstract, series, sponsor, identifier. List is hard-coded so can be modified. Limited booleans. Wildcards supported. No searching of content (bitstreams).	Searching limited to system metadata and the DC set. No Boolean or stem searches as yet, wildcards supported.	
41.Dynamic saved search facility	Searches are sent as URL parameters so can be bookmarked.	If invoked from a browser searches can be sent as URL parameters so can be bookmarked.	

Requirement	*DSpace*	*Fedora*	*Other options*
42.Queue searching via: -- Partial indexes -- Content of unindexed objects	No. but tasks at a stage in a workflow exist in a pool.	Once a basic object record is created it can be searched for, even while incomplete.	
43.Downloadable results from searches (html, pdf, word processed or spreadsheet documents, send to email)	No, but can develop add-on.	No, but can develop add-on.	
REPORTING			
44.Report generation e.g. usage, audit trails -- Templates – out-of-the-box and user definable -- Ad-hoc/custom/user-defined -- Across all available metadata (i.e. comprehensive choice of parameters) -- Automated system generated reports -- Report scheduling ability (for overnight, frequency)	No.	No, but planned for future releases.	Can develop any necessary reports as needed, using published APIs.
OBJECT AUTHENTICITY AND INTEGRITY			
45.Enable the authentication of digital objects as correct copies	Yes. Checksums stored for all bitstreams.	Mechanism unclear, but checksum field in the METS	
46.Version control – ability to manage different versions of same files	Doesn't appear to.	Yes. Retains all versions of a datastream and allows timestamp-based access.	
47.Automatic checksum utility to maintain integrity and authenticity of files	Yes.	SIP schema validation.	
48.Ability to check digital material collection for duplication	No.	No.	
GENERAL SYSTEMS SUPPORT			

Requirement	***DSpace***	***Fedora***	***Other options***
49.Adequate systems support; helpdesk; effective escalation procedures	Open source so no formal support. Technical support discussion lists available that seem to get some answers. Probably useful to build up direct relationships with developers at institutions with more experience.	As for DSpace.	
50.Systems availability to external users 24 hours per day, 7 days per week; internal availability as per other network applications	No reason why not – no need to close system while loading data etc. Dependent on external factors, e.g. server reliability.	As for DSpace.	

5.3.Technical requirements and platform

	DSpace	***Fedora***	***Other***
Architecture	Fairly integrated system. Data in the repository is exposed through JSP web pages and Java servlets.	Loose, three-tier system. Repository data and functions exposed via Web services (SOAP-enabled APIs), allowing access via the Internet to any application (not just via browsers).	
Hardware	Nothing specific. Scalable – can test on a desktop box. Production size depends on activity. SAN supported.	Nothing specific. Scalable – can test on a desktop box. Production size depends on activity. SAN supported.	
Operating System	Linux/Unix – Windows possible but no real information about it.	Linux/Unix or Windows	
Development language(s)	Java, JSP	Java, JSP, XSLT	
Other software requirements	Java SDK 1.4, Tomcat 4, Apache. Warning of version control issues from one user – have to have just the right combination of software with the right versions for it to work. Implications for upgrades.	Java SDK 1.4, Tomcat, Cocoon, Apache, Saxon	

	DSpace	***Fedora***	***Other***
Database/File storage	Relational database that supports transactions. PostgreSQL suggested, Oracle also already in use, no mention of MySQL (but transactions only available in MySQL from 4.0). Bitstreams stored outside DB – conventional filesystem.	Comes with a native Java database (McKoi) but supports any compliant RDBMS. MySQL and Oracle 9i have been tested.	
Backup facilities	Export facilities	Export facilities	
Staffing requirements:	OSI[10] suggest Unix Systems Administrator and Java Programmer.	Depends on development work needed. Web interface and GUI clients can be developed in any language using SOAP. OSI[10] suggest Unix Systems Administrator for set-up and Java Programmer.	

6.Other issues

1. **Storage of 'websites'**

   It is unclear just how effective either system will be at storing websites in a format which will be 'directly' usable i.e. able to select and display a page from a site in the repository and then follow links within that page.

   It may be that the mechanism for 'viewing' a website object will require that the site is exported to a web server and link references recreated at that time.

   A related and more general issue is how dynamic websites will be /can be stored. HTML pages are straightforward as they are just text documents – pages generated from perl/php/jsp etc. are created as needed. We can store the source of the system (the pages and databases) but how do we store and run the software that actually runs the system? How are other institutions handling this?

   Further research will be needed in this area.

2. **Package development and updates**

   Both DSpace and Fedora are still in the very early stages of development and adoption. Much still has to be done to develop the feature sets. As a consequence it is likely that new releases of the software will become available fairly frequently over the new few years, and any project plans for pilot and live implementations must allow for the time needed to apply these upgrades, which may be quite complicated in some cases. Migration guidance is usually provided, however.

   Care must also be taken in preparing development procedures to ensure that modifications and enhancements are made in such a way as to ensure that any local modifications to the code are not over-written when applying new releases.

3. **“Long term” preservation**

   “Long term” is difficult to predict in IT – both systems will need regular migration to new media and systems etc. over time. There will be associated costs with each migration, which need to be planned for. The initial system will evolve over time but realistically a life-span of much more than 10-15 years is unlikely before it will become necessary to consider a new system, (assuming the initial system has worked as anticipated).

   Provided the data is physically secure and properly defined in the first place then a migration program should be relatively straight forward.

7.Users

With both packages the user community is still small, and the packages are both still in quite early stages of development and continue to evolve rapidly. As a result there are very few, if any, full-blown live implementations to look at.

It is difficult to tell exactly how many projects are under way on each platform: the nature of open source code means that anyone can download it and there is no central record of 'sales'. DSpace seems to have the advantage in terms of numbers of active projects. It is being developed by Cambridge/MIT in conjunction with the universities of Columbia, Cornell, Rochester, Ohio, and Washington in the United States, plus Toronto in Canada. A number of other smaller institutions are also now rolling out DSpace systems on varying scales, e.g. Duke uses it to store student portfolios. Rochester includes the following list of projects on its website:

Institutions promoting and using DSpace:

Alliance for Innovation in Science and Technology Information (part of LANL)

Cornell University

Drexel University

Edinburgh University's Theses Alive! project

Erasmus University, Netherlands

European University Institute

Hong Kong University of Science & Technology

Ibero American Science and Technology Education Consortium

Kansas University

Roskilde University, Denmark

University of Calgary

University of Cambridge (description, not yet live)

University of Oregon

University of Tennessee

University of Toronto

Many of these seem to be in the early stages of development, and all seem to be concentrating on the primary aim of theses and academic publishing. There is no evidence that any of the institutions above have attempted bi-lingual interfaces (even Roskilde is in English)

Institutions developing and using Fedora

Five universities have been working with the early development stage, testing different types of materials. These are:

University of Virginia: *TEI-marked-up books plus some images; EAD finding aids, some with images; artwork images; early maps.*

Northwestern University: *Images + metadata; MPEG videos+associated jpegs and metadata; historical anatomy atlas (TIFF, SVG + metadata) + TEI encoded text that references the images; TEI-encoded Shakespeare that links to a grammatical dictionary in a relational database.*

Indiana University: *Image collection – existing metadata in a variety of formats; music – linking audio, scanned scores, digital notation scores, video*.

New York University: *Video collections.*

Tufts University: *TEI-encoded SGML texts; GIS maps and datasets; QTVR panoramas; VRML models – delivery through existing interface; University records.*

There is also a document describing the use of Fedora for storing Public Opinion Poll data.

Other organisations using or evaluating Fedora include the National Library of Portugal, the British Library, Monash University, National Archives of Australia, Cornell University (part of the main development team) etc.

8.Summary and conclusions

8.1.Summary

Clearly neither package offers a complete off-the-shelf package that meets all of NLW's digital asset management needs. Whichever package is chosen (if either) it is clear that additional development work will be needed to extend and tailor the package to NLW.

At one stage I considered allocating some numerical weighting to each requirement and then a weighted score for each application against each requirement. This would have given some (possibly spurious) mathematical authority to any decision.

However it became clear that the decision on how to proceed should really be based on a few of the most fundamental issues: how complete the match is on the various detailed requirements probably will not affect the final decision, but merely assist in identifying the areas in which additional development work must be done.

The fundamental issues appear to be:

- Which application is most suited to storing the wide range of data types that will need to be stored by NLW over the coming decades?
- Which application is most suited to distributing the wide range of data types that will need to be stored by NLW over the coming decades?
- Which application can best support the collection structures at the NLW?
- Which application is most scalable?
- Which application has the most 'future-proof' architecture?

To an extent, if there is one clear winner on the above questions then the other issues are secondary, no matter whether the 'loser' does better or not. If the loser is easier from the point of view of technical support or is cheaper to run that doesn't matter if the loser isn't up to the basic job: it merely indicates that the higher costs must be factored in to the plans.

8.2.Project issues

A number of points about the overall project have become clear during the development of this report:

Project Planning: careful, detailed project planning will be essential.

Staffing issues: a project of this scale will require serious commitment of staff, both technical and 'business'. It is impossible to estimate the numbers at this stage, but a dedicated joint development team with seconded library staff and appropriate technical staff will help to ensure a successful implementation.

Cost: although the basic software is free, the project will not be cheap.

Time-scale: the initial development will be a lengthy process, and needs to be, with well planned pilot implementations to ensure that the wider needs of the Library are handled in the basic architecture. Time-scales for wider roll-out will then be dependent on resources available to carry out the digitisation work and the loading of data into the system. Given the rapid rate of development of both products there may well be an advantage to taking things slowly at first, to allow for more sophisticated functionality to be available for the live roll-out.

8.3.Next stages

Once basic decisions to proceed have been made I suggest the following tasks should be considered:

- **Continue investigations:** the limited information available so far suggests that the VITAL package from VTLS may be a very useful extension to Fedora, filling many of the gaps in the necessary features list. Efforts should continue to find out more details, costings etc. Further investigation of other add-ons (i-TOR, MyCoRe) may also be useful.
- **Detailed hands-on evaluation:** install the chosen product (or both products) on a server at NLW to identify exactly what it can do, to assist in the planning and design of the next stages. Can also evaluate quality and maintainability of code, available support etc. before committing to full pilot.
- **Development of detailed use cases:** NLW has a wide range of materials that are candidates for storing in the DAM system. To assist in planning and evaluation of any pilot it is essential to have a range of use cases that cover the full range of requirements of the Library.
- **Pilot scheme:** develop a full working system using a few pilot collections.
- **Continued education:** Dspace@Cambridge are organising a series of seminars (LEADIRS II) on planning digital repositories. It is worth considering sending appropriate staff to these. Further details at http://www.lib.cam.ac.uk/UKseminars/. The series covers: Organisational and Implementation planning; Technology Requirements; Legal and Regulatory Environment; Cost and Funding models. "The aim of this project is to assist in the development of individual institutional level planning for the implementation of sustainable institutional repositories."...."A series of professional seminars along with working materials will be provided to senior managers of institutions in the United Kingdom that are currently planning for, or in the midst of, the implementation of an Institutional Repository"

9.Selected references

(1) DSpace Federation website http://dspace.org

(2) "MIT's DSpace Experience: A Case Study" – on the Dspace.org website

(3) DSpace System Documentation: http://www.dspace.org/technology/system-docs/

(4) "DSpace - An Open Source Dynamic Digital Repository" *D-Lib magazine* Jan 2003 http://www.dlib.org/dlib/january03/smith/01smith.html

(5) "MIT's DSpace Experience: A Case Study" http://www.dspace.org/implement/case-study.pdf

(6) "DSpace and ETO-db Comparative Evaluation" - ThesesAlive! at Edinburgh - www.thesesalive.ac.uk/archive/ ComparativeEvaluation.pdf

(7) "Requirements for NLW Digital Assets Management System" John Watts-Williams - NLW Internal document (GofynionDAMS02.doc)

(8) "The Fedora Project" - *D-Lib magazine* April 2003 http://www.dlib.org/dlib/april03/staples/04staples.html

(9) "The Mellon Fedora Project: Digital Library Architecture meets XML and Web Services" Payette and Staples, ECDL

(10) Open Society Institute – A Guide to Institutional Repository Software ver 2 Jan 2004

Appendix A. DSpace Metadata (Qualified Dublin Core)

(http://dspace.org/technology/metadata.html)

Dublin Core with Qualifiers

Element	Qualifier	Scope Note
contributor		A person, organization, or service responsible for the content of the resource. Catch-all for unspecified contributors.
contributor	advisor	Use primarily for thesis advisor.
contributor	author	
contributor	editor	
contributor	illustrator	
contributor	other	
coverage	spatial	Spatial characteristics of content.
coverage	temporal	Temporal characteristics of content.
creator		Do not use; only for harvested metadata.
date		Use qualified form if possible.
date	accessioned	Date DSpace takes possession of item.
date	available	Date or date range item became available to the public.
date	copyright	Date of copyright.
date	created	Date of creation or manufacture of intellectual content if different from date.issued.
date	issued	Date of publication or distribution.
date	submitted	Recommend for theses/dissertations.
identifier		Catch-all for unambiguous identifiers not defined by qualified form; use identifier.other for a known identifier common to a local collection instead of unqualified form.
identifier	citation	Bibliographic citation for works that have been published as a part of a larger work, e.g. journal articles, book chapters.
identifier	govdoc	Government document number
identifier	isbn	International Standard Book Number
identifier	issn	International Standard Serial Number
identifier	sici	Serial Item and Contribution Identifier
identifier	ismn	International Standard Music Number
identifier	other	A known identifier type common to a local collection.
identifier	uri	Uniform Resource Identifier
description		Catch-all for any description not defined by qualifiers.
description	abstract	Abstract or summary.
description	provenance	The history of custody of the item since its creation, including any changes successive custodians made to it.

Element	Qualifier	Scope Note
description	sponsorship	Information about sponsoring agencies, individuals, or contractual arrangements for the item.
description	statementofresponsibility	To preserve statement of responsibility from MARC records.
description	tableofcontents	A table of contents for this item.
description	uri	Uniform Resource Identifier pointing to description of this item.
format		Catch-all for any format information not defined by qualifiers.
format	extent	Size or duration.
format	medium	Physical medium.
format	mimetype	Registered MIME type identifiers.
language		Catch-all for non-ISO forms of the language of the item, accommodating harvested values.
language	iso	Current ISO standard for language of intellectual content, including country codes (e.g. "en_US").
publisher		Entity responsible for publication, distribution, or imprint.
relation		Catch-all for references to other related items.
relation	isformatof	References additional phsical form.
relation	ispartof	References physically or logically containing item.
relation	ispartofseries	Series name and number within that series, if available.
relation	haspart	References physically or logically contained item.
relation	isversionof	References earlier version.
relation	hasversion	References later version.
relation	isbasedon	References source.
relation	isreferencedby	Pointed to by referenced resource.
relation	requires	Reference resource is required to support function, delivery, or coherence of item.
relation	replaces	References preceeding item.
relation	isreplacedby	References succeeding item.
relation	uri	References Uniform Resource Identifier for related item.
rights		Terms governing use and reproduction.
rights	uri	References terms governing use and reproduction.
source		Do not use; only for harvested metadata.
source	uri	Do not use; only for harvested metadata.
subject		Uncontrolled index term.
subject	classification	Catch-all for value from local classification system; global classification systems will receive specific qualifier.
subject	ddc	Dewey Decimal Classification Number

Element	Qualifier	Scope Note
subject	lcc	Library of Congress Classification Number
subject	lcsh	Library of Congress Subject Heading
subject	mesh	Medical Subject Headings
subject	other	Local controlled vocabulary.
title		Title statement/title proper.
title	alternative	Varying (or substitute) form of title proper appearing in item, e.g. abbreviation or translation.
type		Nature or genre of content.

Appendix B. Documents referenced

The following documents were referred to during the writing of this report. On completion of the report they will be filed in the Computer Department at LLGC.

Volume 1: LLGC documents and DSpace

1. LLGC – Digital Preservation policy and strategy – Draft – March 2003
   LLGC – Requirements for NLW DAMS - JWW
2. LLGC – Interview notes with Refugee Studies Centre and ThesesAlive
3. DSpace 1.1.1 System Documentation
4. "DSpace-An Open Source Dynamic Digital Repository" D-Lib article Jan 2003
5. DSpace Federation website:
   "Elevator Pitch"
   Introducing DSpace
   FAQ
   Planning for and Implementing DSpace
6. DSpace@Cambridge
   Website: Outline
   Project proposal
   Timetable
   Contact details
   Powerpoint presentation
   LEADIRS – Seminars website intro
   DSpace training workshop agenda Oct 2003
7. Metadata
   DSpace – Qualified Dublin Core metadata list
   Dublin Core Metadata Element Set – Reference Description
   Fedora – Dublin Core schema
8. DSpace at MIT
   MIT's DSpace Experience: A case Study
   Website: Community Development Guidelines
   DSpace Planning – Lessons learned (& Project Plan!)
   System Manager Job Description
   Useability test results Sept 2002
9. Comparisons and Evaluations
   UKOLN Metadata Resources – OSS Software
   University of Arizona – DSpace evaluation
   Sunsite – DSpace for dummies (installation instructions)
   Open Society Institute – A Guide to Institutional Repository Software
   ThesesAlive – DSpace and ETO-db Comparative Evaluation
10. DSpace misc documents
   Cornell: list of supported file formats
   E-mail on risk of technical problems with software versioning.
   Sourceforge: outstanding feature requests
   University MSH-Alples powerpoint
   Univ. Washington: Installing DSpace powerpoint
   MSH-Alpes DSpace home page
   Links useful during DSpace installation at KU

11.Browser compatability examples

12.DSpace – Related software

CNRI – HandleSystem Overview

Lucene – Search engine

SIMILE – Website intro

Introduction

History events

Volume 2: Fedora and Miscellaneous documents

13.Publications

"The Fedora Project – An Open-source Digital Object Repository Management System" D-Lib April 2003

The Mellon Fedora Project: Digital Library Architecture meets XML and Web Services

14.Fedora Technical Documentation

Technical Specification Dec 2002

Fedora Features

Fedora Digital Object Construction Guide

Fedora 1.2 Release notes and features (Dec 2003)

Versioning in Fedora

Administrator Documentation (GUI user guide)

Local Services User documentation

Web Exposure of the Fedora Access Service

API-M-Lite Interface Client User Documentation

Search Interface User Documentation

API-A Sample SOAP client

Client Command Line utilities

Installation and Configuration Guide

Demo Manual – description of Demo objects

15.Fedora – Presentations

Library of Congress – Corey Keith

Indiana University – Jon Dunn

DLF Forum 17/11/03

U. Virginia to JA-SIG 9/12/03

U. Virginia VLA conference 11/03

16.Fedora misc documents

Twiki – FAQ

Example SOAP client – Perl

Example SOAP client PHP

Linux/Unix installation tips

Fedora development priorites – April 2003

Development testbed descriptions

Public Opinion Polls and Digital Preservation: an application of Fedora – D-Lib Nov 2003

Deployment Testbed descriptions – Pilot test projects

Fedora at Northwestern – Planned test cases

"This Fedora's big enough for any DAM project" - Econtent article Oct 2003

17.E-mails - general

18.Tomcat/Java/JSP

Jakarta – Tomcat introduction

   "Servlets and JSP: An Overview"
   Tutorial – installing Tomcat
   How SOAP compares to REST

19. Fedora – University of Virginia
    Text Object Model Committee Recommendations
    Image Object Model Committee Report
    Issues for the design of Fedora Repositories
20. Fedora related software
    VTLS – VITAL
21. Miscellaneous documents
    InterPARES2 – project summary
    PRO – Digital Archive home
    "Data Preservation: Great data, but will it last" - Research Information
    SCOOP – automatic metadata generation
    PAD – related projects
    "Policy Enforcement for Complex Digital Objects"
    METS: Overview and Tutorial.
22. Other software – website overviews
    ARNO
    CDSware
    Eprints
    i-TOR
    MyCoRe