



# UNIVERSITY OF BIRMINGHAM



## **MIDESS Workpackage 8: Repository Implementation and Population**

### **Executive Summary**

The MIDESS Project is a JISC and CURL project, majority funded under the JISC *Digital Repositories Programme*. MIDESS explores the management of digitised content in an institutional and cross-institutional context through the development of digital repository architecture. The project addresses how support can be provided for the use of digital content in a learning and research context, in an integrated manner. The partners in the project are the University of Leeds, University of Birmingham, London School of Economics (LSE) and University College London (UCL).

This work-package describes the work undertaken to create prototype repositories at three of the partner institutions and to populate them with content. Key decisions, issues and outcomes are described for each repository, including how the collections were structured, the degree to which the repository met initial expectations, staffing implications arising from the choice of software and how the repository will be taken forward at the end of the project. An appendix gives screenshots from each of the software installations for a number of standard functions.

At the University of Birmingham, the project provided an opportunity to create a pilot institutional repository and thereby explore the long-term requirements and issues. Support for open standards was seen as essential, and there was particular interest both in storing and making available the outputs from the Library's digitisation programme and also in holding learning objects which could then be made available through a VLE. DSpace was selected as an open-source product which fitted well into the existing technical infrastructure at Birmingham. Sample files in various formats were

## MIDESS: WP 8 – Repository Implementation and Population

loaded, including digitised images derived from the University coin collection, playscripts, maps and video. An attempt was also made to load learning objects, complete with IMS/LOM metadata; however the facilities for metadata handling within DSpace did not make this a viable option. At the end of the project, Birmingham felt that DSpace did not meet their long-term needs, not least because of restrictions such as the lack of thumbnails, poor metadata support and the difficulty of modifying the default interface.

The London School of Economics chose to install Fedora, another open-source product. Fedora seemed to offer a powerful and very flexible structure within which to hold digital objects, together with multiple metadata datastreams where required. It also offered support for METS, for version management and for a wide range of protocols which offered the possibility of easy integration into the broader information architecture. LSE concentrated on migrating two major collections into Fedora: one comprised archival photographs associated with Malinowski's ethnographical fieldwork, the other a collection of digitised recordings of television programmes to support learning and teaching within the institution. In both cases, rich metadata was available in an external database, and METS was used to successfully create a datastream for ingest into Fedora. However, the work required to move it from being a pilot system to a fully functioning repository was too great; in particular, the lack of a satisfactory interface for end-user searching was a major limitation. However it was decided to retain the Fedora repository in the short term as a storage medium for large multimedia files while potential alternatives platforms are investigated.

Unlike Birmingham and the LSE, the University of Leeds had secured additional funding in order to establish a multimedia repository service for the University. This permitted installation of a commercial product, offering the advantage of richer in-built functionality, including a fully-developed interface to support resource discovery. Endeavour Curator was selected because of its easy-to-use public interface and its support for a wide range of metadata standards including MODS and METS. In the event, both Endeavour and the rival software company Ex-Libris were purchased by a private equity company late in 2006. As a result, development work on the Curator platform ceased and Leeds was obliged to migrate to DigiTool in the first half of 2007. This had a severe impact on project timescales, and delayed the loading of live data; however a wide selection of test data – digitised images, audio and video files - was successfully loaded into both Curator and DigiTool. The intention remains to offer a full service on this platform after the project has ended.

## Contents

<b>Introduction</b> .....	<b>4</b>
<b>The University of Birmingham Digital Repository</b> .....	<b>5</b>
Choice of Repository .....	5
Support for metadata schemas .....	5
Choice of digital content.....	6
Collection structure .....	6
Experience of using DSpace .....	8
Staffing Requirements and Training Requirements.....	8
Future of the Digital Repository at the University of Birmingham .....	8
<b>Implementation of the Digital Repository at the London School of Economics (LSE).</b> .....	<b>10</b>
Choice of Repository .....	10
Support for metadata schemas .....	11
Choice of digital content and collection structure .....	11
Resource discovery .....	12
Experience of using Fedora.....	13
Staffing Requirements and Training Requirements.....	14
Future of the Digital Repository at the LSE.....	14
<b>University of Leeds digital repository</b> .....	<b>16</b>
Choice of Repository .....	16
Support for metadata schemas .....	16
Choice of digital content.....	19
Collection structure .....	20
Browsing and Searching within Curator.....	21
Structure and Deposit in Curator .....	21
Experience of using Curator .....	22
Migration to DigiTool.....	23
Structure of data and collections within DigiTool.....	24
Browsing and Searching in DigiTool .....	24
Experience of using DigiTool.....	25
Overall software performance .....	26
Staffing Requirements and Training Requirements.....	27
Future of the Digital Repository at the University of Leeds.....	28
<b>General Conclusions</b> .....	<b>29</b>

### Introduction

This work-package describes the work undertaken to create prototype repositories at three of the partner institutions and to populate them with content. The University of Birmingham chose to install DSpace, the London School of Economics (LSE) selected Fedora, while the University of Leeds purchased the commercial repository product Endeavour Curator (subsequently replaced by DigiTool).

Key decisions, issues and outcomes are described for each repository. Areas covered include:

- The reasons for the choice of repository software platform
- Support for metadata schemas
- How the collections were structured
- The experience of using the software platform
- The degree to which the repository met initial expectations
- Staffing implications arising from the choice of software platform
- How the repository will be taken forward at the end of the project.

An appendix gives screenshots from each of the software installations for a number of standard functions. The in-depth work undertaken on information architectures, inter-institutional resource sharing and digital object re-use is described in detail in Work-packages 9 and 10 and is not repeated here.

At each site, a working group was set up to guide and coordinate the work undertaken. At Birmingham, membership was drawn from across Information Services in order to ensure a thorough evaluation of institutional requirements for a digital repository. At LSE, the working group included staff from the Library and from the Centre for Learning Technology (CLT): CLT involvement was important not only because they already held many of the digital objects which would be used in the project, but also because of their expertise in the use of digital objects to support learning and teaching. Because Leeds intended moving as quickly as possible to a live repository service, membership of the working group there included not only staff from the Library and from Information Systems Services but also significant academic representation, drawn from several different schools.

Communication and coordination between the three sites was facilitated by a regular email “bulletin” from each site (weekly during the main implementation phase of MIDESS), distributed to all project staff. Partner meetings, held at 6-monthly intervals, enabled in-depth sharing of experience and expertise through formal presentations and discussion. A number of smaller, more specialised sessions were also organised to explore particular issues and themes.

### The University of Birmingham Digital Repository

#### Choice of Repository

The choice of a repository at Birmingham was primarily determined by the following requirements.

- A pilot system was required to enable the University of Birmingham to evaluate functionality required for multimedia repositories.
- Open standards were required where possible.
- The solution needed to be sympathetic with existing Java application support available within the University of Birmingham
- Budget limitations excluded any platform requiring additional cost for licenses etc. typical of commercial repositories.

The University of Birmingham's objective was to use the MIDESS project as an evaluation of the features and functionality that would be required for any future repository at the University. In particular it provided an opportunity to explore the emerging metadata standards and draw up a full requirements specification for any future repository/system. DSpace, an open source product, was chosen for its anticipated adherence to open standards and relative ease of installation and use.

The goals for Birmingham therefore included:

- Evaluating DSpace as an open source repository for holding images
- Evaluating metadata standards, e.g. Dublin Core, METS etc. with a view to determining which were most likely to meet future needs not only for images, but for the full range of potential digital materials that may need to be housed in a repository.
- Experimenting with image scaling to decide the optimum file size and graphics format for the 28,000 digitised images already in existence.

#### Support for metadata schemas

For some classes of material which it was hoped to ingest, it was relatively easy to identify a suitable metadata schema. For example, EAD is the de facto standard for archival materials and IMS/LOM is widespread for learning objects. All the MIDESS partners had an interest in holding learning objects within the repository, and this had been suggested as an area where the sharing and re-use of digital objects would be of considerable value. In consequence, considerable effort was invested in investigating how DSpace might be used as a learning objects repository and in particular ways of handling the metadata available as part of an IMS content package. However DSpace is extremely restrictive in the metadata schemas it supports – essentially it is only capable of handling Dublin Core. Work-package 4b details the efforts that were made to investigate the mapping of IMS CP onto qualified Dublin Core. However this proved to be a dead end, and it was accepted that, for this reason, DSpace is unsuitable as a learning objects repository.

Therefore, for all the objects loaded into DSpace, any metadata available had to be tailored to Dublin Core – usually qualified Dublin Core. This operation in itself generally required a metadata specialist, and still resulted in a significant and frustrating lack of richness and specificity. This can be seen particularly with the coin collection, for which various additional metadata elements had to be specified in order to hold information essential to an understanding of the actual coins e.g. the requirement to distinguish between features on the obverse and reverse sides of each coin.

### Choice of digital content

Content loaded into DSpace was largely drawn from material generated by digitisation projects previously undertaken within the University of Birmingham. Since the repository was intended as a pilot, only selections of collections were loaded rather than entire collections: it was felt that loading complete collections would be unproductive until a decision had been made regarding the final choice of a digital repository for the University.

The collections created and populated include:

#### *The Artworks Image Library*

This collection holds a few images from the Barber Institute of Fine Arts collection, provided by the Institute's slide library. A sample image is in Appendix 1, fig.7.

#### *MIDESS metadata presentations*

These were two lectures given by Adrian Dover on "The MIDESS project – its aims and goals" given to an audience from Birmingham University Library.

#### *Shakespeare Institute Library Scripts*

A photocopy of an unpublished continuity script of Christine Edzard's 1992 film *As you like it*.

#### *Learning Object collection*

A pilot collection of learning objects of varying complexity.

#### *Urban Morphology Research Group historical maps and urban images*

Various image files including maps and photographs. A sample image is in Appendix 1, fig.8.

#### *Special Collections*

This collection included digitised images of medieval manuscripts held in Special Collections and Archives - part of Information Services. A sample image is in Appendix 1, fig.9.

#### *Roman Imperial coins & Roman Republican coins*

Images in these collections are a small sample from the coin collection held in at the University of Birmingham. Details about the coin collection at Birmingham are available at <http://www.coins.bham.ac.uk/> . A sample image and the associated metadata can be seen in appendix 1, fig.6.

For this coin collection, no pre-existing metadata schema could be identified which would provide an appropriate structure to hold all the available data in the manner required. Qualified Dublin Core was therefore used, with local modifications to meet the specific requirements. Fuller details are available in Work-package 4b.

### Collection structure

The University of Birmingham subdivided the digital material in the repository to mirror the school/departmental/institute structure of the University. In the DSpace vocabulary, these are communities (and sub-communities), each of which has a community page and any of which can own collections. Sub-communities and collections can be browsed by selectively moving down the branches of the hierarchy. It is typical of DSpace implementations to mirror the structure of the parent organisation and this results in a very hierarchical presentation to the end user. However, this structure does allow the user to see the structure of the schools, departments, Institutes, etc and to

## MIDESS: WP 8 – Repository Implementation and Population

establish where in the hierarchy responsibility for the digital material actually exists i.e. which community is “responsible” for which collection.

When this hierarchy is presented for browsing, a lack of highlighting indicates that material actually exists at that level of the hierarchy within the repository. In the following illustration, the collections “Metadata presentations” and “Scripts” are shown in normal text because they contain digital materials, while the remainder of the hierarchical structure is shown in bold, as shown in fig.1.



Fig.1. Part of the communities and collections menu

Alternative browse modes are available, including title, author, subject and date, as illustrated in fig.2.



Fig.2. Alternative browse modes in DSpace

Since there was no intention to deliver a live service to users from this particular implementation of DSpace, no attempt was made to modify this default structure and behaviour and no formal evaluation took place from an end-user perspective.

Searches can be carried out across the entire repository, an individual community/sub-community or in just one collection. It is also possible to browse by title, author, subject or by date at each of these levels.

Screenshots illustrating the DSpace resource discovery interface are shown in appendix 1.

## MIDESS: WP 8 – Repository Implementation and Population

### Experience of using DSpace

DSpace initially seemed to perform in line with expectations of the MIDESS staff at the University of Birmingham. It was expected - and found in practice - that DSpace was a relatively simple system to install and use but which provided much less flexibility and functionality than other digital repository software such as Curator, DigiTool or Fedora. The primary restrictions were the inability to tailor the interface, the lack of thumbnail searching and the poor interoperability with other digital repository systems. Thus while DSpace provided an excellent repository system for storing text-based digital material, the inability to visually browse multimedia material was felt to be a major handicap.

During the installation of the repository, issues were primarily with the set up and installation of the server rather than with the DSpace software. Initially the software was installed on a PC rather than a server to ensure that there were no major problems with the installation. Installation of the software on a server took longer than expected because of the delay in purchasing and installing a server of sufficient capacity to support the repository while at the same time ensuring that the repository was accessible from the Internet.

In the course of the project, other, specific limitations of the software became apparent. It was felt that its exporting functionality could specifically be improved since there were significant problems with this functionality during the lifetime of the project. METS support could also be improved. There were issues around the lack of ability to arrange the hierarchy within the system for display of items and the lack of IMS support for potential learning objects. Overall, DSpace failed to offer the flexibility and breadth of functionality that Birmingham required for its repository system.

For system documentation, Birmingham used the standard DSpace documentation at <http://www.dspace.org/> and a DSpace "How to" walkthrough published by MIT and available at: [http://cwspace.mit.edu/docs/PeopleAndOrganizations/DSpace/Dspace\\_HowTo-Linux.txt](http://cwspace.mit.edu/docs/PeopleAndOrganizations/DSpace/Dspace_HowTo-Linux.txt)

### Staffing Requirements and Training Requirements

During the lifetime of the MIDESS project at Birmingham only part-time technical support was required for the support of the repository as DSpace seemed to require relatively low maintenance on an ongoing basis (unlike e.g. Fedora). It should however be noted that some issues could have been addressed had staff resources been available to undertake significant modification and development of the underlying code – this is true of any open-source product.

It is also clear that a digital repository requires significant investment in the creation of metadata. Identifying (and on occasion creating) suitable schemas and creating mappings between schemas are both highly complex tasks. Even creation of metadata for material to be ingested requires significant expertise. During the MIDESS Project, resource was available for this purpose, and one of the ongoing challenges is to identify such resource on an ongoing basis. It has been suggested that it might be possible to draw this from the pool of cataloguers currently employed in the Library (though their expertise currently resides in the use of MARC and some re-skilling in Dublin Core and other XML schemas would be required).

### Future of the Digital Repository at the University of Birmingham

The DSpace digital repository deployed as part of the MIDESS Project enabled the University of Birmingham to examine the functionality available within repositories and to explore the issues that would arise in installing and supporting a long term digital repository. It is readily accepted that a repository is required to store digital material since there are large quantities of digitised images which

## **MIDESS: WP 8 – Repository Implementation and Population**

need to be preserved. However the exact details of how a repository sits within the infrastructure of the Library/University have not yet been determined with any clarity.

Birmingham is, therefore, currently considering the next steps in terms of software platform, function and support for a multimedia repository. It is unlikely that DSpace will be used. A pilot repository for research publications (UBIRA – University of Birmingham Research Archive), previously running on Eprints 2, is being expanded and relaunched as a live service using Eprints v.3, with separate silos for research papers, for other research material and for e-theses. It is likely that this same platform will be developed to support a parallel repository for multimedia materials, running on the same server. Initially, this is likely to be targeted primarily at the needs of Information Services, but it would be hoped to expand the user base in due course.

# Implementation of the Digital Repository at the London School of Economics (LSE).

## Choice of Repository

The reasoning behind the selection of Fedora as the repository of choice at LSE was that Fedora had the following key features:

- Powerful digital object model whereby the digital objects, or units of information, in Fedora may combine any number and any variety of data streams, either internal or external to the Fedora system.
- Extensible metadata management: any number and variety of metadata formats may be stored as data streams, alongside content, in a digital object
- Support for version management – Fedora stores a history of all modifications to digital objects
- Support for rich metadata and ability to use METS.
- Support for OAI-PMH
- Potential ability to link different databases in the institution to provide a single combined repository.
- Felt to have strengths in Web Services, security and good compatibility with Shibboleth.
- The ability to include relationships between objects and deal with complex digital objects. It can automatically assess parent/child relationships.
- Felt to have better provision for preservation metadata than Dspace.
- Has system plug-in for Persistent ID generation.

These requirements are in line with the functional and technical requirements specification prepared as part of MIDESS Workpackage 2 and available on the MIDESS website at [www.leeds.ac.uk/library/midess/Workpackage2](http://www.leeds.ac.uk/library/midess/Workpackage2)

Initial installation was of Fedora version 2.1.1. This version was superseded by Fedora 2.2 , released in January 2007, though no significant changes were noticed in terms of out-of-the-box functionality. The Fedora system was restricted via IP address to staff working on the repository at LSE and those MIDESS staff at other Institutions (Leeds, Birmingham, UCL).

Initially, Elated was investigated - an open-source application developed to work on top of Fedora – as an easier and web-based interface for ingest, repository management and resource discovery. According to its development team, Elated “could be used as a digital assets management system, an institutional repository, or to meet other collection archiving, publishing and searching needs”. However, there proved several obstacles to its use within the LSE environment including:

- Lack of compatibility in several areas between data input directly into Fedora and functionality for accessing/exploiting that data in Elated
- Erratic searching capability in Elated
- Malfunction of various software features
- Lack of sophistication and flexibility in metadata handling
- Limited support for collections
- Not possible to discover the permanent ID (PID) for objects created through the Elated interface

It was therefore decided to discontinue use of Elated and carry out subsequent work through the standard administrative interface available within Fedora, an approach which demands greater technical competence and knowledge. A [report](#) on the work carried out on Elated is available on the MIDESS web-site.

### Support for metadata schemas

Upon ingest, metadata from the Dublin Core metadata datastream plus Fedora System metadata (i.e. administrative metadata) are indexed in a relational database within Fedora and can be searched using the native search interface. If no DC datastream exists, then the system creates a simple DC datastream consisting of dc:title and dc:identifier elements only. DC is handled as an “Internal XML datastream”, with XML validation. Other metadata schemas can be stored as either “Internal XML datastreams” with validation or as “Managed content” – the same format which is also used for internal storage of digital objects – and there are no formal restrictions on content.

The search interface permits the 15 DC elements to be searched explicitly, along with various elements of administrative metadata. For the objects retrieved, the title and description are displayed, and also the persistent identifier (PID) which functions as a link to the “object profile view”. This intermediate screen permits access to the “dissemination index” and also to the “item index” for the object. The item index in turn allows any attached metadata records to be displayed, and also provides a link giving access to the actual digital object. These stages are illustrated in appendix 2. fig.11-20.

From this brief description, several key features relating to the handling of metadata within Fedora can be discerned. On the one hand, Fedora treats each metadata record associated with the digital object as an object in its own right, accessible from the item index. Fedora is therefore very flexible in its approach to storing metadata and able to handle any schema, whether standard or bespoke. However, retrieval via the search interface within Fedora is dependent on the presence of a Dublin Core datastream; without this, access is so limited as to make the material almost invisible.

### Choice of digital content and collection structure

The LSE Centre for Learning Technology (CLT) had developed the CLT media database in order to store and manage some of the video and audio content it had digitised for teaching and learning purposes. This content comprises mostly digitised lectures and television programmes (recorded under the ERA Licence).

705 objects in this collection were ingested to Fedora, using METS as content packaging tool for the creation of Submission Information Packages (SIPs). The resulting Fedora objects (Ise:2000 to Ise:2705) include CLT metadata, Dublin Core metadata (derived from the CLT Metadata in order to facilitate discovery via the Fedora Search Interface), and the audio or video objects (in Quicktime, Real Media or Windows Media format). Parent/child relationships were also recorded between digital objects where necessary.

Video files were not ingested as binary content, but carried instead a reference to the file as held on the LSE’s streaming server (which permits more efficient network delivery of the content). This used the ‘Redirect’ facility within Fedora described in Fedora official documentation as follows: “Fedora will send clients a redirect to the URL you specify for the datastream. This is useful in situations where the content must be delivered by a special streaming server, it contains relative hyperlinks, or there are licensing or access restrictions that prevent it from being proxied.” In this instance, both conditions applied:- the data was stored on a streaming server, and also the ERA licence under which many recordings had been made explicitly prohibits delivery to off-campus users.

Towards the end of the project, a further 66 programmes, where the original recordings were held by University College London’s School of Slavonic and East European Studies, were identified for digitisation and ingest into Fedora. This was to explore inter-institutional resource sharing and interactions with the ERA licence about the Cold War and these aspects are discussed in Work-package 9. Metadata for these programmes had previously been held in a series of flat html web

## MIDESS: WP 8 – Repository Implementation and Population

pages; these were manually edited in order to provide an xml file of qualified Dublin Core records which were conformed in most respects to the EBU Core application profile and were also compatible with other CLT metadata.

The LSE Archives hold a substantial collection of material relating to the anthropologist Bronislaw Kasper Malinowski, including a large collection of black and white photographs from his fieldwork to the Trobriand Islands (1915-1918). These photographs had been digitised in recent years. Part of this collection was ingested to Fedora (lse:2721 to lse:2772) using METS as a content packaging tool. The resulting objects include full EAD metadata record, the JPEG and TIFF images, and Fedora's default Dublin Core record.

The METS structure map was used to establish parent/child relationships between the 3 hierarchical levels of "series", "sub-series" and "illustrations" into which the Malinowski collection is organised. The XML structure supporting the Fedora "isMemberOfCollection" relationship is illustrated in Appendix 2, fig.16 (the datastreams for CLT metadata and for Dublin Core metadata for the same object are shown in figures 17-18).

One METS manifest was used to encode the collection, zipped together with the associated folders and files. However, while the whole Malinowski collection could in theory be wrapped into one ingest package (SIP), DirIngest does not process zip files larger than about 600MB. The initial SIP which was put together, and which includes the JPEG and TIFF images as well as the EAD metadata output from the LSE Archives CALM system, thus represents only part of the collection, encoding the first 2, of 34 sub-series (Malinowski/3/1 and Malinowski/3/2 only). This SIP was successfully ingested into Fedora using the Fedora Directory Ingest Service or DirIngest

Fuller details of the ingest process into Fedora are available at:  
[http://www.leeds.ac.uk/library/midess/clt\\_media\\_database\\_to\\_fedora\\_ingest.pdf](http://www.leeds.ac.uk/library/midess/clt_media_database_to_fedora_ingest.pdf)

With only two collections to ingest into Fedora – and those very distinctive – the structure of material within Fedora was determined by this content. The structure within the Malinowski collection does however illustrate how a more complex hierarchy, with collections nested within collections, could be built if required.

### Resource discovery

Fedora is explicitly presented as open-source software for creating a repository infrastructure which can then be used to store, manage, and deliver digital content. However, it is not a complete digital repository solution in itself, but simply an underlying architecture upon which other applications can be built. For instance, while Fedora 'out of the box' is a powerful piece of software, it does not come with a web or user interface, though it does have a simple search interface, enabling searching and browsing of the Fedora repository. There also exist some open-source add-ons which provide or include a web interface options, such as Elated and Fez, but these are separate products which need to be installed individually.

It was clearly desirable to be able to present the collections within Fedora to the end-user through a web interface, and initial hopes centred on Elated. When this proved a dead-end, it was decided to implement Fez, but this too proved problematic. Initial attempts to get Fez working in Autumn 2006 failed. After the upgrade to Fedora 2.2 in January 2007, renewed efforts were made, but with no greater success. The conclusion drawn was that greater technical expertise is required to implement Fez than was available to the project at that time.

With no public interface, it would still be possible to use Fedora to manage the digital objects – with resource discovery provided by an alternative front-end system – but this was not viable within the constraints of the actual project.

## MIDESS: WP 8 – Repository Implementation and Population

The Fedora repository at LSE does not therefore provide the ability to easily browse the material in the repository. The default search screen, illustrated in fig.3, is designed to meet the needs of a repository manager, providing either a general search across all fields or the ability to restrict it to specific Fedora and Dublin Core fields.

The screenshot shows the 'Fedora Repository Find Objects' search interface. It features a 'fedora' logo on the left and a search form on the right. The search form includes two input fields: 'Search all fields for phrase:' and 'Or search specific field(s):'. Below these is a 'Maximum Results:' dropdown menu set to '20' and a 'Search' button. On the left side of the search form, there is a section titled 'Fields to display:' with a grid of checkboxes for various fields. The 'pid' checkbox is checked, and the 'title' checkbox is also checked. The other fields listed are: label, fType, cModel, state, ownerId, cDate, mDate, dcmDate, bDef, bMech, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights.

Fig.3 Fedora search screen

The displays of retrieved records are similarly opaque to the non-specialist and are oriented to management of the objects in the database. There are generally options for each of the datastreams associated with the object and metadata datastreams (both administrative and descriptive) are displayed in parallel with those for the actual digital object. Appendix 2, fig.11-20 give examples of the various screens involved. Quite clearly neither the “method index” (fig.12) nor the “item index” (fig.14) are suitable for display to the end-user.

If the MIDESS project had had a greater budget then an additional programmer could have been employed to undertake further work on providing a public resource discovery interface for Fedora. However, as matters stood, there was no easy way to overcome the difficulties encountered in this area.

### Experience of using Fedora

Fedora is undoubtedly a complex product with rich functionality to support a wide range of repository functions. Within the framework of the MIDESS Project, it was only possible to explore a limited sub-set of that functionality. Usability is not one of Fedora's strengths, particularly Fedora “out of the box” which does not include a web interface enabling any user, whether administrative or public, to fully interact with the repository. The Fedora Search Interface is clearly limited in its functionality, and the only administrative tool, the Fedora Admin Client, is mostly limited to creating, ingesting, and exporting, purging, searching and retrieving Fedora data objects.

For this reason, early efforts concentrated on Elated, which it was hoped to use as a front-end tool to exploit Fedora. When this failed, efforts concentrated on ingest of material into Fedora and

## MIDESS: WP 8 – Repository Implementation and Population

on provision of a public resource discovery interface through Fez. Ingest proved idiosyncratic, complex to master and slow, but ultimately successful and demonstrated the underlying power and flexibility of Fedora. Equally the lack of success in installing Fez demonstrates how time-consuming it can be to build fully operational implementations with Fedora and the level of technical expertise required to achieve this. The MIDESS project officer was not a computer programmer and this undoubtedly limited what could be achieved.

The difficulties encountered in installing either of the two open source front-ends (Elated and Fez) meant that testing could not be carried out with end users and ultimately undermined the viability of further development of Fedora as a repository platform within LSE, despite the quantity of material successfully ingested.

### Staffing Requirements and Training Requirements

Fedora's complexity clearly requires the availability of technical skills (including programming) to get the most from it. The product appears to be developed from a technical programmer's point of view and with the expectation that local expertise will be available to create a working implementation.

Documentation produced by [Fedora](#), the [RepoMMan](#) project team and the [Paradigm](#) project team was consulted during the work. The Fedora mailing lists were also used and found to be of some use although it was felt that technical questions were given a higher priority than user-oriented questions when supplying answers. Detailed documentation, based on the LSE Fedora installation, was therefore produced by the MIDESS project officer to facilitate training of future Fedora administrators at LSE and this is available at: <http://www.leeds.ac.uk/library/midess/FedoraUserDocumentation.pdf>

Working with repositories has also highlighted the issues relating to metadata at LSE. Substantial progress was made in understanding and manipulating METS as a tool for ingest because this in-depth knowledge was essential for ingest of material into Fedora. It became clear that the ability to map between metadata schemas (particularly from the in-house CLT format and from EAD to qualified Dublin Core) is also a key requirement for migrating material into any repository structure. During the project itself, it was fortunate that the expertise in metadata required was available through a staff member who was on the local MIDESS working group at LSE and it is planned to form a small group to examine the broader metadata issues identified as part of the MIDESS project.

LSE did not need to identify training requirements for end users since the lack of a public resource discovery interface prevented direct access to Fedora by end users.

### Future of the Digital Repository at the LSE

LSE is looking at content management systems and repositories overall with the intention of establishing a way forward which will meet the diverse requirements identified. LSE undertook a detailed evaluation of the repository software after completion of Work-package 3 as part of an evaluation for the long term strategy for a repository. This is available at [http://www.leeds.ac.uk/library/midess/Criteria\\_for\\_digital\\_repository\\_software\\_at\\_LSE.pdf](http://www.leeds.ac.uk/library/midess/Criteria_for_digital_repository_software_at_LSE.pdf)

One key driver is the intention to move additional video materials to digital format. Technically, details of these could continue to be held in the CLT database (built in-house with no established standards and from which material was drawn for the MIDESS project). However there are clear advantages to be gained in holding this material within a repository framework, particularly with respect to both preservation and access. There is no shortage of other material which could

## **MIDESS: WP 8 – Repository Implementation and Population**

potentially be stored in the repository as well, although there are resource implications to undertaking significant ingest activity.

The MIDESS Project allowed LSE to investigate the potential of Fedora as an institutional repository with a specific focus on how the software deals with multimedia and image collections. At the outset of the project LSE was considering Fedora as a possible solution for its institutional repository. Experience from MIDESS has suggested that Fedora required a substantial investment of resources and a high level of technical expertise to produce a working repository. Moreover, during the course of the MIDESS project, significant developments were made to the Eprints software, which improved its ability to manage image and multimedia files. In the light of this and the difficulties of getting a working user interface to Fedora, development work on the Fedora platform has ceased. The repository and its contents remain on a server at LSE (though not publicly accessible) and the resulting documentation has also been maintained. LSE will be now exploring the possibility of using Eprints to manage multimedia and images.

### University of Leeds digital repository

#### Choice of Repository

The choice of repository at the University of Leeds was based primary upon the following criteria:

- **Functionality.** Based upon the user needs analysis, a wide range of potential functionality was required since an institutional repository rather than a departmental or subject based repository was proposed.
- **Easy to use Resource Discovery System (user interface).**
- **Support for a variety of metadata standards.** The range of digital collections indicated a potential requirement for schemas such as EAD, MODS and MARCXML.
- **The repository should not be technically complex** since a technical support post was available to support the repository for only three days a week.

The detailed functional specifications for the digital repository are set out in detail in [MIDESS Workpackage 2](#).

Based upon these criteria and the feedback from a detailed evaluation study and series of presentations, two digital repository software products achieved the highest scores from the evaluation process: - Endeavor Curator and ExLibris DigiTool, both of which were commercial products. Once detailed pricing had been obtained from both vendors, Curator was selected as offering the best long-term value-for-money solution and this is the digital repository software that was initially adopted by the University of Leeds.

Other attractive features within the Curator software that differentiated it from other digital repositories evaluated were:

- **Hierarchical structure** allowing collections added to the repository to be subdivided into a number of smaller collections.
- **Easy to understand deposit facility**
- **Ability to integrate with Virtual Learning Environments**, specifically Blackboard and WebCT.

#### Support for metadata schemas

For an institutional repository, flexibility in the metadata schemas supported was seen as essential and Curator was one of the few repositories that natively supported a wide range of metadata schemas. However multiple metadata schemas within a repository can create potential problems of interoperability when searching across collections since the elements within each metadata schema can be so different.

Curator's solution to the problem of searching across multiple collections (which it calls "repositories") is by making the assumption that each metadata schema has elements which can be 'mapped' to a common metadata schema for the purpose of searching.

Thus within Curator there is a "master" metadata schema (which is usually unqualified Dublin Core) and other metadata schemas within the repository have their elements mapped to this Unqualified Dublin Core. This Master Metadata schema within Curator was called the *Collection Manager*. Elements (fields) in other Metadata Schemas are then "matched" to the most appropriate elements in the Collection Manager.

## MIDESS: WP 8 – Repository Implementation and Population

Thus searching on the element 'Creator' in the collection manager schema would also find matching data in the 'Author' element of collections using EAD and 'Heading' in another bespoke metadata schema providing that the appropriate elements (Author and Heading) had been mapped to the 'Creator' element in the collection manager.

In a Curator repository where metadata schemas are used that differ radically from unqualified Dublin Core, then it is probable that there will be elements in those schemas which do not match elements in the Collection Manager schema. Under these circumstances it would be difficult to search for those elements specifically. However there is also the option for a free text search across all metadata elements, either throughout the entire digital repository or alternatively through a specific collection and this provides a way of searching for any unmapped element. Although free text searching can potentially retrieve a large number of citations (since it searches all elements), the number of matching results returned can be limited by restricting the search to specific collections rather than searching the entire contents of the repository.

It is possible in theory to use a bespoke Collection Manager which contained all the elements from all the different collections in the digital repository. However, given that metadata schemas such as EAD and LOM contain large numbers of elements and that under these circumstances all elements from all possible metadata schemas would need to be included (including those from bespoke metadata schemas) then there could potentially be hundreds of elements to select from for a search based upon a specific element. Adding new bespoke metadata schemas would also be a problem since new elements would be needed to be added to the collection manager each time a new metadata schema was added which contained elements not currently supported in the existing bespoke collection manager schema.

To further illustrate this mapping, consider the case where three metadata schemas exist within the Curator repository, viz.

- Unqualified Dublin Core
- EAD
- A Bespoke Metadata Schema (created by the digital repository developer)

The Collection Manager is configured as unqualified Dublin Core and the other two metadata schemas are mapped to the elements in the unqualified Dublin Core metadata schema. Thus in the example below:

'Author' in the EAD metadata schema and 'Created By' in the bespoke metadata schema can be mapped to 'Contributor' in the Unqualified Dublin Core.

Similarly 'Archival Description' in EAD is mapped to 'Description' in Unqualified Dublin Core

and

'Date of creation' in the bespoke metadata schema is mapped to 'Date' in Unqualified Dublin Core.

Under these circumstances, searches on the element 'Contributor' in the Collection Manager would deliver matching data from the elements 'Author' in the EAD metadata Schema and 'Created by' in the bespoke metadata schema, etc.

## MIDESS: WP 8 – Repository Implementation and Population

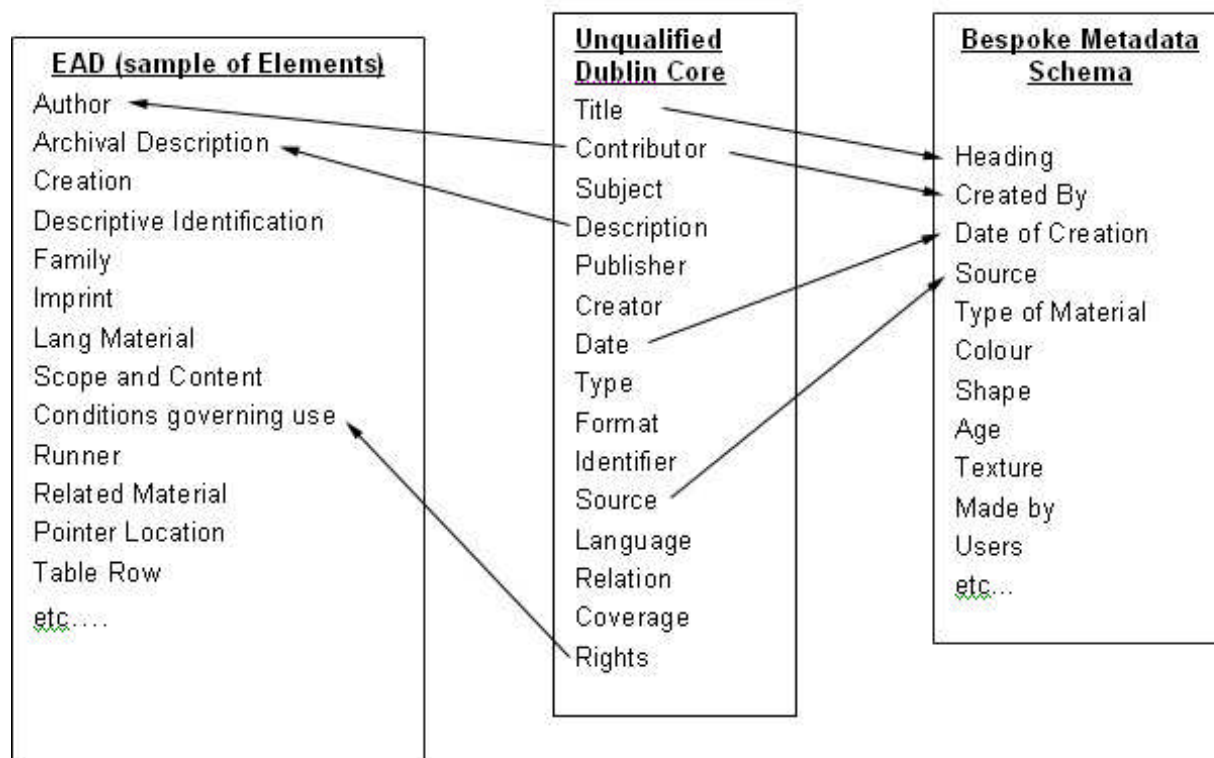


Fig. 4. Mapping between Metadata schemas

The following metadata schemas were supported by Curator version 4.0.

- Unqualified Dublin Core
- Qualified Dublin Core
- EAD (Encoded Archival Description)
- TEI (Text Encoding Initiative)
- SCISERV (a science-specific metadata schema).
- MARCXML
- MODS (Metadata Object Description Schema).
- METS was expected to be available in version 4.1 of Curator when released

### Choice of digital content

The survey of potential users conducted within work-package 3 had identified a wide variety of digital content within the University. It was decided that an important consideration for Leeds – and the project more generally – was to assess how these different types of content could be handled by a repository and the associated issues. Therefore, a range of material would be loaded as part of the initial pilot. The material actually loaded and used for testing included:

- High resolution images and a video from Media Services  
Media services currently had an ageing server on which these were stored and it was hoped that the Repository could be used to store their substantial collection of high resolution images and video files that are used throughout the university for academic and promotional purposes. In consequence, Media Services required some specialised functionality from the repository including:
  - The ability for users of the repository to view the high resolution images but not download them. (it was planned to achieve this via the software product Zoomify (<http://www.zoomify.com/>)).
  - The ability for chosen users only to download high resolution material.
  - A measure of the number of times the digital files stored in the repository had been downloaded to ensure that 'stock' images of the university were not used too often.
  - It was expected that videos would run from the streamed server for fast playback.
- Images from the collection of physics instruments  
The School of Physics has a large collection of physics instruments dating mainly from the early 20<sup>th</sup> Century and wished to make digital images of these instruments more widely available to the public. Several hundred files were supplied on DVD in high resolution TIFF. One image was loaded for testing purposes plus an image from the Bragg laboratory notebooks which were also recently digitised.
- A sound file from the Voices Project  
The School of English worked with the BBC on a project collecting voice dialects from around the UK. A safe and secure storage location is required to store the original sound recordings of this material which are currently stored in digitised form on a removable hard disk at the School of English at Leeds. The sound files are in windows WMV format.
- Images from the Medical Slide collection in the School of Medicine  
The School of Medicine has a very large collection of medical slides which need to be shared amongst medical staff at the University of Leeds. It is expected that access to the material will be heavily restricted with only authorised users within the School being able to see it. Given the nature of the material, subject matter experts (such as medical staff) will be required to detail the metadata. A total of 50 slides have been provided by the medical school and loaded as part of the pilot.
- An image of a medieval manuscript (at different resolutions) from the Library's Special Collections section  
Leeds University Library's Special Collections has recently digitised illuminated pages from its collection of medieval manuscripts. It is proposed that the collection be stored both in a low resolution form with universal access and a high resolution form to which access is restricted. It is expected that Zoomify will be used for this purpose. As well as a requirement for multiple versions of the same object (the individual page), it was also a requirement to be able to link and navigate all pages from the same manuscript. Digitool can deliver this via "parent-child"

## MIDESS: WP 8 – Repository Implementation and Population

linking, although this is unfortunately a manual process which has to be undertaken post-ingest.

Substantial work was undertaken on preparing further data for loading, including preparation of complete files for loading the Physics and Medieval Manuscripts collections; however technical problems with the repository software prevented this work from being completed within the timeframe of the project. As the project became more widely known across campus, there were approaches from a number of other academic departments about the possibility of using the repository, and these will be taken forward as the repository moves forward to a live service.

Consideration was also given to storing PDF files of journal articles and book chapters digitised under the recently extended CLA licence. However the terms of the licence prohibit access by any students except those registered for the specific module for which the material has been registered, and this restriction extends to public availability of the metadata. This could perhaps be achieved by full integration of authentication and authorisation mechanisms for the repository with other relevant campus information systems, but was not achievable within the limits of the MIDESS Project. Work on this was therefore deferred until a later time.

### Collection structure

Leeds decided to adopt a structure which organised the digital material on the basis of the collections identified for loading (Physics Collection, Medical Slide Collection etc). Within each collection, material is further subdivided by the type of material in the collection (image, video, sound, text etc). These collections were presented for browsing in the resource discovery interface via a brief description and thumbnail image, as illustrated in fig.5.

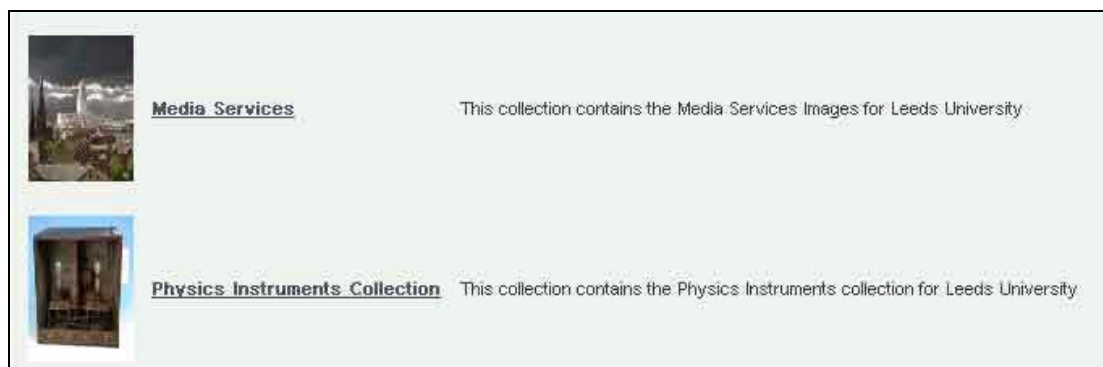


Fig.5. Presentation of collections on the Curator home page

Reasons for this included:

- There were a small number of collections and so these could be easily represented on the initial repository home page. This permits the user to easily browse the entire list of collections available without resorting to the search box to find a collection.
- If collections are subdivided by department/school/faculty etc, subjective decisions must often be made about the location of a collection where the collection is linked to more than one department (for example the Bragg physics collection derives from the School of Physics but is used by History and Philosophy of Science).

## MIDESS: WP 8 – Repository Implementation and Population

- Mirroring the organisational structure of the University could also encourage a very proprietary attitude to the content and undermine the desire to create a central service for the whole University.

This structure - in which all the collections are visible on the initial page of the repository - works well when there is a small to medium number of collections within the repository since it is easy to browse through the complete list of collections and has worked well at this pilot stage. Curator did however also support an infinite number of nested hierarchical levels which could be used to subdivide each collection into one or more sub-collections and it is recognised that some reorganisation may become necessary as the number of collections increases.

### Browsing and Searching within Curator

Both browsing and searching can be used in the Resource Discovery interface in order to find and view digital objects. Screenshots illustrating Curator's Resource Discovery interface are shown in Appendix 3 & 4.

Searching within Curator can be executed by simply typing the required string into the search box. If the user wishes to specifically search for a match from one specific element then they can choose from a drop-down menu next to the search box. It is important to note that not all elements will be listed: on the assumption that the Collection Manager is configured using unqualified Dublin Core, then the choice of elements will be restricted to those available within that schema. The user is unaware from the search box which metadata schemas are used within the individual collections.

Elements in other schemas may (or may not) be mapped to elements within the Collection Manager's Unqualified Dublin Core Schema. Unless the user is aware which elements from a particular schema (such as EAD or MODS) have been matched to elements in the Collection Manager metadata schema, the user is may prefer to choose 'free text' as the option of choice from the search form in order to find all matches to their chosen criteria, regardless of the element and schema in which their search term is located.

Endeavor Curator also has the ability to limit searches to one collection or any combination of collections. Other digital repository software (DSpace, DigiTool etc) restricts the user to search either all the collections or just one specific collection.

Curator displays the results in a form which illustrates both the name of the collection and the underlying structure within Curator where the data sits. Detailing the underlying repository rather than just the specific collection was found to confuse the typical user undertaking searches.

### Structure and Deposit in Curator

A particular feature of the Endeavor Curator system was the concept of containers. A container record was a record that acted as an organisational entity that grouped object records by subject, media type or other logical categories. Containers could contain other container records and must be linked to a collection record to order to be made available for public display in the web client. In effect a container acts a 'holder' for other entities and associated with that container is specific metadata.

The metadata associated with a container could either follow the metadata schema of the underlying objects and containers within that container or else it could contain data from a metadata schema that was different. Thus while the underlying digital objects in a container may contain material in the MODS metadata Schema, the metadata for the container may be in a

## MIDESS: WP 8 – Repository Implementation and Population

different schema such as Dublin Core. A container record (with the description of its contents) is shown in figure 6; in this case, the container record uses Dublin Core.

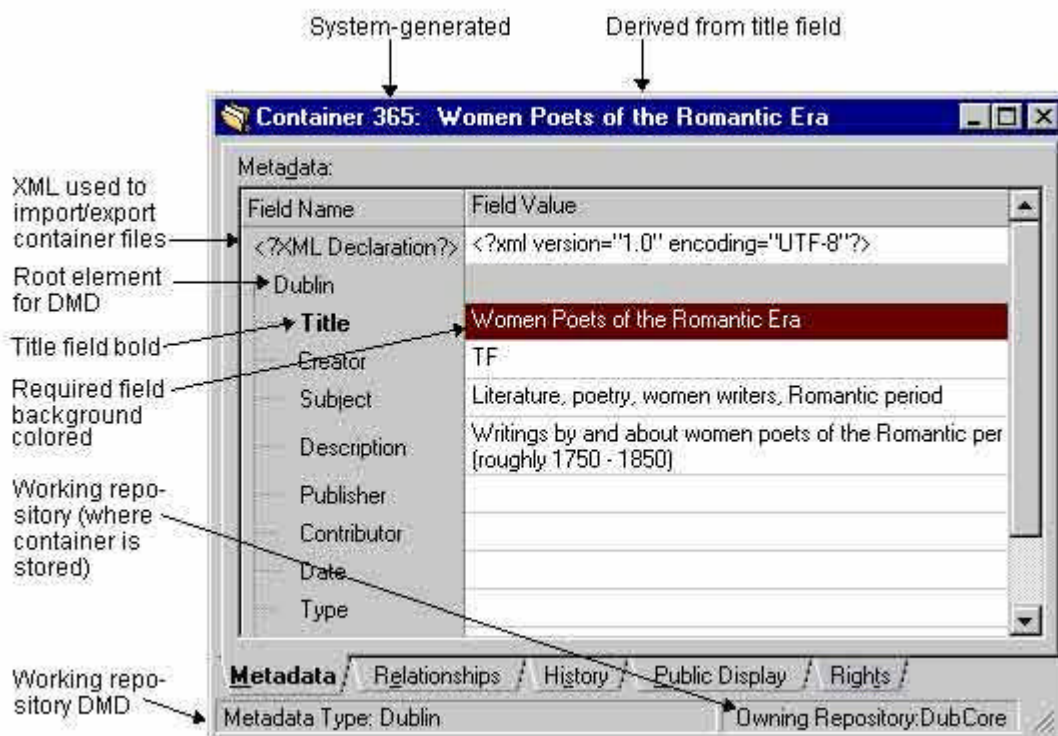


Fig.6. A container record in Curator

Depositors usually enter information into Curator via the staff client rather than through the web interface (although both are possible). The staff client is an application that is installed on each computer on which material is to be uploaded and added to the repository. Each collection manager will likely consist of a number of repositories which are mapped to a specific collection manager. An example of these separate repositories is displayed in Appendix 5. Objects in these repositories could be added to multiple collections if required. Thus an object in the Demo Repository may be added to both the Media Services Collection and the Medical Slides Collection.

Thus in effect the collections act as 'virtual' holders for the objects actually stored in the repository. Once a repository has been selected then the contents of the repository can be seen. Files from the repository can then be added to the collections (by "copying" files from the right-hand side of the page to the left). A single file in a repository can be added to more than one collection in this way, and thus can appear in multiple collections. Digital material from the repository on the right of the page should be added to collections with the same metadata schemas on the left of the page.

Using a similar approach, entitlement rights can be determined for individual collections so that only specific users of the digital repository system can see particular collections.

### Experience of using Curator

The Curator digital repository system had both significant strengths and weaknesses. Its strengths lay in its ease of deposit of material, the logical way in which the system was organised for system administrators and staff managing the system, and its ease of use for both staff and for resource discovery users. The support provided by the company Endeavor was also excellent.

## MIDESS: WP 8 – Repository Implementation and Population

However a particular weakness of Curator was the complexity involved in customising the resource discovery interface, which effectively required programming skills. For example, the entire resource discovery module was developed in XML rather than HTML and the lack of detailed documentation associated with the system made it difficult to tailor the system to the requirements of the University of Leeds. The complexity involved in linking the XML structured interface to the system also implied an expert knowledge of the product and an expert knowledge of XML if the public interface was to be modified or customised.

Thus while the back-end administrative system of Endeavor Curator was well designed and easy to use, the user interface and the linkages to the objects in the repository from the interface was difficult to develop and with poor documentation. The project officer employed to support the repository did have basic XML skills but the level of skills required to tailor the repository to the requirements of the University of Leeds exceeded his knowledge and the MIDESS project had to rely upon Endeavor programmers and support staff (working in the USA) to make all but the most minor of changes to the XML. The documentation explaining how XML was used within Curator to link between the Resource Discovery interface and the underlying system was also poor.

The version of Curator initially installed at the University of Leeds was Curator 3.6. Given that the University of Leeds was operating a pilot digital repository system rather than a full digital repository service, the University of Leeds volunteered to become one of the beta testers for the next release of Curator – version 4. As a beta tester for version 4, there was enhanced support provided by Endeavor technical staff in response to bugs found in the software by the Leeds staff. Moreover, since the University of Leeds was at this stage only developing a pilot rather than a full service, migrating to version 4 at an early stage provided an opportunity to exploit new features of functionality which were pertinent to the needs of MIDESS and, more generally, of Leeds. These included:

- Better support for metadata standards such as METS (important for MIDESS)
- A new Resource Discovery interface which would include support for integrated browsing by thumbnails (important for rolling out to actual users)

Version 4 of Endeavor Curator was installed on the MIDESS server in the Autumn of 2006 but the initial version had several significant bugs present which greatly impeded progress. Project staff spent a lot of time investigating and reporting features which malfunctioned, and it was sometimes difficult to determine whether the failure of a particular operation (e.g. ingesting a METS file) was due to a problem in the data or a bug in the system. Since Version 4 of Curator was very much a beta product, Leeds was not in a position to conduct a full usability testing procedure.

However a simple System Documentation manual was produced as a technical guide to support staff working on the system and this is available at:

[http://www.leeds.ac.uk/library/midess/MIDESS\\_Guide\\_Leeds.pdf](http://www.leeds.ac.uk/library/midess/MIDESS_Guide_Leeds.pdf)

### Migration to DigiTool

In December 2006 – just as Curator version 4 was looking to come out of beta – it was announced that a third-party finance company (Francisco Partners) was buying both Endeavor and ExLibris and merging the two companies. Both companies had their own digital repository products, the ExLibris product being DigiTool. The new company then decided that only one digital repository product would be supported long term and that this product would be DigiTool rather than Curator. Support would continue to be available for Curator 3.6 but all development work on Curator 4.0 would stop and beta-testers (such as Leeds) would be migrated to DigiTool.

Leeds was given a migration path from Curator to DigiTool and all digital contents that had previously been held in the Curator digital repository was migrated by ExLibris staff to the DigiTool platform. It was also agreed that the DigiTool interface would be modified to Leeds requirements where this was possible within the constraints of the system functionality. Unfortunately, contractual negotiations delayed implementation and training in the new system,

## MIDESS: WP 8 – Repository Implementation and Population

which was only commissioned in late May 2007, with training finally delivered in early July (postponed from June because of travel disruption due to floods). Even after this, Leeds staff had to familiarise themselves with the new environment, spend time on tailoring the parameters and workflows and cope with various idiosyncrasies and bugs. An upgrade to release 3.5 was also scheduled: this was critical to the Leeds implementation as it introduced support for MODS for the first time and also enhanced support for METS. Effectively the migration to DigiTool lost over 6 months to the development work at Leeds and delayed the planned transition from pilot project to live service. It did prove possible within this period to test some functionality around METS and OAI-PMH and so contribute to a better understanding of the issues involved in implementing services based on these standards; however even here, progress was limited by the fact that by this time project officers at LSE and Birmingham had completed their contracts and left.

### Structure of data and collections within DigiTool

Whereas within Curator there was the concept of Containers holding objects, DigiTool uses a simple hierarchy of “collections”. A collection type defines the collection’s characteristics. There are three collection types in DigiTool: Node Itemised and Logical.

- A Node collection defines a descriptive collection level in a hierarchy and contains one or more sub-collections linked to it – not objects.
- An Itemised collection is at the lowest level of any hierarchical branch, where relevant digital objects records are chosen for itemised inclusion in the collection.
- A logical collection is also at the lowest level of any hierarchical branch. This is a dynamic collection defined by a search term (or terms). A relevant search query is linked to this collection to enable automatic generation of the collection’s contents.

To be visible in the Resource Discovery Interface, collections must be “published” i.e. marked as available. Without this, a collection is not available for public searching.

The creation of a collection is shown in appendix 7. Unlike Curator, a collection does not have a specific metadata schema associated with it. Rather a thumbnail can be linked to it (if required) along with a short and long description for display in the Resource Discovery Interface.

DigiTool supports a number of schemas – MARCXML, Dublin Core, and Qualified Dublin Core, METS and EAD, with release 3.5 also introducing support for MODS, VRA and the digital preservation standard PREMIS. Some problems were caused by the fact that DigiTool 3.5 only supports MODS 2.0, whereas Curator supported MODS 3.1. Plans were already well advanced for loading metadata for the medieval manuscripts collection using MODS 3.1 and this had to be backwards mapped onto 2.0, with a consequent loss of some precision in the metadata structure.

### Browsing and Searching in DigiTool

The DigiTool system is currently available at <http://midess1.leeds.ac.uk:8881/> (02/08/2007)  
Examples of the Resource discovery interface for DigiTool are shown in Appendix 6

Leeds chose to keep the repository home page as simple as possible and it differs from the default DigiTool homepage both in layout and by the inclusion of “thumbnail” images. This page was specific tailored for the University of Leeds and based on the closest match possible to the previous implementation in Curator. However this does now mean that every time a collection is added, the HTML of the repository homepage screen must be modified in order to link the thumbnail on the homepage to the collection. At various presentations on campus, both this home page and, more generally, the simplicity / ease of use of the resource discovery module have elicited favourable comment.

## MIDESS: WP 8 – Repository Implementation and Population

The Leeds DigiTool homepage presents the option to either browse a collection or to perform a simple word search. For searching, the choice is between searching one collection and searching all the collections; there is no option to select a combinations of collections to be searched (in contrast to Curator). Simple icons on the left of the page are used to represent the collections. There is also the facility to provide a paragraph of information about each collection alongside the thumbnail image and link. Where required this can include links to accompanying or related websites. For example in the case of the Physics Instruments collection, the paragraph of text detailing the physics instruments collection would link to the page on the Library's website which gives details of the Bragg notebook.

The advanced DigiTool search screen allows the user to search collections within very precise parameters. Individual metadata elements (Title, Creator, Subject, Metadata, Full text), specific media types (Text, Image, Audio, Video) and specific file formats can all be specified and also combined using Boolean operators. Searches can also be carried out against specific tags for XML objects encoded in METS and EAD, allowing an extremely detailed approach where required.

Unlike many other digital repositories (both open source and commercial) DigiTool also has the ability to browse by thumbnail rather than solely by a simple text description. The Leeds instance of DigiTool has been designed to keep the thumbnail browsing as simple as possible with minimal metadata on the thumbnail browsing screens. At the thumbnail browse level the only two metadata fields (elements) shown are the Title and the Subject. Clicking on the title enables the viewer to see the full metadata record.

At the thumbnail level the user has two primary choices for each specific object. They can either click on the title of the individual object and see more detailed metadata or they can click on the thumbnail of the object to view/use the object itself.

It is possible to offer a number of different instances of the object (where they exist e.g. different resolutions, formats, etc.) and these are shown by an icon to the right of the thumbnail. On ingest of high resolution TIFF graphic files - unlike in Curator - image thumbnails are automatically created (generated on the fly), along with a JPEG file. For most image files, if the user clicks on the thumbnail then the DigiTool image viewer is invoked and the image is displayed. This viewer includes options to manipulate the image in various simple ways, including changing the magnification, rotating the image, and selecting options to display at actual size or with best fit to the page.

The user also has the ability to log in to the system with a personal account and this provides a personal working space in which they can "collect" objects and annotate them. These personal collections can be organised into folders and – given suitable authorisation – the user can choose to make individual folders public. This could be particularly useful for academic staff members and several have already commented on this feature. At time of writing, the facility for login by University members via LDAP or Shibboleth has not yet been implemented due to time constraints. Implementation of this feature will also permit restriction of resources to specific users or groups of users – which is a key requirement for several of the collections.

### Experience of using DigiTool

Impressions of DigiTool so far are that the DigiTool repository has a wider range of features than Curator. It seems to lack the easy to use back-end administration properties that were present in Curator. However against this needs to be weighed the presence of very sophisticated workflow control features associated with web deposit. This potentially opens the way to allowing individual academics to enter material into the repository and controlling, for each individual, how the ingest process operates, which options regarding metadata format and rights permissions are offered and how closely the ingest is checked and approved. In this way, a trusted individual may be permitted to have access to several separate workflow options which may result in material being

## MIDESS: WP 8 – Repository Implementation and Population

fully ingested and made public, whereas another may only have access to one workflow, which requires checking before the material is fully ingested.

Limitations in DigiTool encountered at Leeds include the following:

- The DigiTool system can automatically create both a low resolution JPEG file and a thumbnail on ingest into the repository. However there is no flexibility as to the size of the JPEG file or thumbnail created. Thus if a large TIFF file is ingested, this produces a large JPEG file which results in only part of the file being first seen when the JPEG image is viewed in the repository. In order to view the whole JPEG file the magnification needs to be decreased each time one of the JPEG files is viewed, so that the entire image can be seen rather than just a small part of the image.
- There is little flexibility as to how the thumbnails within a collection are displayed on the page. Only two thumbnails can be displayed horizontally at one time. Where a collection is very large this may require a great deal of scrolling or page turning before the entire collection can be viewed by thumbnail.
- Currently, video stored in the repository is physically stored on a streamed server rather than in the repository. Thus the streamed server functionality is not actually contained within the repository and only the associated metadata files are stored in the repository. Ideally we would like be able to stream the files directly from the DigiTool software program rather than simply linking to these files on the streamed server.

Some problems have also been experienced arising from the way in which material is made available to the public resource discovery module. At present, when material is added to the repository, it does not become visible in resource discovery until it has been “harvested” by a process which normally runs at timed intervals (the default is every 2 hours but this can be modified by the local site), although the process can also be invoked at any time via the staff client. This delay in public availability makes it more difficult for staff to load and view test data e.g. in preparation for a large ingest. It also impacts on the end-user deposit workflow – which otherwise is a strong point of the system – in that the depositor is unable to see immediately whether the deposit has been successful (presuming, of course, that the depositor in question has authorisation to fully complete the submission of new material and that their deposits are not subject to manual approval before being released to public view via resource discovery).

### Overall software performance

Both Curator and DigiTool are complex digital repository software products. Both products have (or, in the case of Curator, had) the potential to provide fully functioning digital repositories and both repository systems preformed largely in line with expectations at Leeds. However both repository systems require trained technical staff to support the repositories and neither can be considered an ‘out of the box solution’.

The experience at Leeds suggests the following comparison between the two systems in terms of their overall strengths and weaknesses:

## MIDESS: WP 8 – Repository Implementation and Population

	Endeavor Curator	ExLibris DigiTool
<b>Strengths</b>	<ul style="list-style-type: none"> <li>• Good back end Interface for deposit, authentication, control of collections, etc</li> <li>• Support for Integration with VLE's (WebCT)</li> <li>• Good browse facility</li> </ul>	<ul style="list-style-type: none"> <li>• Greater overall functionality than Curator</li> <li>• Wider range of Metadata schemas supported</li> <li>• Workflow control included.</li> <li>• JPEG2000 supported.</li> <li>• Resource discovery interface written in HTML ensuring Interface can be tailored more effectively.</li> <li>• Resource discovery interface simple to understand</li> <li>• System can be extensively tailored.</li> </ul>
<b>Weaknesses</b>	<ul style="list-style-type: none"> <li>• Complex to create metadata schemas</li> <li>• Resource discovery interface was poor (no thumbnails)</li> <li>• Resource discovery relied on XML, with complex integration to the underlying system</li> <li>• Web deposit was poor</li> <li>• No ability to automatically create an thumbnail and JPEG file on Ingest of Image files.</li> <li>• Lack of good METS support (3.5 and 4.0)</li> <li>• Lacks the functionality of DigiTool</li> </ul>	<ul style="list-style-type: none"> <li>• No Integration with VLE's</li> <li>• Deposit of digital material in staff interface complex when compared with Curator</li> <li>• 'Has a harvest repository into Silo' function which prevents users directly entering material into resource discovery.</li> <li>• No browse facility (all browses are in fact searches based upon common criteria)</li> <li>• Inability to specify resolutions and size thumbnail and JPEG file on ingest of TIFF image file.</li> <li>• Detailed training essential to understand the system</li> <li>• Requirement to build bespoke metadata schemas in raw XML</li> </ul>

### Staffing Requirements and Training Requirements

The migration from Curator to DigiTool coincided with a changeover in project officers at Leeds, which makes comparison between the 2 systems a little more difficult. The DigiTool repository seems to require more system administrative functions to be performed at the server level than was required for Curator. However the public interface in DigiTool is in HTML rather than XML which makes it much easier to tailor to local requirements without specialist knowledge. Overall, training in the use of DigiTool seems to be absolutely essential – much more so than for Curator.

For both Curator and DigiTool, a reasonably high level of expertise is required in server based systems (Apache, backups, maintenance of servers etc.), as well as a detailed knowledge of UNIX command line (both Curator and DigiTool repository at the University of Leeds ran on a UNIX Sun server). Basic Oracle/relational database knowledge was useful for both systems and an appreciation of JBOSS is useful for DigiTool. As with all repositories, it is also useful to have a good understanding of metadata schemas as well as a basic understanding of protocols such as OAI-PMH and Z39.50.

## **MIDESS: WP 8 – Repository Implementation and Population**

During the majority of the lifespan of the MIDESS project the University of Leeds was able to employ two project officers, one specialising in analysing and developing the metadata associated with collections while the other concentrated on the technical issues associated with the repository. Even so, it was clearly unrealistic to expect a detailed understanding of all aspects of the role upon first joining the MIDESS project. Thus training the project officers to the necessary level of expertise took several months.

Metadata expertise emerged quite clearly as a key requirement for successful implementation of a repository. Leeds had access to staff with some prior experience of metadata, but there was nonetheless a steep learning curve in terms of familiarity with Dublin Core and MODS, the two schemas required for the main collections being ingested. Like many university libraries, the Library at Leeds has a Metadata Team – but one whose role has until now focused on MARC cataloguing of traditional library materials. The project proved a useful stimulus for developing skills within this team and building expertise in the use of the most commonly used XML schemas.

Training requirements for public users of the Resource Discovery interface were minimal and its use was broadly self-explanatory. Deposit of material into the repository could take place either via a web interface or via software directly installed on the PC. Training took a maximum of 1 hour.

### **Future of the Digital Repository at the University of Leeds.**

Overall, Leeds has gained greatly from the opportunity which MIDESS provided to explore the ways in which a multimedia repository can be used to support learning, teaching and research. Very positive relationships with academics have resulted from the discussions, and both sides have clarified their requirements and expectations. There are manifestly a wide variety of potential collections that could be stored in the repository and the current contents are recognised as only representing a pilot collection of materials from across the University. The general feedback from staff across the university who have seen the repository has been very positive, with a number of people expressing interest in both using the repository and depositing material into it.

However many staff expressing an interest in the repository also wish to adopt a 'wait and see' attitude to see to what extent it will be funded and supported long term by the University. These staff are therefore reluctant to provide material or develop the accompanying metadata while the repository is still considered to be in a pilot stage with no long term funding available from the university. The work required to gather material, digitise material (where necessary), and add suitable metadata to each of the datastreams is substantial for the depositor. In particular, provision of adequate metadata is an issue since much material has little or none pre-existing, yet the material will remain invisible in the repository without this.

During the final year of the project, the Library submitted a bid for ongoing staffing to take forward the work begun through the MIDESS Project (the hardware/software platform is secure until 2009). A case was made for developing a comprehensive repository service which would support learning, teaching and research, fully embedded within the University's core processes. Interim funding for one year has been identified and it is hoped to obtain a positive decision for ongoing support beyond that.

### General Conclusions

The implementation phase of MIDESS resulted in 3 functional repository platforms, each containing a mix of material to test how different types of multimedia content can be ingested, stored and made available. In working with the metadata for the various collections, each of the partners came to understand the importance of being able to draw on metadata expertise in handling multimedia materials and also gained significant experience of the practical issues which can arise.

It was very disappointing that none of the partner sites was able to progress the work to the point where testing with live users became possible, each for very different reasons:

- Birmingham had always viewed the work as a pilot which would allow them to explore the issues involved in handling multimedia materials. The project identified very real limitations in DSpace and led them to consider Eprints v.3 as a platform for the next stage of repository development
- LSE succeeded in ingesting a substantial amount of material into Fedora, but the difficulties encountered in installing a web-based front-end application for public resource discovery was a major impediment to developing a full service. In consequence, they are reassessing their options, and in particular the functionality which may be available through Eprints v.3.
- Developments at Leeds were unexpectedly delayed for approximately 6 months by the merger of Endeavour and Ex-Libris, which obliged them to abandon their Curator installation and switch to DigiTool 18 months into the project. However, their intention is still to launch a full repository service based on the DigiTool platform in the session 2007/2008.

Nonetheless, the response of the academic partners to the possibilities opened up by using a repository was extremely positive, the scenarios explored within MIDESS gained a significant degree of acceptance within the user community and it would have been very beneficial to have moved forward to test working models for the delivery of digital objects within and across institutions

Several key themes and lessons emerge from the combined experience:

- For open-source products, another site may have succeeded in implementing a particular feature, but this does not imply that it will install easily as part of your own repository. This can result from differences in version, in hardware platform, or simply in having local expertise to tweak the software module.
- Even with a commercial product, bugs and other unpredictable behaviour are likely to emerge because most products are currently fairly new to the marketplace and complex in what they are trying to deliver. Comparison can usefully be made to the experience of Library Management Systems during the previous decade.
- Particularly for multimedia and/or unusual collections, it is unwise to underestimate the staff effort involved in the range of processes required to achieve a successful ingest. Not only staff time is required, but also the particular expertise necessary to resolve the issues concerned – and these can be very varied.
- Following on from this, nearly all repository implementations require technical expertise to be available locally if a full range of services is to be delivered.
- Metadata expertise is crucial for successful implementations, and staff with the relevant expertise may not always be readily available locally.
- The fact that a product supports a particular protocol does not necessarily mean that implementation will be easy. Version numbers and interpretations may differ, impeding integration and interoperability.

## **MIDESS: WP 8 – Repository Implementation and Population**

Finally, the project's experience of using Fedora and DSpace does open up a number of questions about whether open-source repository products are suited to the complex requirements of multimedia objects, at least in the typical HE institution. It could be useful for a standard implementation to be developed for these products – and perhaps for others as well - which would address some of the deficiencies in functionality which MIDESS encountered in the default installations, and thereby make them a less costly option in terms of staff time and expertise.