Technology Watch Report

Institutional Repositories in the context of Digital Preservation

Paul Wheatley
University of Leeds

DPC Technology Watch Series Report 04-02
March 2004

1.0 Scope

This report will focus on the requirements, functions and use of digital preservation in an institutional repository context. It will also provide an overview of existing institutional repository software as well as details of working systems and their core aims and purpose. Software that could be installed and used to perform the role of an institutional repository will be covered by this report.

A number of existing publications provide comparisons between existing institutional repository software. A specification document for the DARE Project contains a comparative discussion of DSpace, ARNO and NCP [1]. The Open Society Institute has, at the time of writing this report, released a systematic comparison of repository software entitled "A Guide to Institutional Repository Software" [2]. This OSI guide provides a detailed checklist of features present in open source institutional repository software. These documents do not discuss digital preservation in any detail (if at all). A publication by the DAEDALUS Project [3] describes the experiences at Glasgow University in implementing, running, configuring and building on both the ePrints software and the DSpace software.

Rather than duplicating this work, this DPC report will concentrate on issues of crucial importance in achieving long term digital preservation in the context of institutional repositories. The reports described above only briefly touch on digital preservation related features.

2.0 Institutional repository - definition

Clifford Lynch defines his view of an institutional repository as "…a university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution." [4]. The crucial word Lynch uses here, is “community”. The term institutional repository implies a community based service although this is interpreted by repository developers in different ways. Some embody a cross-subject, cross-department service which requires flexibility to meet the requirements of many different types of users. Some focus more specifically on a particular subject and possibly type of material to be archived, while still delivering an institution-wide service.

Although, as Lynch implies, the term institutional repository suggests a higher education context, this is not always the case. Institutions with requirements to store, preserve and provide access to digital materials may require their own repositories, and this is likely to become more common over time.

In the context of this report, the term institutional repository will be used simply to refer to an actual instance of an institutional repository or the software that enables an institutional repository.

3.0 Technology Watch

The development of institutional repositories is very much in its infancy. The specific requirements for repositories are still being defined and existing software is only just beginning to attempt to fulfill these needs. With respect to the application of digital preservation in institutional repositories, current implementations are at an early stage, but understanding of the problem and the solutions to address it are advancing rapidly. This makes the subject of Institutional Repositories in the context of digital preservation an important area for this current series of DPC Technology Watch reports.

The subject has relevance to the DPC membership with regard to:

- Understanding the key issues relating to digital preservation and longevity
- Where the contribution of requirements, expertise and best practice can be made to institutional repository development
- Awareness of purpose and suitability of current products and technologies

Readers should also be aware of a related DPC report “The Open Archival Information System Reference Model: Introductory Guide” [5].

4.0 Repository purpose of use

Institutional repository software has been developed with a range of aims and purposes in mind, and there is not a consensus across the available products of what those roles are. The contention between improving access and dissemination now, versus protecting that access for posterity (preservation) is discussed in detail by Pinfield and James [6]. In practical terms there is an apparent difference in focus of institutional repository implementations, where some prioritize digital preservation as an aim and others do not (for example DSpace and ePrints respectively). Experiences of the DAEDALUS Project [3] suggest that repository software like DSpace and ePrints are not really in direct competition. In fact, these differing implementations have quite distinct aims but similar attributes.

Differences can also be found in repository software with regard to types of materials the software is primarily designed to hold. Some repository software concentrates (primarily if not exclusively) on document type materials and others (perhaps more embodying the true sense of the term institutional repository) are designed from the outset to attempt to realistically meet the requirements of holding any type of digital object.

The overall purpose and aim of an institutional repository will have a significant bearing on the selection policy and practice for that repository. Materials may be selected for

archiving by those responsible for administering the central repository, by devolved community based management or even by individual users (or creators of digital materials). Institutional repositories may not even have an overall selection policy, but instead simply provide the mechanisms and flexibility for the development of community or department level policy.

The Cedars project suggested that issues of digital preservation (eg. data format, dependencies, etc) should not affect initial selection decisions [7], but repository implementations that limit the range of acceptable submission formats are common. It remains to be seen whether digital preservation understanding and support will advance to a sufficient level that choices in selection and submission really can be format independent.

5.0 Current best practice?

A consensus in best practice in performing digital preservation is yet to be reached. While recent research and testing is now getting closer to practical implementation and real world usage, the definition of best practice is some distance away. OAIS (Open Archival Information System [8] provides a useful framework of terminology and understanding, describing the key elements of a digital repository design. The RLG/OCLC work on Trusted Digital Repositories [9] builds on the OAIS structures, detailing the requirements for effective repository based digital preservation.

For now, the foundations provided by OAIS, the RLG/OCLC work and the practical experiences of those who have developed digital repositories and preservation systems must guide archival and preservation work where possible. Some of these key developments are described in more detail below.

6.0 Requirements and aims for effective digital preservation

6.1 Understanding the problem

Where the long term accessibility of archived objects is a key aim of a particular digital repository, a number of requirements must be met in order to effectively meet this goal. In order to understand the process necessary to achieve the long term digital preservation of objects placed within a repository, it is first useful to break down what we understand as effective preservation itself. The key functional goals can be summarised as follows:

1. Data can be maintained in the repository without being damaged, lost or maliciously altered.
2. Data can be found, extracted from the archive and served to a user.
3. Data can be interpreted and understood by the user.
4. Goals 1, 2 and 3 can be achieved in the long term.

Goal 1 is easily recognized as a key requirement for any form of digital repository and has been well fulfilled by fundamental techniques of computer science for many years. In recent times, developments in areas such as security, authenticity, verification and storage have made significant advances and it is clear that current repository software and actual repository implementations address this goal well. As technology advances, procedure, software, and technology will also need to progress.

Complications arise when digital objects are tied to the media upon which they are stored. But as Holdsworth points out, “The media is not the message” [10]. By mapping representations of digital objects to simple bytestreams, while ensuring that all the required significant properties are retained, media independent preservation can be achieved.

The second goal has received much attention in the last few years as the "information revolution" has sought to improve access, finding and integration. The process of Access and Dissemination (in OAIS terms) integrates these wider search and retrieval issues with the user authentication and the extraction and delivery of the data to the user. Repositories will need to support a number of methods of searching and retrieval dependent upon their subject and user base. These methods are likely to change over time as standards and widely accepted techniques go in and out of fashion. For example, the Open Archives Initiative Protocol for Metadata Harvesting is currently an essential access route to support and is seeing wide uptake across a number of communities.

An issue which remains unclear is the underlying requirement for the unique and persistent identification of each digital object in a repository. There must be a system for resolving these identifiers and directing queries to the physical location of stored digital objects. In most cases this will be the underlying technology on which searching and finding aids depend. RLG/OCLC noted in 2001 [9] that a standard which supported the requirements of digital preservation was yet to emerge and recommended that efforts be made in this area. Since this report was produced, standardisation in this area has still not been achieved. Unique identifiers will be addressed in more detail below.

The third goal can be seen as the core digital preservation aspect of the repository question. Although the first two goals must certainly be achieved in order to provide real long term digital preservation, the third goal ensures the work involved in the first two is still useful in the *long term*. Ingest, archiving and then providing access to a digital object in the space of for example, 2 years, is unlikely to tax the success of meeting of goal 3. Add another 10 years or more between ingest and access and a user will struggle to make sense of the digital object they have been provided with as technology change makes the original hardware and software obsolete.

OAIS loosely terms the process required to achieve this goal as "Preservation planning", but rather than effectively forming a separate function within the archival model, it plays an important integrated role within most of the key archival processes which OAIS also describes (eg. ingest, administration, dissemination). This integration aspect must be

recognised and understood. Repositories must provide the flexibility to allow digital preservation functions to be incorporated as they are developed.

This is the least understood of these 4 goals, and continues to see a trickle of research into techniques and consequently best practice.

The fourth and final goal suggests that some thought and effort needs to be given to ensuring that the first 3 goals can still be achieved successfully in the future. This implies a degree of continuity which ideally should be attained without undue expenditure. While goal 3 implies consideration of the long term perspective, the first two goals do not and thought must be given to sustaining them over time. This is particularly important in an institutional repository context.

6.2 Preservation processes

Ultimately, achieving the third goal requires a process that adds preserves, interprets and adds meaning to accessed data. However, as suggested above this is not a stand alone process. It can only be achieved with input throughout what Beagrie and Jones term the "lifecycle" of a digital resource [11]. In repository terms this requires specific preservation type processes from ingest, through to storage, administration and finally access. These functions must capture, store and enable use of various types of metadata. In particular, Representation Information, which describes how to gain access to the intellectual content encoded within a digital object (see glossary).

A summary of the functions and infrastructure required might include:

1. A process of ingest that creates or extracts the metadata necessary to ensure preservation.
2. A framework within which the required Representation Information can be stored, managed and utilized (a Representation System).
3. A process of "technology watch" which monitors technology dependencies and the recorded Representation Information, and takes action to ensure continued preservation where technology obsolescence occurs.
4. A process of rendering (displaying or making sense of) retrieved digital objects.
5. A process and related framework for recording change metadata

Beginning with the ingest of a digital object to an archive, the ingest process must capture an appropriate amount of Representation Information. This metadata must be stored in an appropriate way to facilitate both its maintenance (a process of keeping it current) via a preservation watch function and its use in a representation and rendering capacity (see below). The first of these functions will monitor the representation information and technology it depends upon, to ensure it is still current. The second of these functions will provide a user with the appropriate information to render the digital object, perhaps starting a further distinct rendering process (for example Migration on Request [12] or Emulation [13]). The rendering process itself may require additional thought and resources if the repository in question is responsible for maintaining a relevant rendering

method. This could mean maintaining a current tool or replacing it with a new one. If format migration is chosen as the preservation strategy, rather than changing or updating a tool when technology obsolescence occurs, a migration from format to format of all objects of that type in the repository may have to be made. A related process is to record change metadata describing any changes made to objects in the repository.

As will be discussed later in this report, not all of these functions need to be undertaken by a specific repository, but at the very least support will have to be provided for the integration with external services. In the case of ingest in particular, this is not trivial. Consequently these issues need to be considered in the design of an institutional repository.

As well as enabling these functions to provide for long term preservation, the repository design must also ensure that the repository itself can survive in the long term, again another important design consideration.

7.0 Preservation functions

The report will now examine the design considerations for institutional repositories that will provide for long term digital preservation.

7.1 Unique, persistent identification

Purpose - mechanism for managing digital objects, and for finding them even when their physical location changes.
Method - name allocation, name resolution

As mentioned above, institutional repository developers have recognised the need for unique identification of archived objects and incorporated them into their designs but a range of identifier standards have been used. As is often the case with standards there are several to choose from! Requirements for the design of the identifiers and related systems will vary from institution to institution but there is consensus on the obvious benefits of interoperation in finding aids which a single all encompassing standard would provide. While the design of the identifiers itself can be made sufficiently simple and flexible to satisfy most requirements, agreement on the resolving services may be harder to reach. Leadership from a consortium of institutions (perhaps the Digital Preservation Coalition [14] or Digital Curation Centre [15] in the UK) may be able to champion an appropriate standard and make progress in this area. Repositories may be able to integrate their existing identifiers with a flexible but perhaps different leading standard that emerges at a later date. Prefixing existing local identifiers with an institutional label may facilitate this.

The various standards for unique identification (eg. URN, DOI, ARK, etc) are discussed in detail on the PADI website [16].

7.2 Ingest

Purpose - An ingest process fulfils a range of aims, but this report will concentrate on the capture of Representation Information during ingest
Method - Modular tools for identification and verification of file formats and also for the automated extraction of metadata

Ingest processes which aim to capture metadata are recognised as a crucial area to develop and automate to reduce this potentially high effort, high cost function. For most repositories, it is unrealistic to gather and or extract sufficient metadata to enable preservation. A Range of institutional repository ingest functions will need to be developed. These include:

- Automated extraction of metadata
- Automatic identification of file formats
- Verification of an objects compliance to a relevant file format specification

The highly specialised nature of these functions will necessitate modular solutions which can be plugged into institutional repository systems as required. Integration with other key preservation systems such as those that address the storage and use of Representation Information will be crucial. Research, development and evaluation work in these areas is currently being undertaken by MIT, the University of Pennsylvania and the UK National Archives.

7.3 Representation Systems

Purpose - Representation Systems provide a mechanism for storing and utilising Representation Information. Representation Information provides the knowledge of how to gain access to the intellectual content encoded in the digital object (see glossary).
Method - There are two emerging and related methods for addressing semantic Representation Information - file format registries and OAIS Representation Networks. There are a range of techniques for addressing structural Representation Information, from the use of existing technologies to newly defined standards like METS [17].

While much importance has been placed on the definition of preservation metadata schemas, the hardest part of the metadata problem is certainly the technical metadata or Representation Information (see also Recording Change Metadata below).

Examination of existing preservation metadata schemas suggests that the non-Representation Information component of the problem is well understood. Achieving standardisation for the non-Representation Information component of preservation metadata will improve and simplify interoperability and searching. This is now primarily

an organizational problem and will depend on co-operation and agreement between the major players in the field, who are able to develop and promote an appropriate standard.

In the case of Representation Information, a set of basic fields within a metadata schema will not be sufficient. Where this has been attempted (for example in the recent NLNZ Preservation Metadata Schema [18]) the Representation Information fields have been defined very weakly. Even where technical dependencies can be listed in fields of this type (eg, format, rendering application, system the application runs in, etc) the task of maintaining and updating this information over time as it becomes obsolete is colossal! Changes will have to be applied to the metadata of every applicable object in a repository. Moving this to a referenced external system where only one entry (representing possibly thousands of objects) has clear advantages.

A range of approaches have been suggested for describing structural Representation Information. The Cedars Project [19] took a pragmatic view, and utilised existing technologies to describe simple file structures and relationships. The use of the TAR file structure and associated tools for unpacking TAR to a usable file system (details of which were described in a Representation Network (see below)), provided an effective way to address objects composed of multiple files. The more recent development of the METS standard [17] shows promise for describing more detailed structural Representation Information and is being explored in detail by various institutions and communities, including DSpace.

There is a growing consensus that an external system or repository of referenced Representation Information will provide a more manageable and effective solution than raw repository based metadata fields when dealing with semantic Representation Information. For the purposes of this report these repositories of semantic Representation Information have been termed "Representation Systems". These systems store the technical metadata independently from the digital objects in a repository, allowing several objects of the same format to point to the same single piece of metadata. Monitoring and updating the metadata then becomes a much simpler and more manageable task.

Representation Systems will play a crucial role in achieving long term digital preservation and data curation. So far little work has been devoted to their development and only one system is currently in operational use, PRONOM at the UK National Archives [20]. Both PRONOM and the proposed Global Digital Format Registry [21] broadly follow a "file format registry" approach, which is based around a simple database of file formats. A defined categorisation of file formats (ideally far more specific than for example, MIME) is used to structure the recorded Representation Information. In its most simple form, the file format registry will be held in a database at a single location. It is likely that access would be provided to remote sites sharing the registry via the internet (this facility will be present in the next version of PRONOM). The simplicity of this approach is its strong point, but it is unclear if sufficient format detail can be maintained without creating lengthy and unusable categories of file formats (what is classed as a format?). The next release of PRONOM and expected external take up will provide an

invaluable practical evaluation of these issues. In particular whether the PRONOM format categories have reached a suitable compromise between practical use and detail specifics.

The second approach was suggested in the OAIS framework as a network of technical metadata, termed a Representation Network. Much of the information about a particular file format will form some overlap with other file formats. So the Representation Network approach utilises a network of nodes of metadata and a system for pointing from one node to another. For example, nodes for the Word format and the PDF format will both point to nodes describing ASCII and UNICODE.

Initially a Representation Network for a current file format may be relatively simple. For example, a node describing the Word format will have links to documentation of the file structure, a link to a node describing Microsoft (which in turn will point to the Microsoft website) and a link to a node describing a method of rendering; perhaps using Word viewer. In ten years time when Word viewer becomes obsolete, the representation network will be updated with new nodes to describe new methods of rendering, in this case perhaps migration or emulation tools. So Representation Networks will grow over time as technologies are monitored and action is taken to ensure access to data formats over time.

The CEDARS Project and subsequent Representation and Rendering Project developed this concept into a working demonstrator [19]. Cedars argued that by allocating unique identifiers to each node of Representation Information and providing a form of name resolution service, efficient networks of Representation Information could be constructed. Preserved digital objects reference appropriate Representation Information using the identifier and resolving system. A key advantage is the distributed nature of the Cedars Representation Network, where different institutions could provide Representation Information in their areas of format expertise. These would then be available to all other institutions using the same system. A community of digital preservationists could then be created who share, maintain and test each others Representation Information. The network structure facilitates efficient technology watch functions (see below) by highlighting the key dependencies where technology obsolescence will occur [22]. Representation Networks have not been developed to an operational level but were tested using functional demonstrators as part of the Cedars project.

Discussions on the Global Digital Format Registry suggest that even with a working format registry, some materials will be too unique to be adequately described by the Representation Information contained in the registry and so perhaps inevitably institutionally controlled Representation Networks may also be needed. The flexibility and distributed nature of the Representation Network approach could well prove important.

Populating Representation Systems is a major challenge in itself. Gaining access to accurate proprietary file format information can be very difficult, and even open file format documentation can be incomplete or inaccurate [23]. The scale of the problem will

certainly require co-operation and Representation Information sharing across the digital preservation community.

Whichever of these two approaches is widely adopted (and possibly both could work in tandem) Institutional Repository software must be designed with the flexibility to incorporate linking to Representation Systems as they become available. Given that most of the physical metadata is referenced, the main issues will reside in the integration of the ingest and dissemination processes of the repository software. Integration and interoperability with repositories will require open, flexible designs and some degree of standardisation.

7.4 Technology Watch

Purpose - A Technology Watch function monitors Representation Information and related rendering capabilities and provides alerts when the Representation Information is no longer current due to technology obsolescence.
Method - Unclear at the current time, but will likely involve a range of techniques from automated processes to manual surveys and evaluations.

Technology Watch is a frequent function that must be performed to ensure Representation Information is maintained in a current state. Representation Systems are likely to incorporate integrated Technology Watch functions and also rely on external Technology Watch operations like that of the DPC. As well as the primary role of maintaining Representation Information, technology watch must also be provided for the software and hardware on which repositories themselves depend (see Overall Repository Structure, below). Although the shape and form of adequate technology watch functions is yet to be fully understood it seems clear that as with Representation Systems, co-operation and community integration will be important, where sharing of results and expertise will be required.

7.5 Rendering

Purpose - To turn a bytestream into meaningful information or to gain access to the intellectual content encapsulated in the raw data.
Method - Many rendering strategies have been proposed, including migration and emulation.

Rendering will not be addressed in detail in this report as it has been discussed in detail elsewhere [24] [25], and the implications for the design and integration with repositories are effectively covered under Representation Systems and Recording Change Metadata.

7.6 Overall Repository Structure

Purpose - To ensure a repository survives technological change.
Method - Layered design and the choice of stable technologies in the construction of the repository.

Just as Representation Information will need to be monitored and updated as technology obsolescence impacts upon it, repositories themselves will change and develop over time. Change may occur to ensure continued function of the archive in the face of dependent technologies becoming obsolete, or alternatively may be instigated to provide new or different services at the point of use. If long term survival of a repository is required, it must therefore be accepted that significant parts of the repository will change. For example, in the case of the LEEDS long term file store [10], migrations across 3 different storage technologies were required in the space of just 12 years.

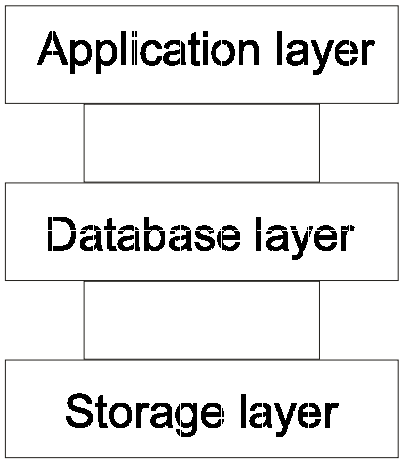

Figure 1 : repository abstractions

Figure 1 shows a simple break down of abstracted repository layers. By careful design of the interfaces between these layers, the providing technologies for the layers themselves (primarily the top and bottom layers) can be changed without major impact to the repository as a whole. As the technical, functional and user paradigms of modern computing change over time (and go in and out of favour) we have to accept that the applications which depend on them will also change. Clearly, the current front end implementations of the repositories described at the start of this report will not survive in their present form for five years, let alone one hundred. Choosing a sensible high level design can simplify this inevitable change and hopefully prevent any data loss in the process. Browsing through the many institutional repository implementations on the web reveals several with warning labels about adding content to systems that may close down without hope of migrating data to a new replacement system. The dangers of not addressing this issue are all too apparent.

7.7 Recording Change metadata

Purpose - To record changes made to digital objects in a repository in order to assist in answering questions of authenticity and to inform future maintenance and preservation actions.
Method - Requires a repository function to record and update metadata.

Digital objects in a repository and their respective metadata can require changes to be made to them for a number of reasons. Changes to a digital object may occur following maintenance or preservation action (eg. format migration), and new versions of a particular digital object may be created through redaction or revision. These changes must be recorded in the metadata record as a form of history or change metadata. Note that recording an audit trail to changes in the metadata as well as the digital object itself is a sensible course of action [26].

The process of recording change metadata will by necessity have to be quite an integrated repository function given the range of processes that may alter or update objects or metadata. Mechanisms for recording change history exist for many other purposes and are relatively straight forward, but a question remains as to the quantity and detail of change metadata required in order to fulfill the purpose.

8.0 Institutional repository software

This sections describes open source institutional repository software, listed in alphabetical order.

8.1 ARNO

Project home : http://www.uba.uva.nl/arno
Example installation : http://dare.uva.nl
Core purpose : document storage and dissemination

ARNO is a university archive server based repository system that has seen significant take up amongst Dutch HE institutions. With a strong focus on repository integration, ARNO's focus is on the stewardship and dissemination of scholarly publication and supports the OAI protocol. ARNO is written in PERL and is structured around a document store and metadata database. Long term digital preservation is not cited as a key aim of this development. ARNO is open source.

8.2 CDSware (CERN Document Server Software)

Project home : http://cdsware.cern.ch/
Example installation : http://cds.cern.ch/
Core purpose : document storage and dissemination

CDSware was developed to store and disseminate scholarly publishing from CERN as well as providing a unified, portal like interface to existing repositories and information sources. The software is a Python/PHP implementation incorporating a MySQL database server and provides support for OAI metadata harvesting. Long term digital preservation is not cited as a key aim of this development. CDSware is open source.

8.3 DSpace

Project home : http://dspace.org/index.html
Example installation : https://hpds1.mit.edu/index.jsp
Core purpose : storage, preservation and access to a wide range of institutionally derived materials.

DSpace was developed by MIT and Hewlett Packard to address the preservation and dissemination needs of MIT. DSpace has at its core a model which attempts to address the real needs of a flexible institutional repository. This "Communities and Collections" structure aims to allow devolved control and administration from a central service to individual departments. DSpace is written in JAVA and utilises a PostGreSQL database layer. Long term digital preservation is a key aim of the DSpace system and MIT (and others) are investing considerable research in enhancing this aspect of the respository software. Version 2 of DSpace is expected to offer a more modular design and extended support for digital preservation functions. DSpace is open source.

8.4 ePrints
Project home : http://software.eprints.org/
Example installation : http://libeprints.open.ac.uk/
Core purpose : document storage and dissemination with flexible support for different format types

ePrints was developed by the University of Southampton with the aim of enhancing open access to scholarly materials. The software was developed in PERL and manages a MySQL database layer. It has now reached version 2. Uptake of the ePrints software has been considerable. Long term digital preservation is not cited as a key aim of this development. ePrints is open source.

8.5 FEDORA (Flexible and Extensible Digital Object and Repository Architecture)
Project home : http://www.fedora.info/
Core purpose : storage and dissemination with flexible support for different uses

FEDORA is a comprehensive repository and digital library system developed from the FEDORA architecture at Cornell University and the University of Virginia. FEDORA is currently being tested by a variety of institutions across the US and UK including the Library of Congress. The software is implemented in JAVA and the system relies on a range of developing standards including SOAP and METS. Long term digital preservation is not cited as an initial aim of this development but technology watch functions have been mentioned as development goals for new versions of the software. A related project "PRISM" is investigating digital preservation and is utilising the FEDORA architecture. FEDORA is open source.

8.6 MyCoRe
Project home : http://www.mycore.de/engl/index.html
Core purpose : storage and dissemination with flexible support for different uses

MyCore is a flexible repository system designed by a group of German universities working in collaboration. MyCore is written in JAVA and incorporates a flexible database layer allowing different database backends to be used with it (see Overall Repository Structure, below). Long term digital preservation is not cited as a key aim of this development. MyCore is open source.

9.0 Future and Related Developments

A range of institutions are involved in the development and maintenance of digital repositories which have a relevance to this discussion. Two key current developments are described below.

9.1 UK National Archives

Project home : http://www.pro.gov.uk/about/preservation/digital/archive/default.htm

The UK National Archives has recently developed a digital repository system, the Digital Archive, which is planned for an open source release. Digital preservation is a key aim of this development. Further details, including links to articles about the Digital Archive and the other work of the TNA Digital Preservation Department, can be found at the TNA website above.

9.2 DAITSS (Dark Archive In The Sunshine State)

Project home : http://www.fcla.edu/

DAITSS is an open source digital repository system designed and implemented by the Florida Centre for Library Automation. The Centre is developing the software to support over 50 public and university libraries in Florida. DAITSS is written in Java and incorporates a DB2 database layer. Digital preservation is a key aim of the system and an initial release is expected in the near future.

10.0 Current support for digital preservation in institutional repository software

The OSI guide to institutional repository software [2] lists only DSpace and CDSware as having a "defined digital preservation strategy". The CDSware strategy is reliant on migration on ingest to an archival format, in this case Adobe's PDF format. DSpace goes much further in defining categories of "Known", "Supported" and "Unsupported" formats [27]. Instances of DSpace can give different levels of commitment to preserving formats depending on the category within which they fall. Currently research and development work in a number of areas (in collaboration with Cambridge University and others) is being conducted with the aim of meeting these preservation commitments [28]. The current version of DSpace supports only very basic Representation Information including

minimal information like the deposited object's MIME type. While this is certainly a starting point, MIT are quick to acknowledge that this is not adequate for the purposes of long term digital preservation. Again, MIT is concerned about this issue and with Harvard University and the Digital Library Federation is leading the Global File Format Registry initiative to address the issue of Representation Information.

Digital preservation is not currently addressed as a key aim of the other repository software listed above. Clearly the provision of support for digital preservation in institutional repository software is at a very early stage. The key for current repository software is to provide flexible and extensible designs that can adapt to take advantage of digital preservation developments as they become available. As long as the main digital preservation issues described above are understood, this should be possible.

Addressing all aspects of digital preservation at the repository level is unlikely to be achievable due to the scale of the task at hand, and a degree of cooperation, sharing and external support will be required. In recognising this need, the JISC and eSCP are at the time of writing engaged in the establishment of a Digital Curation Centre that aims to provide support to existing digital repositories [15].

11.0 Recommendations

The key recommendations from this report are for the continued development of specific requirements for trusted digital repositories, and also for the creation of independent certification services for digital repositories that will evaluate how repositories meet these requirements. A clearer picture can then be presented as to how well institutional repository software, as well as specific digital repositories, can deliver effective digital preservation.

The report also makes the following recommendations:

- Preservation functions require integration with institutional repository design and must be considered from the outset both in the development of repository software and in the establishment of a given repository.
- Digital preservation developments are at an early stage in many areas so where possible, developments in institutional repository software should be made as modular, flexible and extensible as possible to allow integration with digital preservation developments as they become available. If an element of fore thought is given to the demands of digital preservation as described above, this process can be considerably simplified.
- Careful consideration must be given to the preservation needs of materials to be archived within an institutional repository. Very good reasons must be identified for not addressing digital preservation.
- Community wide efforts must be invested in developing the solutions to identified requirements for digital preservation in a repository context. The following areas are considered to be crucial:
    - Ingest

    - Representation Systems
    - Rendering
- Where possible concentrate development on *distributed* preservation functions which offer community wide sharing, and community based ownership and maintenance.
- Continue to build on the OAIS model (particularly with respect to Representation Networks, the value of which has been ignored in many sectors).

12.0 Glossary

Technology obsolescence : Where current hardware and software is superseded by new technology, which may not be compatible with older systems. This can lead to the loss of the ability to make sense of or "render" (see below) data.

Media obsolescence : Where storage media is superseded by newer media. Note that although much emphasis is often placed on the readable lifetime of digital media, it is almost always the obsolescence of the hardware that reads the media that prevents access to the data (see above). For example, the videodiscs upon which the BBC Domesday Project data are stored were designed to last a hundred years, but the special LVROM readers which read the discs have not been manufactured or supported for over a decade (and surviving units are now prone to breaking down). Long lived media does not equal long lived preservation.

Digital preservation : An organised series of actions taken to ensure continued use of digital objects is possible over time. The key elements of the solution include ensuring digital objects are : never lost or damaged, can always be found, and can always be understood. For example, placing a digital object into a repository where it will be backed up to prevent loss, where it will be given a unique identifier so it can always be found and where it will be linked to Representation Information which will describe how it can be rendered.

Representation Information : Metadata which describes how the bytestream of a digital object can be turned from a meaningless series of numbers into a human readable representation. This could include a simple textual description of the type of data in question, a detailed breakdown of a specific file format or a description of the tools which render that format.

Rendering : The process of displaying a digital object in a human readable way. For example, using WordViewer to display a Microsoft Word file, or running a BBC Micro emulator to render the BBC Domesday Project software.

Technology Watch : The monitoring of software and hardware dependencies to ensure that when technology obsolescence occurs, appropriate action is taken to update the relevant Representation Information and associated Rendering process

12.0 References

[1] “The Case for Institutional Repositories: A SPARC Position Paper”, Crow, R, http://www.arl.org/sparc/IR/IR_Final_Release_102.pdf

[2] A Guide to Institutional Repository Software, Open Society Institute, http://www.soros.org/openaccess/software/

[3] “DAEDALUS: Initial experiences with EPrints and DSpace at the University of Glasgow”, Nixon, W, Ariadne, issue 37, http://www.ariadne.ac.uk/issue37/nixon/

[4] “Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age”, Lynch, C, A, http://www.arl.org/newsltr/226/ir.html

[5] "The Open Archival Information System Reference Model: Introductory Guide", Lavoie, B, http://www.dpconline.org/graphics/reports/index.html#intoais

[6] “The Digital Preservation of e-Prints”, Pinfield, S, James, H, http://www.dlib.org/dlib/september03/pinfield/09pinfield.html

[7] Cedars : collection management, http://www.leeds.ac.uk/cedars/colman/colman.html

[8] OAIS, http://ssdoo.gsfc.nasa.gov/nost/isoas/

[9] “Trusted Digital Repositories: Attributes and Responsibilities”, RLG/OCLC , http://www.rlg.org/longterm/repositories.pdf and the follow up work undertaken by the Task Force on Digital Repository Certification http://www.rlg.ac.uk/longterm/certification.html

[10] “The Medium is not the message”, Holdsworth, D, http://www.personal.leeds.ac.uk/~ecldh/paper2.html

[11] “The Handbook”, Beagrie, N, Jones, M, J, http://www.dpconline.org/graphics/handbook/

[12] Migration on Request, University of Leeds, http://www.leeds.ac.uk/reprend/migreq/migreq.html

[13] Emulation, http://www.nla.gov.au/padi/topics/19.html

[14] Digital Preservation Coalition, http://www.dpconline.org/graphics/index.html

[15] Digital Curation Centre, http://www.ucs.ed.ac.uk/bits/2004/february_2004/

[16] Persistent identifiers, http://www.nla.gov.au/padi/topics/36.html

[17] METS, http://www.loc.gov/standards/mets/

[18] “Metadata Standards Framework : Preservation Metadata”, http://www.natlib.govt.nz/files/4initiatives_metaschema_revised.pdf

[19] Cedars (technical), http://www.leeds.ac.uk/cedars/archive/archive.html

[20] PRONOM, UK National Archives, http://www.records.pro.gov.uk/pronom/

[21] Global Digital Format Registry (GDFR), http://www.erpanet.org/www/products/Rome/Dale_**registry**_pres.pdf

[22] A Blueprint for Representation Information in the OAIS Model", Holdsworth, D, Sergeant, D, http://esdis-it.gsfc.nasa.gov/MSST/conf2000/PAPERS/D02PA.PDF

[23] “Survey and assessment of sources of information on file formats and software documentation”, Wheatley, P, http://www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf

[24] CAMiLEON, http://www.si.umich.edu/CAMILEON/

[25] Testbed Digitatle Bewaring, http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=181&categorie=2

[26] “Preservation Metadata : Pragmatic First Steps at the National Library of New Zealand”, Searle, S Thopmson, D, http://www.dlib.org/dlib/april03/thompson/04thompson.html

[27] DSpace supported formats, https://dspace.mit.edu/help/formats.html

[28] “A way forward for developments in the digital preservation functions of DSpace : options, issues and recommendations”, Wheatley, P, http://dspace.org/news/articles/DpAndDSpace.pdf