Edinburgh University Library

Digital Archives Research Project
A report and recommendations

Najla Semple

24/01/03

Contents

1 - Introduction

1.1 Abstract

The preservation of digital resources is of great importance to the academic community and society as a whole. We are in danger of losing much of our academic and cultural heritage through neglect of these burgeoning resources. This report, initiated by the Special Collections Division of Edinburgh University Library, focuses on research methods of preserving information generated in digital form, and provides a set of guidelines and recommendations for the benefit of the University of Edinburgh as a whole.

This report covers the main themes in Digital Preservation, legal dilemmas, metadata and methods of preservation such as the Open Archival Information System (OAIS) model, which is now considered to be the internationally recognised process model for Digital Preservation. It also focuses on a pilot project and demonstrates how some electronic university records have been archived successfully. Finally it outlines guidelines and best- practice in creating of a digital object for long-term preservation.

Acknowledgements:
As part of CURL, Edinburgh University Library sought to consult colleagues on the Cedars project and in that context we are very grateful. The reports Cedars[1] have published have been indispensable, as have the process models that the Nedlib[2] project have so clearly created and disseminated.

Chapters 1 and 2 cover the background to the project and how digital objects are prepared for archiving. For those primarily interested in the practical implementation of Digital Preservation strategies, the Guidelines at the end of the report, in Chapter 4, will be most useful. However, for fuller background and explanation, prior consultation of Chapter 3 of recommended.

[1] See Glossary in Section 1.4
[2] See Glossary in Section 1.4

1.2 Key recommendations

For The University of Edinburgh:

- On-going access to key electronic university records must be secured, despite the difficulties presented by rapid technological change.
- Digital Preservation techniques should follow internationally recognised digital archive standards and methodologies.
- Long-term access considerations must be considered when selecting digital file formats.
- A 'life-cycle' approach must be adopted when managing digital records – this will enable records to be managed, stored, retrieved, access-controlled and preserved throughout their existence.

For the creator of digital objects:

- Create files in non- proprietary, vendor-neutral formats.
- Use of structured XML files.
- Metadata about digital material should be recorded. This should include administrative and technical information, as well as details of the preservation process.

For web authors:

- Use of the Dublin Core standard in web pages.
- Use of valid HTML, and where possible, it should be XML compliant.
- Web document templates should be used where possible when authoring sites to be located on the university site.
- Any major changes to a site should be noted in the accompanying metadata.

1.3 Background to the project

Recognising the growing need for a Digital Preservation strategy within The University of Edinburgh, a Digital Archives Research Officer was appointed in June 2001 to research the preservation of core University publications.

Digital resources are part of the responsibility of Edinburgh University Library, and are therefore subject to the same preservation policies as other media. Policies for digital preservation should be co-ordinated with those for the preservation of analogue material, such as manuscripts, microfilm, or paper journals.

The material that the University of Edinburgh produces in electronic format is extremely wide ranging - from committee papers to cutting edge scientific databases. This project focused on data produced by administrative departments that support the running of the University. In terms of current archiving techniques within the university, there is both a University Archivist, and a University Central Filing System[3]. As yet, these cater for non-digital materials only.

Much University administrative information can be found online, and can be grouped loosely into the following: websites, committee papers, minutes, and the yearly University Calendar. Electronic student records exist in the NESI[4] database, due to be expanded over the next few years as part of a wider student management system. It is vital that a digital preservation procedure is put in place to ensure that this ever-increasing amount of digital material is retained for the future.

Scope

The timeline for the research project was 12 months, giving sufficient time to undertake the research; to pilot a project and to focus on a practical outcome.

It was not within the scope of the project to look into the contentious issue of how libraries are to provide long-term access to electronic journals. There are other projects addressing this, such as LOCKSS[5] and JSTOR[6]. The Mellon Foundation's e-journal archiving programme [7] was also set up to work a distributed system for the long-term retention of electronic-journal files, and has provided some defining work on the topic.

Methodology

Visits were made to similar Digital Preservation projects and initiatives.
Conferences were attended, including the Cedars dissemination event[8]. Talks on this project were delivered to Scottish Cultural Resources Access Network, Society of Archivists and to groups within

[3] See http://www.sec.ed.ac.uk/ExtRel/Filing/cfs_guidance.htm#Archive policy
[4] The University of Edinburgh student record system- Information can be found at: http://www.registry.ed.ac.uk/StudentSystems/nesidocumentation.htm
[5] See http://lockss.stanford.edu/
[6] See http://www.jstor.org
[7] See http://www.diglib.org/preserve/ejp.htm
[8] See http://www.leeds.ac.uk/cedars/dissem.htm

The University of Edinburgh. A large Digital Preservation seminar was organised in September 2001, which was well attended.
Courses attended were the UCL Electronic Records Management summer school, Edinburgh University XML course.

Digital archiving technology advances very fast - it is important that preservation strategies should be made flexible and adaptable. The preservation method recommended is that of keeping digital objects data format neutral, because a neutral format is not dependent on the file format that created it.

Archiving techniques developed are loosely based on the OAIS model, fully explained in Chapter 3. It is recommended that the project could benefit from the implementation of this model on a larger scale.

Pilot project

The project decided to archive a corporate University record to demonstrate how files can be successfully archived and that access can be guaranteed in perpetuity. The record that we chose to concentrate on for the purposes of the project was the University Calendar, which carries an important legal status at the corporate level and this continues the present function of the Library in providing a preservation copy of the printed Calendar through the University Archives.

In addition to producing general guidelines and recommendations, this approach also meant there has been a practical aspect to the project through initiating the digital archiving of the online version of the current University Calendar. The project worked in liaison with the Calendar team and facilitated their planning processes in the future implementation of the University Calendar, which is planned to be exclusively online. As this document has legal status, it was identified as a suitable pilot project for preserving access.

Currently the recognised legal version exists as a paper copy, but there is an online version available on u website. This consists of both HTML and PDF files. There is as yet little ‘dynamic’ information, or complicated web-based features. Corrections made form a separate file within the site. There is a plan to alter the face of the Calendar entirely and make it fully electronic as part of a wider integrated student management system. It will be updated regularly and it will feed into other areas of the online management system; courses list, prospectus, student portals etc. This will pose challenges for the future archiving of the digital Calendar, so it is essential that some process is set in place to archive it.

Outline of digital archiving model

To support the pilot project, a small-scale digital archive was established. This facilitated research into archiving methods and checked its suitability for future archiving.
The model I suggested adopts and simplifies the themes and recommendations of digital archiving models that other large institutions have developed. I have purposefully made it as simple and as interoperable as possible, so that at a later date it can:

a) Function alongside other digital archives.

b) Be made more sophisticated and be built into an automated system.
c) Be altered as digital archiving techniques develop.
d) Be user-friendly for those who find themselves faced with archiving digital material.

The archive will not provide online access to material - it will be stored off line and the only access to it will be via the archivist. At a later date, access might be via file transfer protocol (FTP) with the appropriation of strict access authorisation for users of the archive. It will be possible to locate, but not access, the archive via the library OPAC.

This problem of access might be solved by continued involvement in the OCLC project – (*see Chapter 3*), which will provide:

a) Storage of archived files on the OCLC server (duplications of content of our local archive)
b) An appropriate user interface for access and irretrievability of digital objects.

1.4 Digital Preservation

Digital Preservation within Edinburgh University can be defined as the actions taken to ensure enduring access to digital resource content over time to support both academic teaching processes and support group services.

In the light of rapidly changing technological infrastructures, the growth of digital information in UK Higher Education institutions far outweighs the precautions taken to ensure continued access to this data in years to come. This issue needs to be tackled so that this information is available in the future.

Electronic information is forming an increasingly large part of Edinburgh University's academic and administrative activities. The prospect of losing this data through poor management has begun to raise awareness of the fragility of this medium.

For any organisation or department that produces electronic information, a strategic preservation plan is vital. This has a bearing on the life cycle of a digital object; correct procedures followed at the creation of a document or web page can ensure it can be preserved for future access. The volume of electronic data created means that preservation of all information is impossible - selection procedures and retention schedules should be built into the preservation process.

The Threats to long-term access

1. Decay of the storage medium

This occurs when the medium on which the object is stored becomes unreadable. This could be the case of portable carriers that do not have readers for example, 12" floppy disks.

2. Technological obsolescence – software and hardware

More problematic than medium decay is the threat that digital files are created using software that soon becomes defunct or unavailable. Digital data exists as a series of bits and bytes and without software the information cannot be accessed. A digital object is also heavily dependent on the hardware environment in which it was created. Nearly all commentators agree that technological obsolescence of both hardware and software represents a far greater threat to the preservation of digital archives than does media storage.

Responsibility for digital archiving

Edinburgh University Library, which has always had a responsibility for preserving scholarly material, owes it to the scholarly community to preserve its digital heritage for the future. However, given the expensive nature and expertise of maintaining electronic material, there is a view that the creators of digital data such as electronic journal publishers have an onus to ensure continued access to back-files of digitally held material.

Kevin Guthrie puts this succinctly,

"There is not yet an equivalent system in place to protect the electronic literature being published today. How can we be sure that such a system will evolve? Where will the resources come from to support it on an ongoing basis? Who will accept responsibility and accountability for such a system? These are just some of the challenges that lie ahead if the academic community's commitment to archiving is to make the transition to the digital age."[9]

In the USA, the Mellon Foundation has taken the view those actions necessary to maintain access to digitally published material should be undertaken in a distributed way, in order to share the burdens of responsibility[10]. It might well be that digital archiving will have to be carried out on a national level, whereupon libraries share big repositories of historical information.

[9] Guthrie, K (2001) *Archiving in the Digital Age*, Educause Review, November 2001 Available from: http://www.educause.edu/pub/er/erm01/erm016w.html

[10] See http://www.diglib.org/preserve/ejp.htm

1.5 Glossary

Digital preservation uses some specialised terms. As it involves a number of different stakeholders of varying professions, it is important that definitions are clarified from the outset.

Archiving	Computer archiving - data that has been stored and is no longer immediately accessible to users through networks. Traditional archiving - providing access to a managed collection over a period of time. Thought and preparation is given to what resources are stored, how they are maintained and accessed, and by whom.
Audit trail	Stored knowledge of all the actions and changes that occur to an archival record, both before and after preservation.
Binary data	A file that contains codes other than ASCII characters. Thus this is data which is computer readable but in some cases not human readable; unlike text files, where there is mapping between bytes and characters. Formatted files are usually binary files.
Byte stream	This is the smallest unit of computer information. It is made up of bits, and exists of a series of 0's and 1's.
Cedars Project	This 4-year Leeds University based project researched the practical issues of Digital Preservation – it has released a number of invaluable recommendations and guidelines. See www.leeds.ac.uk/cedars
Checksum	A method of checking a file for any changes. Each object is allocated a unique number. When it is extracted from the archive, it should generate the same number – if not then it is clear the file has been tampered with.
Compression	Reducing the size of a digital file, without changing its contents, for transmission or storage purposes.
Digital object	The document, record or unit to be archived. It is a term used frequently in this report.
Digital signature	Appendage to an email or digital file that allows a recipient to verify the source of the data.
Document	A piece of information used day to day and can exist as many different versions.
Encryption	Conversion of a data file into a secret code to allow it to be transported safely.
File Transfer Protocol (FTP)	This is a method of transferring, sending and receiving files across the internet.
Information management system	Any electronic system that processes information in digital form, can allow information to be accessed and manipulated.
Legal Admissibility	An institutional record has to have proof that it is what it purports to be, and is the original record that was created. This is pertinent in the digital age given the ease of altering documents.
Metadata	This can exist in digital or non-digital form. Metadata is information about information; it often accompanies (or is embedded in) a record and it is necessary to yield information about a digital file and audit its history, as well

	as providing valuable resource-discovery.
Nedlib Project	A European project, set up to create a practical archiving model primarily for national deposit libraries. It has released useful documentation and booklets. See http://www.kb.nl/coop/nedlib/
OAIS Model	Open Archival Information Systems is a model created by the US Space community, which defines a process model for digital archiving and archival concepts. It is applicable to any archive, and should be interoperable with any system.
Provenance	The origin of a record and its custodial history. This should be recorded.
Record	Documents created by an institution in the course of its business and retained as evidence of its activities.
Stakeholders	Institutional departments involved in the life cycle of the digital resource, they have an interest in policies for the creation, access and destruction of files.
Underlying Abstract Form	The basic properties of a file that are left when the file is separated from the software that created it.
Web Harvester	A machine that automatically searches a given web page and downloads web pages by use of a 'spider' which crawls all the links in each web page. It can also be used for collecting metadata.
WORM	Write Once Read Many- A form of storage that allows digital media to be stored once and read many times- but never deleted or updated.

Chapter 2

Practical solutions to Digital Preservation

There are much information available detailing current digital preservation methods. The aim of this report is not to duplicate this information, but to raise awareness of the issues and provide references where appropriate to more detailed research.

2.1 File types

It is first worth defining suitable file types for preservation:

Non-proprietary file types

This means that the software used to create the files is open-sourced and freely available. It is not reliant on a company to manufacture, sell and patent the software.

ASCII	American Standard Code for Information Interchange- The basic character sets to represent current day computer languages. It is an 8-bit code – each byte represents a character. This is different to a binary data, as it is a text file. A computer programme would not be stored in an ASCII file.
RTF	Rich Text Format- this is a Microsoft standard. It is a simple text format but expands on ASCII as it defines formatting information.
PDF	Portable Document Format, a format created by Adobe inc., is an ideal format for delivery and presentation of data in a fixed layout. Although this format is proprietary in that a company owns it, the software is free and open-sourced. The PDF standard has been published and is freely available. It is a suitable storage medium because even if Adobe stopped making it available, it would be easy enough to write the programme from its publicly available source. There is nothing to say however, that another company, which might make it proprietary, might buy up Adobe. Research into preserving PDF is vital- a large majority of publications submitted to libraries are in PDF format. The new PDF format will incorporate the functionality of XML.
TIFF	Tagged Image File Format - this is considered a robust medium for storage of images. It is considered very robust and unlike PDF, it is not reliant on proprietary software to read it.

Mark-up languages

These are vendor neutral and non-proprietary. Provided that a specification is stored and a browser is preserved, then mark up language files should be readable into the medium-term future.

SGML	Standard Generalised Mark-up Language- a standard for defining mark up languages. The tags used indicate structure rather than formatting.
HTML	Hyper Text Mark-up Language. A subset of SGML, this is the language of the web, which creates web documents by marking-up documents by use of a series of tags and attributes –

	It supports hyper links and is readable in all browsers. Note that there are many different versions of HTML.
XML	eXstensible Mark-up Language – This is more structured and rigid than HTML. This application-independent format enables any number of document types to then be re-created from the core data (e.g. Word, HTML, PDF), thus ensuring that future generations are not hardware or software dependent. *See extended section on XML in section 2.2.*

Databases

A database is a method of storing structured data in sets of columns, which determine type, and rows, which represent each occurrence of the data type. In terms of long-term preservation, databases pose great problems, both because of their complicated nature and because they are often proprietary. In order to interpret information residing in a database, one is reliant on the database software that created it.

Websites and databases

It is increasingly the case that much of the web is now driven by databases, and gives it its interactive quality. Whenever a user poses a query at an information service, a dynamically generated page is returned that holds information extracted from a database.

Data on the web is accessed by database queries, and in turn web pages are created 'on the fly' to a user's browser. It will be nearly impossible to archive these immediate pages. However, it might be possible to archive the database that drives the website, and in turn re-run it via a browser. This becomes a problem however when the database is regularly updated.

Recommendations for File Types:

- File formats must be non-proprietary.
- The more simple the file format, the longer the likelihood of long-term preservation.
- If a file type is proprietary, as much information as possible must be recorded about the software package.
- XML files should be used for content creation, if possible.

2.2 XML - eXstensible Mark-up Language

A great deal of information about digital preservation promotes the use of XML files as a medium suitable for storing information for long-term access and preservation. A substantial amount of this chapter is devoted to the benefits of XML, as it is increasingly being implemented in a digital archives context. XML is a way of storing information and processing it to make it readable in a variety of formats,

"What one can see on a desktop computer in a single screen must be rendered differently to the user of a portable telephone. Thus, I believe that we may see more movement toward the storing of information in application-independent and structure-independent forms" [11] Barry, 2001

For the purposes of this report, it is worth outlying some key points in favour of using XML as a medium for marking up and storing records.

Positive points about XML

XML:

Provides an open standard for long-term accessibility.

Isolates structure from content - this means that the content information can be kept separately from formatting information about how the resource should be presented.

Structured XML marks up text so that components can be individually identified.

Re-usable content People with different needs can use the same content - e.g. information can be transformed into a Braille browser or multimedia applications.

Supports a wide variety of applications - XML can be read on any current or future operating system.

Other Features of XML

Document Type Definition

A DTD is a prerequisite for every XML document and a link to one is needed in the XML document header. It defines exactly what fields can go into a document, and it is the DTD that gives XML its rigid structure. A DTD, for example, will insist that all Dublin Core metadata elements be populated.

[11] Barry, R (2001) *Making a Difference –Revisiting the IU ER Project Five Years Later*: Available from: http://www.rbarry.com/

Style sheets

XML needs to be *processed* in order to make it understandable and readable. For example, one needs to translate a set of plain XML files into HTML, or PDF to render the data. This involves writing a standard script in XSLT, a language for transforming documents. XSLT is two languages in one; it is both a transformation language and a description language. It can be used in place of Cascading Style Sheets to indicate how the document should be formatted, although the latter can be used in XML documents to indicate the formatting.

XML and databases

Ideally one would want to have the functionality of XML, but in a database format. XML is better equipped to handle complex tree-structured data, such as documents. Its non-process specific format is not optimised by programmes on which databases are designed.

It would be feasible however, to map a database to an XML structure by storing the XML in a series of tables, giving each entry an identifier and then building links between them. This is possible to some extent, but it will never be possible to have the full functionality of a relational database, although big database suppliers (e.g. Oracle, Sybase, Informix) claim that they will be able to do this in the future.

XML for metadata storage

For the purposes of this project, it has been established that the preservation metadata should be kept in XML format. A DTD has been adopted to support the implementation.

How easy is it to convert records to XML?

It might be possible at a later stage to convert existing records to XML to ensure permanent preservation. If this was easy enough, it might be the answer to many archiving problems. However at present no software exists to do this, and it would have to be a manual exercise of marking-up documents.

The best solution, however is to recommend web authors to use XHTML (or valid HTML) which can then be transferred to XML at a later date.

Recommendations for XML:

- Use DTDs available on the University website.
- XML is easy to create - this can be done using any text editor. Ideally however, one would want to create XML data via some other application. An example of this would be to type into a data entry package with a user interface, meanwhile feeding data into XML files, which in turn will lead to a stable storage base, and a source for other delivery formats.[12]
- A whole range of programmes are needed in order to make XML files understandable e.g. XSLT, XPATH.
- Images and links are not directly supported, however these can be embedded. Note that XML doesn't read *binary* data.
- XMP - Adobe eXtensible Metadata Platform is a new format that will allow embedding of XML structure to assist structure-based searching. An XML-based form of PDF might be easier for existing XML-based search and analysis software to process. [13]

Read more on XML:

www.w3c.org- The world Wide Web consortium has had a part in assisting the creation of XML and provides authoritative sources on the technology.
http://msdn.microsoft.com/library/default.asp?url=/library/enus/dnword2k/html/odc_expwordtoxml.asp- A good article on how to convert MS Word to XML.
http://www.adobe.com/products/xmp/pdfs/whitepaper.pdf This details the new Adobe PDF/XML technology

[12] See University of Glasgow 'CDocs' project for an excellent example of creating metadata Available from: http://www.gla.ac.uk/InfoStrat/ERM/Reports/

2.3 Legal implications

Legal issues are an important consideration when embarking on the preservation of digital content. One major area of concern for information professionals is the admissibility of records in a court of law. A large aspect of this project therefore, has been to research the legal implications of digital archiving.
With the pilot project, it had to be ensured that the Calendar was archived as a 'legal' document, and could be retrievable in that future as a legally admissible record of the University's actions.

Corporate publications act as legal proof of the actions of the institution, created in the regular course of business. In a records management sense, the degree to which a record can be considered reliable is dependent upon the level of procedural control exercised during its creation and management.

With regard to *legal admissibility* and trustworthiness, there remains some ambiguity about the reliability of digital records. This is largely because they are open to manipulation and alteration. Thus an object's authenticity is validated by the fact that it hasn't been tampered with or changed while it is in a digital repository.

However, preservation methods often mean that files have to be migrated to different formats to keep them accessible in digital form – in doing so, it may be that certain significant properties of the file are lost. This challenges the nature of admissibility. A file is thus more admissible if all actions taken upon it have been recorded in accompanying metadata, (see Chapter 3 for more information on metadata). Regular audits trails should be documented and should reflect requirements for the legal acceptance of records. The change history of the object must be maintained, as well as retention schedules. File security information should also be recorded, such as *checksum* strings (see below) to prove that no change to the bit stream has occurred.
The question of authorised access control must be addressed –archives must ensure that unauthorised people cannot access digital records, especially if they have strict copyright regulations associated with them. This will verify a trusted and secure provenance.

Edinburgh University has to be aware of the legislation for archiving digital content. It also has to ensure that information is secure for legal accountability.

[13] http://www.adobe.com/products/xmp/pdfs/whitepaper.pdf

Legislation

The following statutes and laws bind any digital archive:

Copyright, Design and Patents Act 1988

Our legal system has not yet fully taken into account documents in the digital domain, yet we still have to conform to laws made for analogue material; this is especially the case with the Copyright law of 1988.

This law applies to literary, musical, dramatic, architectural or artistic works. The creator of material published in the public domain, or in some cases the publisher, owns all the rights to the object- it cannot be copied without the full consent of the creator. Copyright is automatically understood at the point of creation. Copying digital materials into an archive will therefore be a problem for any archive that Edinburgh University sets up.

The law has been amended considerably however, and one amendment of note was made in 2001, by a European directive, which gives exemptions for libraries to copy and store digital publications;

"Member states should be given the option of providing for certain exemptions or limitations for cases such as educational and scientific purposes, for the benefit of public institutions such as libraries and archives…" European Parliament 2001[14]

Archiving software in order to be able to read a file in the future can also be problematic- one might need to keep a copy of the software to re-run a file in later years, or to be used in the emulation process (*see Chapter 3*).

Web archiving: Archiving any website is essentially making a copy of it and this will come under the framework of Copyright law. This will not be a problem with public material, but it should be noted that websites do sometimes contain a ' robot exclusion file', which will prohibit harvesting and this should be respected. Owners of sites will have to be contacted and permission gained.
In many cases, copyright regulations hinder the process of Digital Preservation and prevent certain valuable digital materials from being archived.

[14] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 Official Journal L 167, 22/06/2001 p.10-19.

Intellectual Property Rights

A major issue in relation to copyright is Intellectual Property Rights and ownership. When one hands over a document to be archived, the author loses control over what happens to it- this can be an issue of concern in the digital domain, as a file can be so easily be distributed to any number of people who access the archive. IPR therefore has to be clarified at the point when the document is given over for archiving. The author can request that access is disallowed until the given date, and this is written into the metadata. Decisions can be made as to who can view the document and access levels have to be agreed upon, before ingest into the archive.

Another option is that 'ownership' of the object could be handed over to the archive,

"*When a work is handed to an archive to ensure its long-term preservation, the author loses some control over the document since it, in a way, leaves his or her sphere of influence. Basically, two points have to be clarified between the management of the archive and the holder of the intellectual property rights."* Aschenbrenner, 2001[15]

Data Protection Act 1998

Data protection will affect the archive in the following two ways:

1) It is the archive's responsibility to protect access to personal data, if the data refers to living persons.

2) The archive also has to prove that the data has not been tampered with or altered in any unauthorised manner.

In some cases, owing to a court order or such like, it may even be necessary to amend or delete specific information from an archive, to meet the requirements of the Act. In that case, information has to be fully and easily accessible to do so. The Act also calls for the effective management of retention schedules.

Personal data should not be kept for any longer than needed. Duplication of data should be avoided, even within storage mediums or on university servers. This is made complicated in the digital arena, given the ease of copying. An attempt must be made to ensure that copying is strictly controlled.

Although access to personal data within the institution is necessary, especially material such as Edinburgh University student records, the Data Protection Act stipulates that the data has to be stored and fully. Appropriate access controls have to be established.

Edinburgh University has a Data Protection Officer who can be contacted for more information about this act; http://www.dataprotection.ed.ac.uk/principles.html

[15] Aschenbrenner, A (2001) *Long Term Preservation of Digital Material- Legal issues* Available at: http://www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/Legal_Issues.html

For more on the Act, see- http://www.dataprotection.gov.uk/

Freedom of Information Act (Scotland) 2001

Note that this Act differs to the English Act of 2000. This recent Act has a bearing on any public authorities in Scotland in that members of the public now have the right to access information that is held by these authorities. Edinburgh University is one of these institutions.

There are certain exceptions to the categories of information accessible, but the categories which are, have to be organised in such a manner that if a request comes through to see a record (regardless of the material being analogue or digital), the record has to be instantly retrievable and come attached with a verified *provenance.* In short, the archive will demonstrate that all information on a specific topic is available for review at any time.

This means that, in effect, strict *electronic records management (see Section 2.5)* procedures have to be in place. Managing all the requests will be time-consuming for Edinburgh University, thus e-records have to be managed properly from the start. The Code of Practice [16] sets out the conditions for which public bodies should create, manage, retain and dispose of their records. Retention schedules should be allocated, and all records should be indexed.
Edinburgh University has its own FOI committee group;
http://www.planning.ed.ac.uk/FOI/pubbackground.htm

See-http://www.hmso.gov.uk/acts/acts2000/20000036.htm,

Electronic Communications Act 2000

This Act does not directly affect the archive, but it is worth being aware of it as it acts as a formal recognition that electronic records act as legal records if verified by *digital signatures* (see section 2.4).

See: http://www.hmso.gov.uk/acts/acts2000/20000007.htm

2.4 Security

Accountability and integrity of digital information

In terms of Edinburgh University's records, a large institution such as Edinburgh is held accountable for the records it has to keep. This is demonstrated by the recent information acts that are now incumbent upon us. Accountability is key in record keeping and in preservation; one has to be able to prove that the records that are accessed in years to come are the same ones that were originally created, and that no one has altered them in any way. Documents have be able to prove that they are what they purport to be.

[16] *Code of practice on the management of records under the Freedom of Information Act*, Public Records Office Available at: http://www.pro.gov.uk/recordsmanagement/freedomofinformation.htm

At this point, it is important to differentiate between a 'closed' archive and andopen' on-line archive. Having an online archive brings in other considerations. For example, if ingest methods are supported over a public network, then security is a issue to consider as files, potentially, could be accessed if networks are 'hacked' into.

For the purposes of this study, it is proposed that an archive should be established that is stored off-line, access being controlled via the archivist. This trusted digital vault would prevent access to unauthorised modification of objects, the archiving processes of which will ensure the maintenance of legally acceptable records.

The DISC PD 0008 1999- 'Legal admissibility and evidential weight of Information stored electronically'

The 'Legal admissibility and evidential weight of Information stored electronically' Code defines best practice for electronic storage of material and access to and from the archive. The procedures recommended are now de facto and are recognised in a court of law.

The code is mainly concerned with an electronic records management *system,* and for ensuring that records can be verified in court as admissible and authentic records. All aspects of the system must strive to be fail-safe.

The document also raises the issues of digitisation of documents and the systems used in creating them; this does not affect our digital archive at this point in time.

These are some of the points that concern a digital archive.

Part 4 – Procedures and Processes

Part 4.11 - Back up

- The code argues that back up data may be used as evidence of security of the stored information. Section of the code 4.11 stipulates that the back up data should also include *audit trails.*
- The procedure should include the secure off-site storage of these back ups. These back ups should themselves be tested regularly. Index files should also be backed up to assist in the rebuilding of the system if it happens to fail.

Part 4.13 - Security and Protection

- Edinburgh University should adopt security guidelines. These should include details of controlled access to various levels of the system such as data input and retrieval. User access controls are very important.

- It also says that the central part of the system (file servers and data storage) should be installed in secure areas.

4.13.2 Digital signatures and Encryption

- The code recommends that creation of electronic signatures be documented. It also says however that managers of the system should first fully understand how digital signatures function, as they are prone to risk.
- *Encryption*, on the other hand, makes clear text unreadable and is based on a set of secret keys. Only the person who has the 'secret key' can read the message. This gives the file *integrity*, as it cannot be modified in transmission, and confidentiality. Encryption keys should also be kept securely and restricted to those whom have authorised access. It is a complicated procedure and watch out- if one loses the key, the record cannot be accessed!

4.13.3 Transportation of data

- This is if information is sent to the archive in physical form, such as in the form of CD-ROM. The code says that the material should be accompanied by a control document stating the identity and number of items sent.

4.16 Self-modifying files

- Some files may contain an automatic code whereupon the file is modified every time it is opened. An example would be the insertion of the current date in a Word document. This has obvious implications for the authenticity of the object and how it originally looked. The code recommends that this code should at least be documented when archiving the files.

4.19 Version control

- Some documents are subject to change and updates or modifications. Several different versions may exist, produced at different times. The code recommends that a version control procedure be adopted. Any changes should be documented, and newer versions should be kept for at least the same amount of time as the original copy.

4.20 Maintenance of the documentation.

- Procedures should be put in place so that the associated documentation (in our case, metadata) should be kept up-to-date. This metadata should be retrievable at any given date.

Part 5 - Technologies

5.5 System integrity checks

- Checks should be carried out to ensure the integrity of the data in the system. It recommends that *checksum*s ought to be carried out on files. To protect a system from 'buggy' software, protection software should also be built in.

5.11 Data file migration

- With regard to Migration processes, it is stipulated that files are stored in 'industry standard format' or if not, that software is always kept available to view the files. When migrating, metadata and *audit trail*s should be kept, and the integrity of the file should not be lost.

5.12 Information deletion

- Any deletions should be audited and authorised. Although it is not possible to delete from the WORM archive *(see Chapter 3)*, even the removal of the index references is synonymous with deletion of the material itself.

Part 6 - Audit trails

- The code breaks up the *audit trail* information that should be provided for each process:

  System- records should be kept of the history of the management system, details of migration procedures.

  Stored information- details should be kept of the actual information capture, date unique ID, size of the file, format, any changes, and destruction of the object.

  Audit trail data itself is subject to the same rigours as stored information, back ups should be kept etc. If the possibility exists that the *audit trail* data could at all be tampered with, then this will reduce the evidential weight of the object it is associated with.

Recommendations for legal admissibility of files:

- It would be desirable if all digital information consisted of the record *and* additional information that is needed to maintain the evidential value of the record.
- Integrity of data should be checked using *checksums*. Digital signatures could be used to this effect to prove they are the original documents. Encryption methods on the other hand, are more complex and are more for use for *exchange* of information over networks.
- Access to storage media will have to be recorded and access authorisations will have to be allocated.
- Note that *compression* of files will also alter a file's authenticity and this should be noted in the metadata.
- Check if online documents contain ' robot exclusion files'- legally one is not allowed to harvest these files.
- Any digital archive has to make sure that there are provisions for frequent backup or duplication and stored *off site* disaster recovery.
- Disposal- to complete the electronic document life cycle, disposal must be adequately carried out ' reformatting or overwriting' N.B. *deletion of files* is not the same as destroying them.

Read more on legal issues:

The World Intellectual Property Organisation provides a good background to the issues of Intellectual property rights, with an emphasis on information in the digital age, and is up-to-date in any changes in copyright law worldwide. Available at: http://www.wipo.org

BSI/Rob Allen et al. (1999) *Evidential weight of information stored on electronic management systems.* DISC PD 0008[Code of Practice] BSI; London

Freedom of Information Act 2000- model action plan for achieving compliance with the Lord Chancellor's code of Practice (2001) Available at: http://www.jisc.ac.uk/info_strat/temp/hemapv_1.doc

2.5 Management issues

Electronic Records Management

Like any asset, information needs to be validated, secured, measured and managed effectively.
Well-managed records are a vital part of an organisation's information resource. They enable to retain a corporate memory of its activities, and demonstrate accountability for its actions.
Records can also prove whether the organisation has met certain legal obligations and senior managers often consult them as proof of activity.

It is worth clarifying the difference between documents and records:
A **document** is material created in the workspace, sometimes on a daily basis, and no particular thought is given as to how long they should be retained.
A **record** is managed by an organisation; it has been carefully selected from the documents as a relevant piece of information, which records an organisation's activity. Records are subject to retention schedules; a large amount of records are *not* kept forever.

In order to facilitate effective management of Edinburgh University's records, it is vital that a records management strategy is put in place. When accessing a record, its context has to be understood. Legally, they must have proof of authenticity, integrity, and non-repudiation (this prevents the owner from disowning the digital object). Records should all follow a 'life-cycle', and should be managed in a controlled, structured way.

Electronic Record Life Cycle
The electronic life cycle would, ideally, consist of the following elements:

- **Capture**
- **Appraisal**
- **Technical appraisal**
- **Retention schedules**
- **Metadata recording**
- **Access**
- **Preservation**
- **Destruction**

Capture
The term capture is used to encompass the processes of registering a record, deciding which class it is to be classified to, adding further metadata to it, and storing it in the ERMS. This could be done during ordinary business workflow.

Appraisal and Retention Schedules

Assessing how long a record should be kept accessible, before it is disposed of.

Technical Appraisal

Is the object readable? Does it have adequate documentation? Be aware of budgetary implications for storage.

Recording Metadata

This is vital to ensure that all the necessary transactions occurring to record have been recorded, as well as a description of a record for cataloguing purposes. Automated *audit trail*s need to be developed that document all business processes, including activities relating to the creation, and change history.

Access

Access may be provided via a website. Security issues and access controls are a factor, as well as searching capabilities- it might be linked to a catalogue.

Preservation

The digital object will sit in a managed repository along with its metadata files, and care will be taken to ensure that it is migrated or refreshed to ensure that files are continuously accessible over time.

Destruction

This could be done automatically according to retention schedules. Personnel must be notified when this is done. Care must be taken to fully delete files- often only the *route* to the file is deleted.

A note on Deleting files

This is not as straightforward as it would seem. Deleting a file simply means that the route to it is being terminated, not the file itself. Deletion requires writing over the original file. The Edinburgh University digital archive as proposed will in fact be based on a technology from which no files can be deleted (see Chapter 4).

Read More on ERM:

Much information is available regarding ERM. The Public Records Office has issued a number of excellent guides and can be found at:
http://www.pro.gov.uk/recordsmanagement/eros/default.htm

The National Library of Australia is also at the forefront of ERM:
http://www.naa.gov.au/recordkeeping/er/summary.html

Another Australian project is the VERS, the report of which details all their ERM strategies:
http://www.pro.vic.gov.au/vers/welcome.htm

This was a study commissioned by JISC specifically focusing on the management of electronic-records in higher education institutions.
http://www.jisc.ac.uk/pub01/records_lifecycle/report_introduction.html

Collection and Stakeholder Management

It is often said that one of the biggest obstacles to ensuring long-term preservation of records is *management* of the digital resource. As management procedures differ from those for analogue material, a separate collection management policy for digital materials is crucial. This should be implemented right at the beginning of the digital object's creation. Collection management for digital materials however, should be able to sit comfortably with policies for analogue materials.

Selection: It has to be ascertained as to who will want to access a digital object and for what purpose. Historians in a university will have little use for locally produced committee minutes, but will have much use of contemporary websites.

Levels of archiving: Decisions have to be made as to how long to keep files. If they are not kept for perpetuity, but simply for a matter of a few years, then it may be acceptable to keep them in their proprietary formats.

Original form: Another issue that acquisition management will have to bear in mind is that the *original form* of the object may well alter in the preservation process. Unlike analogue materials, which keep their shape and form, digital materials may well have to be converted or migrated- the original text only may be extracted from an online document. When data providers are submitting a digital object for archiving, they will have to stipulate how much of the 'look and feel' they want to retain, bearing in mind that the more simple the form, the longer its perpetuity.

Collection Management is similar to the *electronic records life-cycle* (see Section 2.5). Cedars suggest that the seven main areas of collection management are: *Selection and acquisition of digital materials, Organisation, Storage, Access, De-selection and Preservation*[17].

An additional aspect of management is that of dealing with all the ***stakeholders*** who are involved with the creation of the resource. Technical preservation hurdles can be overcome if a holistic approach is taken, and all stakeholders in records management adhere to the life cycle.

If digitisation procedures are carried out, it is worth stipulating preservation metadata fields before the digitisation is carried out.

To some extent, the library could stipulate the electronic formats that it is prepared to archive, advocating open standards. If the library becomes the archive of back issues of e-journals, they will have to arrive to the archive in preservable and non-proprietary formats. This approach will take some time however and may pose a problem with some content providers.

[17] See http://www.leeds.ac.uk/cedars/guideto/collmanagement/

Costing:

There have been a number of detailed investigations into the costs of running a digital preservation service.

Read more on costs of Digital Preservation:
See Mary Warner's 'Why do we need to keep this in print, it's on the web' for a good section on costing: available at: http://libr.org/PL/19-20_Warner.html Hendley, T *Comparison of methods and costs of Digital Preservation 0*(1998) British Library Research and Innovation report 106, British Library Research and Innovation Centre, London, Available at: http://www.ukoln.ac.uk/services/elib/papers/tavistock/hendley/hendley.html Cedars guide to collection management, 2002 available at: http://www.leeds.ac.uk/cedars/guideto/collmanagement/

The overriding emphasis is on institutions tackling Digital Preservation before it is too late- large resources and budgets are used for digitisation programmes- the library should initially ensure that these newly created resources can be preserved in perpetuity. It has been suggested that collaboration with other institutions is an obvious solution- digital archives can be shared, reducing duplication costs.

Among the costs involved will be:

Archival storage - although costs of storage material reduce over time, this is a considerable cost to bear in mind, especially if material is to be kept on servers.

Metadata creation - Much of the work involved with metadata is comparable with traditional library cataloguing, and is likely to be comparable in cost.

Rights management & Legal issues - all data will have to be checked to ensure that there is no infringement of legal acts and intellectual property rights regulations.

Maintenance of the resource - this will involve regular checks to see if the resource is still valid and accessible.

Migration of files - this will be a time consuming and costly exercise. Again the resource will have to be checked after each migration to verify if it is readable.

Delivery to end-users – as the digital archive grows, access to data will have to be via a Graphical User Interface. The 'day to day' archive will have to be kept on servers. Authorisation controls will have to be rigorously implemented.

Recommendations for management issues:

- The key to Digital Preservation is management of digital resources from their very creation.
- Departments should make decisions early on as to what format records are created in.
- Departmental records produced should be grouped into different categories, by subject or by file formats for ease of preservation.
- Committing resources to be archived is eminently justified, as failure to do so will invariably result in their loss or future costly restoration.

2.6 Metadata

What is metadata?

Metadata can be defined as structured information about resources. It is a reference that is kept alongside a data object to show details about it, such as when it was created, where and by whom. It has been applied in the library and information world for many years, but the term has come to be particularly abundant in the digital arena.

Examples of different types of metadata

- A computer manual is an example of analogue metadata – information about information.
- MARC (stands for *Machine Readable Cataloguing*) encoding standard, in the library-cataloguing domain.
- Encoded Archival Description - In most cases, metadata will fit into standardised structures such as MARC (Machine Readable Cataloguing) or EAD.
- Dublin Core (*see below*).
- Educational resources- online- indexing large amounts of academic material- here the IMS standard for metadata is used [18].

Preservation metadata

A digital object, as mentioned earlier on in the chapter, has a *life cycle*. It is vital that documentation about the life cycle should be noted throughout – where it was created, when, by whom, what changes occurred to the documents during preservation, and who is authorised to access the object once it is stored. These details are needed both from a management perspective but also from a preservation perspective, taking into account what future users trying to access a digital document might need to know. Preservation metadata can be stored alongside the object or embedded within it.

"Regardless of which particular strategy is adopted, long term preservation will depend upon the generation and maintenance of data that describe the digital information being preserved and to enable its interpretation." Cedars, 2001 [19]

Preservation metadata is needed for Digital Preservation because it,

Identifies the record.
Determines who created it.
Details the content of the digital object.
Puts the record in a context and links it to others.
Provides technical details about the object.

[18] IMS Learning Resource metadata specification Available at: http://www.imsproject.org/metadata/
[19] Cedars Guide to Preservation Metadata Available at: http://www.leeds.ac.uk/cedars/guideto/metadata/metadataguide.pdf

Provides knowledge about the bit stream and how it was created, so as to be able to render it when accessed.
Proves Legal admissibility if there is sufficient and well documented metadata about a resource's history.

Preservation metadata is divided into 3 main parts:

Administrative	To support the administrative process of managing an archive; this will also support the *legal admissibility* of a digital object.
Resource discovery	To enable a user to locate a preserved object. It may well be that metadata already exists for objects, such as from a digitisation process or bibliographic catalogue entries.
Technical Preservation	Describes the technical processes taken to produce a digital object. This will record the technical specifics of the digital object- resource type, format, and size. Details of the specific application environment such as the configuration details may also be included. Pointers to details about each piece of software archived separately will also have to be incorporated. In some cases, if the object itself becomes inaccessible due to obsolescence, the technical metadata will be invaluable in order to interpret exactly how the object functioned, and it may be possible to recreate it from scratch.

A standard for preservation metadata

Although they satisfy the role of resource discovery metadata, the 15 Dublin Core fields do not hold the full range of information needed for the preservation of a digital object. The Digital Preservation community now needs an accepted de facto standard to which all digital objects can adhere.
The development of the OAIS Model (see Chapter 3) has greatly contributed to this preservation metadata, as have the efforts of the Cedars project and other institutions such as National library of Australia, OCLC/ RLG, Nedlib and Cedars. All of these metadata schemas are loosely based on the OAIS model. It would be ideal if all institutions had interoperable and exchangeable preservation metadata, which also tied into other metadata created from digital information process. The preservation community is near to reaching agreement on a common standard, the use of which will enable all digital archives to exchange information.

Standards are needed for interoperability- content creators should be encouraged not to create their own metadata as such but to follow emerging standards:

Dublin Core

Dublin Core for interoperable standards for web based material[20]. This is the open global standard set of elements for describing information resources. The initiative was agreed upon by a diverse group of stakeholders who came up with 15 resource- discovery metadata fields. It is expressible in MARC, HTML, or XML. For more information, see www.dublincore.org

RDF

This is another metadata framework for web-based resources. The system aims to make digital resources interoperable via metadata linked to XML.
For more on RDF, see http://www.w3.org/RDF/ [accessed: 06 May 2002]

METS

This supports libraries' metadata in digital object management and creation, and is to be extended to support metadata XML Schemas. The scheme focuses on descriptive and administrative metadata. It doesn't as yet encompass preservation metadata but it is hoped that it will do so in the near future.
For more on METS, see http://www.loc.gov/standards/mets/METSOverview.html [accessed: 06 May 2002]

Generating preservation metadata

This is perhaps the most challenging aspect of digital preservation. Asking creators of digital information to populate metadata fields, even at the simple Dublin Core level can be demanding, and the incentive lies in preserving their material in perpetuity. To record the full range of complicated preservation metadata is even harder.

The onus is thus perhaps on the archives and the electronic repository to add the additional technical metadata fields.
However, data providers to an archive should, in as much as is possible, have a vested interest in preserving their material and providing simple metadata.
Authors can also generate their own metadata for networked materials, from the UKOLN site- this can be found at http://www.ukoln.ac.uk/metadata/dcdot/ and will automatically retrieve DC metadata from a URL.
For this project, the OCLC service (see Chapter 3) will also be able to generate relevant metadata.

[20] See http://www.dublincore.org/

Recommendations for metadata:

- Metadata must be recorded during file creation.
- Creators of electronic information should be persuaded that it is in the long-term interest of the material that it is accompanied by metadata.
- Metadata schemas followed should be interoperable with internationally recognised schemes.
- Dublin Core metadata should be entered into all university-related websites.

Read more on metadata:

Cedars has produced an excellent schema based mainly on the OAIS model- it is extremely comprehensive, and concentrates on storing the technical digital information of the digital object, less so on its retrieval http://www.leeds.ac.uk/cedars/documents/Metadata/cedars.html

The RLG have produced a Metadata schema - it is more concerned with digitised images, but gives a good example of how such a scheme can be stored in an XML template
http://www.rlg.org/preserv/presmeta.html

The National Library of Australia have also produced their own schema- it is interesting to compare it to the CEDARS version.
http://www.nla.gov.au/preserve/pmeta.html

The OCLC and RLG have collaborated on a project to produce a preservation metadata scheme. They have also produced a very interesting document which compares the above schemas; 'Preservation Metadata for Digital Objects: A Review of the State of the Art', which can be found on the site.
http://www.oclc.org/digitalpreservation/

Chapter 3

Practical approaches to Digital Preservation

The previous chapter focused on preparation of the object for preservation; this chapter will focus on the preservation process itself.

3.1 Methods of Preservation

There are a number of approaches to solving the issues of digital preservation. However, the case remains that these techniques, with the exception of a few institutions, have not stood the full test of time and it is still debatable which is the best method to embark on carrying out. It is increasingly apparent that Digital Preservation projects are choosing to combine these techniques to suit their own institutional needs.

It is worth briefly clarifying these approaches and providing links to more in-depth studies, of which there are a lot.

Migration

This is the process of keeping files accessible by periodically transferring or copying data onto a newer, more technologically advanced hardware or software platform, while preserving the integrity of the digital information during the copying process.

There are 2 forms of migration-
1) That of *medium migration* of the storage medium that objects are stored on, for example 3 ½" floppy disk to CD-ROM, or CD-ROM to magnetic tape.
2) The other form is *file format migration* whereupon the object is migrated either when a file format becomes defunct, or to a more stable, preservable format.

Examples of file format migration:

Word files	→	Printed paper documents
Word 2.	→	Word 97-2000 File
Jpeg Image files	→	TIFF files
Windows 3.1	→	Windows 95

Some migration processes are facilitated if the file formats are 'backwardly compatible'. This is clear in the case of Word documents – newer versions of the software are able to read files created on older versions. This does not mean to say however that a version of Word in 20 years time will necessarily read the documents we are creating today- the pace of change is very rapid.

An important factor to consider when migrating files is that digital data can easily be *altered* in the process. Migration has been carried out on large scale in libraries, often at the cost a lot of the time of the content.
Frequency of migration is up to the archive administrator. Most studies give the figure of 5 years for migrating file formats.

Problems with this approach:

- Files will change in the migration process, thus quality control will need to be carried out on each file batch.
- One option that is facilitated by migration is conversion of records to appropriate data *standards*. Much of the 'look and feel' of the digital object my well be lost in the process, and its overall integrity is challenged as many of the hyperlinks cannot be maintained.
- It must also be appreciated that in large archives, media migration may take so long that by the time the administrator gets to the end, the first batch of files may have to be migrated again!

Read more on Migration:

A comprehensive and clear study by a new project in the Netherlands, which aims to provide evidence from case studies into Digital Preservation.
http://www.digitaleduurzaamheid.nl/bibliotheek/Migration.pdf

Hedstrom, Margaret & Lampe, Clifford (2001) Emulation vs. Migration- Do users care? http://www.rlg.org/preserv/diginews/diginews5-6.html#feature1

Wheatley, P (2000) Migration – a Camileon discussion paper
http://www.ariadne.ac.uk/issue29/camileon/

Refreshment

This is a process similar to migration, but it relates to storage mediums. It involves copying data onto a newer storage medium, before the old one deteriorates. It cannot be seen as a long-term preservation solution, as it merely safeguards against the deterioration of storage materials, and not file formats.

Emulation

This involves creating programmes for computers so that old software and old data can be read. The emulators enable up-to-date computers to mimic old ones. This is the ability to preserve the 'look and feel' of an object so that it can be read in the way that it was originally created.

Problems with this approach:

- Depth of emulation- one would have to emulate right down to PC specifics- graphics cards, exact processor, each version of the software, even the monitor.
- The emulators themselves might become defunct- eventually one would have to have emulators for the emulator's original programmes!
- It requires advanced technical skills.
- It might be a solution for 'one off' applications such as recreating computer games or old web browsers.
- It could produce copyright issues- how does one gain access to re-running software in the future that is patented?

Read more on emulation:

Jeff Rothenburg's definitive study on emulation, *Avoiding technical quicksand: finding a viable technical foundation for Digital Preservation* advocates emulation as the viable solution to Digital Preservation.
http://www.clir.org/pubs/reports/rothenberg/contents.html [accessed: 06 May 2002]

Stewart Granger's *Emulation as a Digital Preservation strategy* discusses the practical advantages and disadvantages of actually implementing emulation.
http://www.dlib.org/dlib/october00/granger/10granger.html

Lynch, Clifford (1999) Canonicalisation: a fundamental tool to facilitate preservation and management of digital material.
http://www.dlib.org/dlib/september99/09lynch.html

CAMiLEON Project was set up to research emulation as a viable Digital Preservation strategy
http://www.si.umich.edu/CAMILEON/

Cedars' 'encapsulation' approach

Cedars assume that the actual process of migrating files will certainly result in losses of the original properties of the digital object[21]. In their project recommendations, they state that it is safer to preserve the simple *byte stream* of the object separately from the medium it was created on. Thus if the software to run it becomes obsolete, newer software can be applied to the byte stream to make it readable.

This method is also known as **'Migration on Request'**. When a user requests a digital object from the archive, a migration of the *byte stream* is carried out automatically. This, in effect, separates the content from its format. Thus the object is more authentic, as it is always stored in the byte stream in which it was created.

[21] Wheatley, P *Migration –a Camileon discussion paper* Available at: http://www.personal.leeds.ac.uk/~issprw/camileon/migration.htm

This is to some extent the process of combining an emulation approach with migration, as one has to preserve a tool with the digital object to migrate it to its original platform. The tool referenced in the metadata could either be a software specification or an emulator to mimic the hardware or software environment.
This goes against the notion that digital objects should be continually migrated to ensure accessibility; they can lie 'dormant', so to speak, as *byte stream*s, until they are needed and then software can act on them.

Underlying abstract form

When converting a file to a byte stream, it is vital that when it is accessed, the file can be re-converted into a readable format, in other words, a reversal of the Ingest process. An underlying abstract form (UAF) will describe a file system so that it can be correctly broken down into its folders and files so that the byte stream can be easily converted back to its original form. In order to do so, it has to be decided at the Ingest stage, what the 'significant properties' of the file are, so that when it is accessed, the original contents and the way the file was put together can be interpreted. This abstract representation, or UAF is *separate* from the software that it was created on.
For a CD-ROM, the UAF would be its file tree. In a web page, the highest-level entry point should be noted, index.html for example. In more complex objects, it will however be harder to ascertain what the UAF is. Nonetheless, when the UAF for each file type has been decided upon, it can be re-used when archiving the *same* file type again.

Read more on the Cedars approach:

Paul Wheatley explains this in full at:
http://www.personal.leeds.ac.uk/~issprw/camileon/migration.htm

Rothenburg's article also touches on this in his paper, 'Avoiding Technological Quicksand: finding a viable technical foundation for digital preservation', Council on Library and Information Resources ***Available from:*** http://www.clir.org/pubs/reports/rothenberg/research.html#summary

Strict life-cycle control- document format specification

There is a trend in the records management discipline to urge members of institutions to create documents in standard or open formats, which are easier to preserve. As a consequence, it is much easier for archivists to manage documents, records and digital objects right from their very creation.
For example, marking-up word documents [22], populating non-proprietary databases, authoring websites under certain rigours, such as strict HTML, which can then be transferred to XML. This is, however, much easier to do in a strict records management environment with a published policy of retention schedules and a clear knowledge of internally produced records.

Problems with this approach:

- Stipulating a specific file format is harder to carry out in a research environment where a wide range of digital materials are produced and have to be preserved.
- Libraries cannot make demands on publishers to keep to certain rigorous file formats.
- Ideally any digital archive should be equipped to accept a wide range of file formats. Preservation of the byte stream enables this.

Digital archaeology

This is a method of retrieving data from a file format, the software for which no longer exists. It is a process that requires skilled computer technicians to re-build the original environment in order to access the file.

It is not a particularly fail-safe method. An institution should look to preserving digital objects long before their carrier becomes defunct. However, if objects are to be migrated, the original object should be kept, even if it cannot be read today, as there may well be devices to read it in the future.

Conversion of digital materials to analogue/microfilm

One obvious solution is to simply print digital objects to paper. Paper is a medium that has proved robust and, if kept in the right condition, can remain accessible for years.

Another option is microfilm. Microfilm is a very stable and long lasting medium. But this will not be a solution for more complex digital data. The functionality of a multimedia file would not be retained. Integrity would be lost, and some legal issues may well be at stake - the object could not claim to be an exact copy of the original.

Preservation vs. access

Some projects work specifically with an emphasis to preserve documents. Records management projects place an emphasis on accessibility- the Australian Victorian Electronic Records project[23] is a good example of easily accessible, but preserved records.

The Cedars' project methodology, however, is that '*long-term preservation storage should be kept distinct from access'*, Cedars[24].

The preservation process of a digital object may well be kept separate from the process that is used to make it accessible. For example, objects could be kept off-line in a 'dark vault' and a copy of them can be kept on line for users to access it. The master or 'preservation-priority' version will be the copy kept in the 'dark vault'.

File functionality

An important part of Digital Preservation is the decision taken as to how much of the digital object one wants to retain. The premise being that there is a direct correlation between simplicity of format and long-term preservation.

---

[22] See http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnword2k/html/odc_expwordtoxml.asp

[23] VER- See http://www.prov.vic.gov.au/vers/welcome.htm

[24] Russell, K (1999) Digital Preservation : ensuring access to materials in the future Available at: http://www.leeds.ac.uk/cedars/Chapter.htm

If a file needs to be kept 'forever', then some severe strategies may be in store, such as transferring all files to a simple ASCII format and removing all functionality in terms of multimedia, in effect losing its 'look and feel'. Again by losing these vital elements, there may be legal issues raised. Such alteration should surely harm the authenticity of an institutional record. The digital preservation community is just embarking on research to preserve more complicated file formats.

3.2 Storage mediums

It is a feature of a digital archive that files should be written to a safe and secure medium.
Digital archives should have high storage capacity, a life expectancy of at least 20 years, and hold an established reputation.

Magnetic disk

Hard drives: Data stored on magnetic disk can be recorded and erased any number of times. This rewritable quality may make the media open to modification.
Floppy disks: Although these are not considered fail-safe, they remain the most widely used storage medium for individuals. Almost everybody has had trouble with storing files on floppy disks, owing to their at times unstable nature. More worrying however is the pace at which disks change- there are already obsolete formats for magnetic disks; the readers for disks other than 3½" floppy are becoming increasingly hard to find.

Optical disk

DVD, CD-ROM, and WORM: These are the most recent of storage media types and are very popular due to low price and high storage capacity. It is claimed by manufacturers that optical disks have very long life spans- a CD is meant to last for 100 years. It still remains to be seen however if CD-ROM is as robust as tape, given that it has only been in use for the last 20 years.

Magnetic tape

Digital Linear Tape, Longitudinal, DAT tape, and Helical:
Magnetic tapes are still the most widely used medium for long-term back up, and they are considered very robust. Magnetic Tape is inexpensive and can hold very large amounts of data.
These are used more for 'back-up' purposes rather than day-to-day access- access times can be slow; it is sequential access and not random access like a disk.

Handling of storage media:

All media should be stored in stable, dust free environments, with controlled temperature and low humidity. Mediums should be refreshed, especially optical media, every few years.

Each media has its own shortcomings in terms of long-term reliability:

- Magnetic media is subject to deterioration and oxidation [25], and should be kept away from magnetic fields.
- Magneto-optical mediums such as WORM are susceptible to temperature change.
- CDs are subject to deterioration due to their coating as well as scratching
- Drivers in order to read these various portable mediums are also under risk of becoming defunct.

Archiving websites and databases

WebPages

Dynamic web pages pose a problem for potential preservation, not least because they change so rapidly. One could capture the database itself, at periodic intervals, every time it is updated, e.g. at the same time everyday.

Web snapshot- this is a method of preserving a website so that it preserves in total its functionality, yet it can run in a totally different computing environment, and can be preserved off line. It has to be decided, upon capture, to what level of depth one will want to record all the external links, bearing in mind that these sites wont be part of the snapshot and may change in time. Two excellent examples of this are the PANDORA[26] project at NLA and the web archive[27] in the US. Web plug-ins, non-standard HTML tags and complicated programming files may pose a problem, and in some cases these may have to be excluded, thus losing the functionality of the website[28].

Databases

A current solution is to extract simple flat file datasets from the database – that would display the raw data. The database would then have to be rebuilt in order to read the files. However, this will give the user neither the 'look and feel' of how the database operated, nor how it interacted with the initial creators and users.

Alongside the extracted datasets, one can also record the exact 'Representation Information' (*see Part 3.4*) that will enable a computer technician at a much later date in the future to recreate the original database and re-run the stored datasets.

Another solution would be to simply migrate the database application at periodic intervals to a more up-to-date version – e.g. Oracle 7 to Oracle 9i DB. This is time-consuming and costly but might ensure short to medium term survival of the data, and enable the user to view data in its original environment.

There are various institutions in the UK that store these vast data sets- UKNDAD, Essex Data Library.

---

[25] Ross, S and Gow (1999), *A Digital Archaeology: rescuing neglected and damage resources* Available at: http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2.pdf

[26] Pandora project- see www.nla.gov.au/pandora

[27] The Internet Archive – see www.archive.org

[28] See Kenney, A et al, 'Preservation Management for web resources ; virtual remote control in Cornell's Project Prism', Dlib Magazine Jan 2002 Available at: http://www.dlib.org/dlib/january02/kenney/01kenney.html

3.3 OCLC Web Document Digital Archive

This is a working example of how a working model can be applied to preserve digital data.
It was fortuitous that at the beginning of the 12 month project, EUL was asked to participate in OCLC's new pilot project; "Web Document Digital Archive Pilot" (WDDA). As part of their new Digital Content Management and Preservation Division, the WDDA project seeks to develop a service, otherwise known as a 'toolkit', which will utilise web-based tools to identify, capture, and above all, provide long-term retention and access to digital objects.

In their words, *"It will offer a sustainable service to provide long-term access to web documents. The service will fill libraries' basic needs for identification, selection, capture, description, preservation and access to documents that would not be accessible in the future otherwise." OCLC, 2001* [29].

The term 'web document' is loosely defined as a digital document with defined boundaries, perhaps more analogous to a print document than a website. For the purposes of the pilot, the digital objects will mainly be in the public domain, and freely available on the web; this may avoid any initial copyright hurdles.

Preservation Metadata
The project places great emphasis on the value of metadata in digital preservation, and a large part of the service will be dedicated to automatically producing metadata from an online object, thus avoiding the somewhat time-consuming manual entry of metadata.
The *Harvester* is based on the CORC [30] interface, whereby a URL is entered into the interface, and metadata is immediately generated. The CORC metadata will be expanded however into preservation metadata. Technical metadata also needs to be harvested – it will be interesting to see how this important aspect of preservation metadata will be harvested from web documents.

File Integrity and Preservation
In terms of long-term storage, WDDA will provide an offsite repository, and files can be stored on the OCLC server; back up copies can also be kept in a repository at Edinburgh.
The archive promises a Format Protection Service, which provides management of technological change to file formats in order to ensure continued access in this changing technological environment. This will include migration of formats and their functions.

File types
File types supported will be HTML, PDF, JPEG, GIF, BMP, TIFF, and ASCII text.
Although it will support a variety of content and data formats, dynamic data does not play a part in the pilot as of yet.

[29] See http://www.oclc.org/oclc/press/20010717b.shtm
[30] See http://www.oclc.org/corc/

Maps to OAIS functions

The archive is based on the OAIS model (*see Section 3.4)*, which will be interoperable with other digital repositories world wide, in particular any repository here at Edinburgh University. It emulates its major functions such as Capture/Harvest, Ingest, Data Management, and Access.

There will also be a general administration module: reports, access authorities, and retention schedules. This will deal with any IPR issues.

Security of data

Files will be placed on the OCLC server and staff will regularly carry out checksums and virus checks.

Access

Again this follows the OAIS model, and provides strict access controls. Access by users will be via a web browser from OCLC or, in some circumstances, our local archive.

The service will also provide the ability to search for and provide access to this data.

A managed system

As well as being a continuous round-the-clock managed service, WDDA will provide an integrated workflow for individual libraries. One can create, for example, different content groups for any number of documents so that duplication is reduced, in terms of managing them and who can access what objects. Individual access authorisation (a very important concern in a digital archive) can be easily allocated and deleted.

Each object will be designated a service level. Records can be 'placed' either on the OCLC server or onsite in our archive.

How this ties into our project

This pilot project has been very useful, both in defining metadata fields for preservation and in defining our workflow for the digital repository. It will ensure certain features, such as strict authority control, version control, and dividing the groups of records into certain groups for ease of management. It might well ensue that we will use the OCLC WDDA to archive a select few objects - a 'dark archive', and a repository here at Edinburgh for documents that are for day-to-day retrieval, yet still archived for the purposes of permanent preservation.

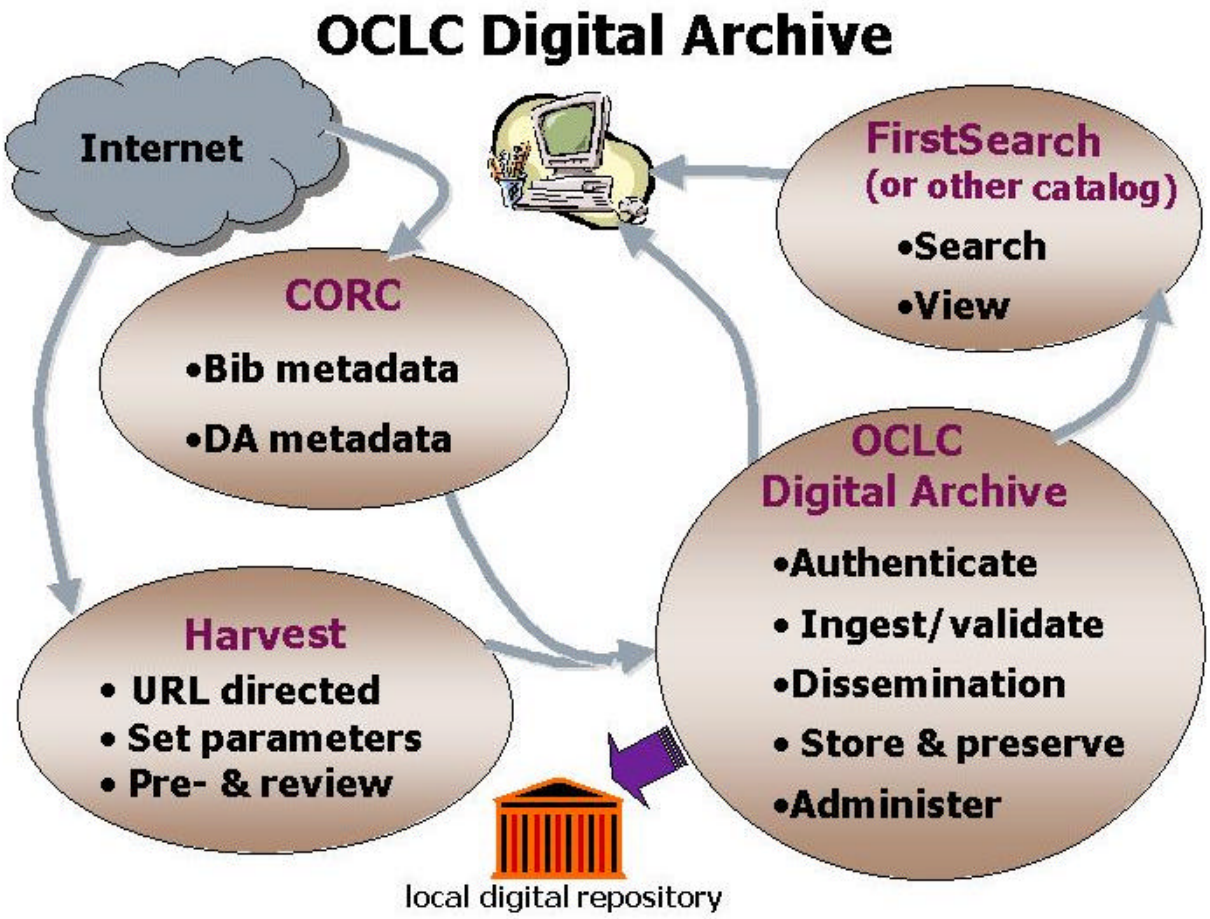


Fig 1- Image of OCLC Digital Archive- Courtesy of OCLC

3.4 OAIS model- a de facto process model for digital archives.

A model that has had a defining influence in development of digital preservation methods is the OAIS model, developed by US Consultative Committee for Space Data systems. The OAIS model (Open Archival Information System) is a conceptual framework for an archival system dedicated to preserving and maintaining access to digital information.

It is important to understand that the OAIS model[31] is applicable to any archive. A number of prominent institutions have based their archiving techniques on the model: OCLC, RLG, National Library of Australia, British Library and the Cedars and Nedlib projects.

The model is not prescriptive, but it offers a series of workflows that can be followed in setting up a digital archive. Although it is very large and complex, it has been invaluable to this project in detailing exactly what processes are needed to set up a digital archive.
The model aims to act as a guide and to create a set of terminologies, which can thus be shared across all digital archives. It is now an ISO standard.

For the purposes of this project, I have decided to follow a very simple version of OAIS, mainly because it will enable the future archive to be consistent with OAIS terminology and interoperable with other archives.

Implementation of OAIS will:

- Make clear standards needed for successful Digital Preservation.
- Create a clear set of terminologies that can be understood by all the preservation community.
- Clarify the procedures to set up a trusted archival service.
- Raise awareness of the Digital Preservation issue.

Nedlib

The Nedlib project has mapped the model very successfully to its own DSEP digital library system[32]. Many of their processes are outside the scope of this pilot archive, but the Nedlib literature has been useful when attempting to extract a more simple set of processes.
The DSEP appreciates that one should build on processes already in place in the library; Edinburgh University Library has very sophisticated and well established methods for dealing with information delivery for end users, as well as circulation and cataloguing.

[31] OAIS Model Available at: http://ssdoo.gsfc.nasa.gov/nost/isoas/
[32] Nedlib- See www.kb.nl/coop/nedlib/

A full application of the model across institutions is questionable, as it will be time-consuming and expensive. However I plan to demonstrate that it can be emulated on a small scale.

Read more on OAIS:
RLG has a good resource page on OAIS http://www.rlg.org/longterm/oais.html See ERPANET training event on OAIS http://www.erpanet.org/www/products/copenhagen/copenhagen.htm

The main processes of the OAIS model are:

1. **Ingest**
2. **Archival Storage**
3. **Data Management**
4. **Access**
5. **Administration**
6. **Preservation**
7. **'Delivery and Capture'**
8. **'Package and Delivery'**

Objects move around this managed system in what are called Information Packages, which change their status depending on where they are in the archive. There are 3 types of packages:

Submission Information Package **(SIP)**

This is the digital object package, which is sent to the archive by the information producer. In some cases, it may well not arrive with adequate metadata.

Archival Information Package **(AIP)**

This contains all the information needed in order to preserve the digital object. All the relevant metadata will be added, and technical specifications to render the object.

Dissemination Information Package **(DIP)**

There is little need for the user to have access to all the complex technical metadata stored in the AIP. The package that the user will receive is the digital object and some of its metadata.

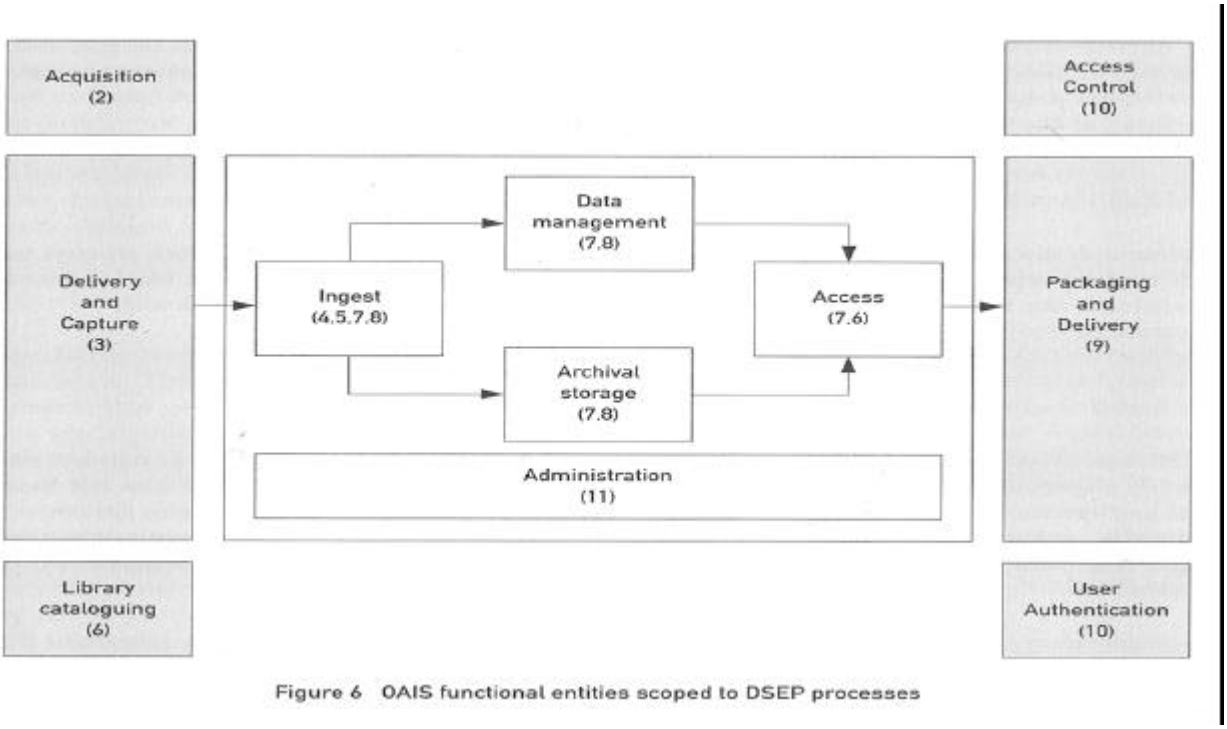


Figure 6 OAIS functional entities scoped to DSEP processes

Fig 2- OAIS functions- courtesy of Nedlib

The processes themselves are complicated; it is worth going through each one in turn, to comment on how they map it to the EUL archive.

Delivery & capture	**SIP**
This is the process of obtaining the record and its associated metadata. Conversion to a more standard file format will take place here. Ideally, it will consist of an interface, which receives the digital object, and can be linked in to other library processes, such as the library catalogue or the acquisitions module. This process also deals with the publishers or users of the digital archive, and stipulates how objects should be delivered, for example, depositing the files via a CD-ROM or using *FTP* to transfer them Here, the Submission Information Package might consist of:	

-The data object itself
-Some descriptive metadata
-Description of the Software

Ingest	SIP

This process receives the data object prepared by *Delivery and Capture*. It then sends the object off for permanent storage.
Any administrative information is sent to Data Management.
The *Underlying Abstract Form* is established.
A Unique ID is assigned.

The archival Information package will consist of:
-Data Object
-Descriptive metadata
-Structural metadata (index of file tree)
-Technical metadata (this will invariably be a link to the software specification, stored elsewhere in the archive)
-Software (optional)
-Bibliographic record

Archival storage	AIP

The purpose of this process is to ensure that the *byte stream* stays in tact.
Nedlib suggests that this part cannot 'act' on anything; it's just a 'meaningless byte stream' and other parts such as data management or dissemination can act on the metadata.

The process however will be responsible for medium migration and back up, error checking etc.
Note that the only migration that takes place is media migration- there is no need for file migration. Files cannot be updated either.

Data Management	AIP

This part of the process manages the metadata itself and the databases that store it.
Updating of metadata poses a problem in the digital archive. Metadata should not be duplicated, as it will be hard to co-ordinate these updates.
Data management can also generate reports about deposit holdings, usage statistics etc. This too is considered metadata, and should be preserved.

Administration	AIP
This deals with the overall operation of the archive system. This can involve negotiating with publishers, developing and checking the day-to-day running of the archive. System performance is monitored.	

Preservation	AIP
This is responsible for the long-term preservation of files and carries out any migration process of the media. It monitors whether files are still readable, and anticipates future changes in preservation strategies.	

Access	DIP
This part of the process includes finding aids, registering of library users, access controls. It will interface directly to the archive. 'Access' will deal with requests from users- it will check the administrative metadata to verify the object's access controls, and then deliver the object to the user. In addition, it will notify the user what technical conditions are needed to view the object. It can also be linked to 'Data Management' for usage statistics. The Dissemination package will consist of: The Object Metadata Software to view it Packaging information	

Packaging & Delivery	DIP
This will deliver digital objects to external systems. It will deal with requests from external users. In addition, this process might add indexes to objects, for search ability. Ideally all access rights are dealt with here, and the package is delivered to the user.	

3.4 Practical implementation Pilot Project

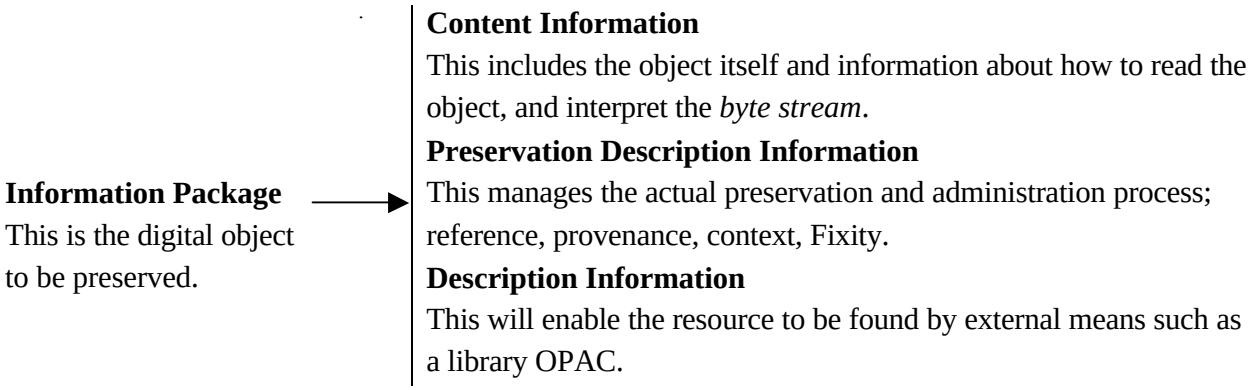
Adapting the OAIS model for Edinburgh University Digital Archive (EUDA)

In order to archive the Calendar and set up our own pilot digital archive, it was necessary to adapt the OAIS model to suit our requirements. The two main OAIS processes were adapted: the metadata schema and work flow model. This chapter details these processes and examines how the Calendar archiving pilot was structured.

3.4.1 Metadata Schema:

As well as presenting a model for a digital archive, the OAIS defines the main metadata fields- these fields are inextricably linked to the model itself and are the 'back bone' of the model.

OAIS breaks down metadata into the following Information Packages:



Overleaf is a table outlining the metadata fields. Note that definitions of these fields, and the XML DTD can be found in Appendix A & B.

Content Information	**Representation Information**	**Structure Information**	Underlying abstract form Description UAF transformer - Platform - Parameters - Render analyse Engines - Output format - Input format
		Semantic Information	Render/Analyse Object
	Primary Digital object		
Preservation Description Information	**Reference**	**Resource Description**	5 DC elements – title, subject, description, contributor, rights, and date.
	Context	**Related Objects**	Related Information Object. - Relationship - Reference
	Provenance	**History**	Reason for creation. Custody history. Change history before archiving. Original technical environments. - Prerequisites - Procedures - Documentation Size of file. Significant properties. Reason for preservation.
		Management History	Ingest process history. Administration history. - action history. Retention period (EUL).
		Rights Management	Negotiation history. Rights information. - copyright statement - date of publication - place of publication - rights warning Actors Actions - permitted by statute - permitted by licence
	Fixity	**Authentication Indicator**	
Packaging Information			
Descriptive Information			

Notes on the Metadata schema:

Archiving software- in order to interpret the byte stream, software will have to be archived within the archive. Instead of archiving the same software with each information package, an archival information class will be allocated to each object:

Archival Information Class - a term coined by OCLC/RLG – as there will no doubt be duplication of data object *types*, it seems logical to assign objects an archival information class, 'type specific metadata', and in the representation information links will be provided to the metadata for each object type. Each information class will also be stored in the archive and there will be pointers to it with the individual information package metadata.

For example, if there were lots of PDFs, then there would be no need to repeat the Representation Information for each one, but instead to draw on a common one.

Storage:
The metadata will be stored in XML files. See Appendix C for outline of DTD.

3.5 Archive structure

Below is a diagram, which shows EUDA.

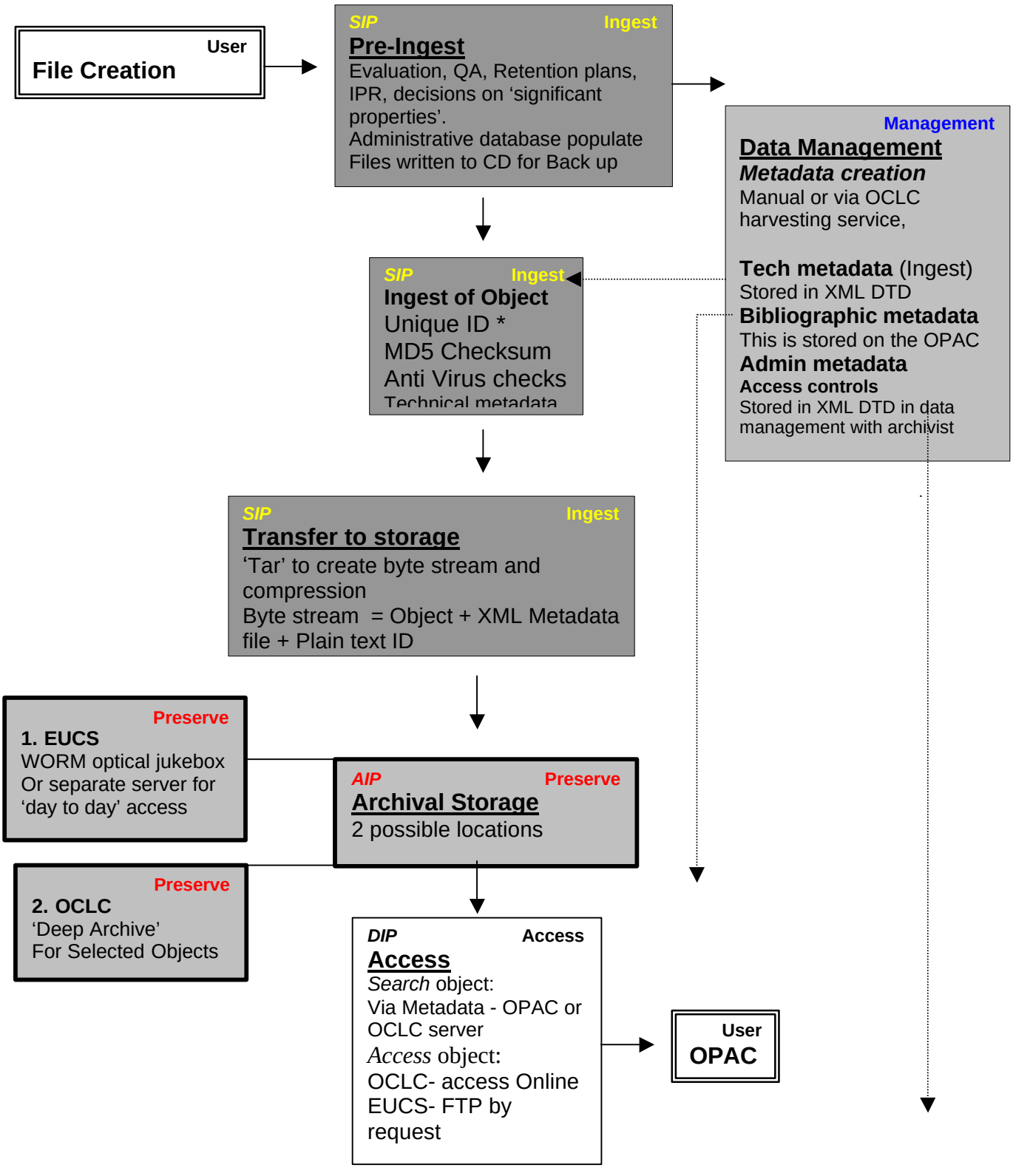
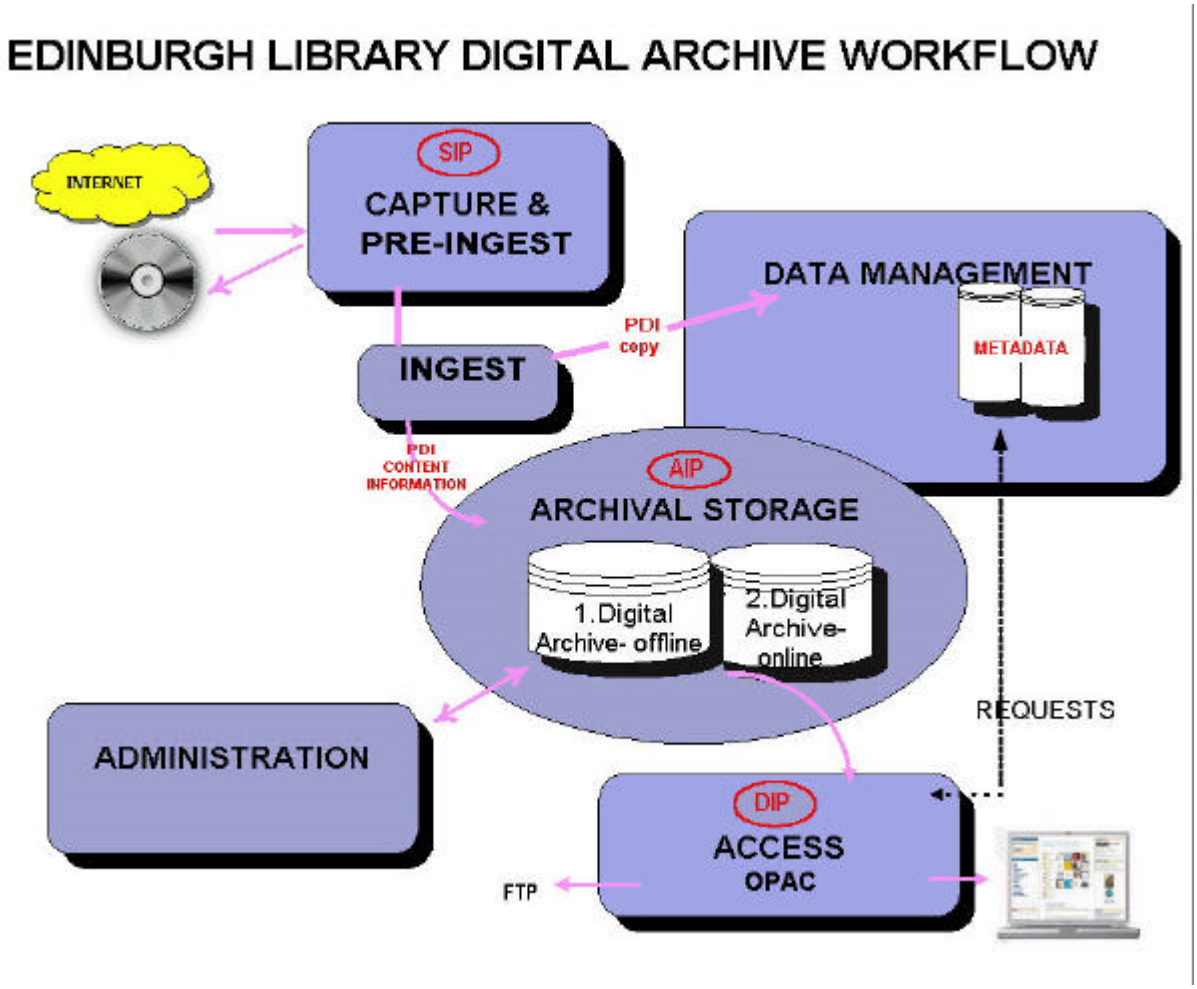


Fig. 3- EUL Digital Preservation Work flow Model



3.5.1 Edinburgh University Digital Archive ~ Individual process explanations:

Although these processes reflect the OAIS model, the processes are taken from the Cedars and Nedlib recommendations. They are based on the assumption that the *byte stream* should be separated from the file format- only on accessing the object from the archive is it transformed into a readable file.

Pre- Ingest	In library terms this is the archive acquisition process. In the first instance of EUDA, there will be no interface; the deposit area is a secure networked space on the archivist's desktop. People can send material to the archive in the shape of CD-ROMS or secure *FTP*. Another option will be to 'harvest' the object from an online source via OCLC harvester. Quality assurance and integrity checks of files will be carried out here; Virus checks and *checksum*s. The object is given its unique ID, and recorded in a local database. The metadata is collected either by harvesting or manually.

	For the technical metadata, Archival Information class descriptions are created (see p.50 of this report). A copy of the actual software the object was created on might be acquired. **Input: Harvesting of metadata, Details of object (ID, checksum) entered into database.**
Ingest	This is where the SIP is on the desktop and is ready to be sent off to the archive. It is transformed into a *byte stream* and packaged up, compressed to be sent to Archival Storage. This will be via 'tar' technology (See Section 3.5.2 below) All the metadata is collected and sent to different locations; the full bibliographic record is sent to the library OPAC. **Input: Data object itself; Technical metadata; Administrative metadata. Copy of the Software to be added; or a link provided to the detailed description the Archival Information Class.** **Descriptive metadata to library OPAC.**
Data Management	This part of the process will manage the metadata. Much of the metadata will be generated by work with OCLC. As the WORM technology disallows updates, metadata corresponding to each file cannot be updated. Therefore separate components of the metadata are kept in different locations, according to their function. Administrative metadata should be entered into a separate database, as it will need to be updated. This will also be archived as data sets at frequent intervals. The WORM archive will store the technical metadata, which will not be updated. Bibliographic metadata will be managed via the library OPAC.
Archival Storage	EUDA storage will be on the EUCS archiving facility; a Maxoptix optical jukebox. This off-line WORM archiving technology (see part 3.5.2) will ensure that items are safe and cannot be altered. No file migration will take place, as the byte stream will be preserved, however it is possible that in the future, EUCS migrates from Optical

	disk to a more up-to-date technology. Details of the software to run file formats will be kept in the accompanying metadata, as well as details of the operating system. As a very secure back up for selected items, the OCLC will be used to preserve objects. Back up will also be via replication of the digital object on a CD-ROM. It might be that when the archive is fully established, further back up copies will be kept on a separate server. It should be noted that the archive would operate independently of any other university process. It will be stored on its own, and it will not be networked.
Administration	This will deal with the ID and *checksum* databases. It will also administrate reviewing of IPR issues, and update *audit trail*s.
Access	The archive is not networked, so access will be restricted. Initially the archive will *not* have a user interface. Objects archived in Edinburgh Digital Archive will be searchable via library OPAC and accessible via FTP. Objects archived at OCLC will be fully searchable and accessible (authorisation permitting). **Output: Digital object via archivist and FTP. Download object from OCLC server.**

3.5.2 Byte stream, Storage and Unique ID

Creation of the byte-stream

As explained in Section 3.1, the archiving method chosen in this project is based on the Cedar's method of extracting a *byte stream* from a digital object.

The best way of creating a byte stream is through the use of a *compression* tool, which produces a file 'on the fly' as it is written to the archive. One would need to use an open format compression tool; the most common one is Zip.

We have however decided to convert files to a byte stream and compress them in the 'tar' format[33].

'tar' is a good format for controlling large amounts of data. Many things in the Edinburgh University Computing Services (EUCS) archive are already packaged with 'tar'. It is usually used on Unix platforms but there are versions available on Windows and it is relatively easy to install.

When an object is retrieved from the archive and decompressed, the file tree is evident and software can be applied to read the byte stream. 'tar' works by gathering together a number of files into a single file to transfer onto another platform. 'tar' will enable the UAF to be identified and we can interpret the original contents and the way the file was put together. On 'decompression', the byte stream can be easily decoded.

N.B. Unlike Zip, 'tar' does not compress on its own, it needs another *compress* or *gzip* command.

Storage- a note on WORM technology

A repository must be able to protect digital objects against technical obsolescence and it must be a trusted storage medium.

The storage chosen by this project will be provided by the already existing EUCS archiving service; Write Once Read Only – WORM. It is very suitable for bulk data storage and is already used as a service offered to university members for personal data storage. There is no intervention on behalf of the administrators of the archive however, and files are written to it in any format, and have been so for the last 20 years. Thus there can be no guarantee that the files can be read in the future.

The project therefore will use the services of the archive, and incorporate further techniques to ensure preservation. Note that a firewall is *not* set up at EUCS for purposes of the archive, as it is accessible only by members of Edinburgh University.

An important point to make is that users will not have the capacity to delete or modify objects once they have been archived. The archivist will be able to alter its access rights externally but cannot physically delete or update an object. With this technology it will be impossible to go back and add more metadata to the object.

[33] For more on tar, See http://www.gnu.org/manual/tar-1.12/tar.html

Therefore administrative and bibliographic metadata has to be kept in a separate place, so that it can be updated when necessary.

Access to the archive

In the first instance access will be via the library only or, if independently archiving an object, via personal FTP accounts.

Again this relates back to the concept of preservation vs. access. Our archive is specifically for preservation. Other projects, namely records management projects such as VERS[34], concentrate on access, and the technology they use allows records and metadata to be updated and easily accessed.

Checksums

Courtesy to the National Archives of Scotland, I have been helped in implementing a database which will register all digital records to provide a *checksum* digital signature, which can be relied upon on to ensure continued long-term preservation.

Checksums are needed to prove that the *byte stream* has not been altered while stored in the digital archive.

Archive back-up procedure

This will be in the form of writing every digital file to CD-ROM before it is sent to the archive, and kept in controlled storage. At a later stage, files could also be kept on a separate server for day-to-day access, and the 'dark vault' at EUCS will be an additional back up.

File size

It is to be expected that very large files will be sent to the repository. The EUCS archive provides each individual with up to 50MB per month, and charges thereafter. For the purposes of a digital archive, capacity will be larger

Version control

In archiving digital objects, we had to decide what version was to be archived, as many digital objects have updates.

In the case of the Calendar, we decided that the 'golden' version would be archived once yearly. A version with additions will be added at a later stage, and stored as a *separate* object. N.B. for the purposes of this project, we archived the Arts faculty section of the Calendar.

Changes/ updates to objects

If changes occur, as they do with web-based material, then due to the WORM technology, one will have to add the change as a *new record.* If the web page is then removed entirely from its host site, an additional metadata note will have to be added to data management, linking it to the previous archived site.

---

[34] Victorian Electronic Records Strategy, See www.prov.vic.gov.au/vers/welcome.htm

Storage Refreshment

The content of the archival information package should not be changed- however the medium on which the object is stored will need to be changed. This will depend on the storage medium itself and knowledge of its life span in the market place. It may be that the archive transfers all the data to another set of CD-ROMS, or DVD, if it is recognised that they will be supported in the foreseen future.

Migrating these large amounts of information will take much management, time and cost.

Note: the archive administrators may decide to change from WORM technology to store objects on a re-recordable media so that metadata can be updated which would greatly facilitate the preservation process.

Unique identification

Each object in the repository must be assigned a unique identifier that never changes. This is the same as a Uniform Resource Name (URN) or a monograph's ISBN number. Each *version* of an object would also have a unique name. The importance of this cannot be underestimated. Any ID system should be designed at the outset so that it will be understandable by the OAIS system and other repositories. This is to unambiguously identify every digital object, even if it is transported to another repository altogether. It will also link the administrative and bibliographic metadata to the digital object.

In addition, each digital object can have a human readable name, which is more easily recognisable and searched. An index of all unique identifiers will have to be stored outwith the archive. As new objects are entered, the index is updated. Note: the long-term survival of the repository is not dependent on this index.

The system chosen is as follows:

Note: it is entirely up to individual departments how they chose to allocate the ID. This is just a recommendation- but each ID has to be superseded with the first 3 fields in order to make it interoperable with other digital archives.

Ed: Mandatory	**Name of repository:** Mandatory	**Department:** Mandatory	**Series/ type of document:**	**Object number:**
	EucsDa stands for EUCS digital archive. If kept on department servers, or tape back up, this will also have to be stipulated.	Abbreviations for departments will follow the central filing system	See table below	

		classification		

Examples:	**Object title**	**Unique ID**
Word document held on departmental server	ME History; Year 2 results; 34	Ed: artsrv0: IAIS: worddocs: 34:
IAIS department web document archived via library		Ed: EucsDa: IAIS: webdocs: 15:

The following document types subdivide digital objects:

Word documents	worddocs
Web objects- HTML files	webdocs
PDF	pdfs
Databases	dbs
Image files	images

Chapter 4

Guidelines and Recommendations

This short guide aims to steer the author of digital material through a series of steps and recommendations in order to facilitate the long-term preservation of files if they are seen fit for preservation.
It is clear that much of the preservation process is during the creation of the digital object.

4.1 Guidelines for archiving

Below is a list of the seven processes to be followed when archiving a digital object. The explanations are below.

Pre-Preservation

1. Retention schedules

What files do you want to keep and for how long?

2. File Format

Which files are easier to preserve?

3. Legal issues

Are there any legal issues to be considered before archiving?

4. Metadata

Start creating the metadata and adding it to the digital object.

Preservation

5. Pre-Ingest

Assemble all the files.

6. Ingest

Send to the package to the archive.

7. Access to object

1. Retention schedules

Assign retention schedules to the object

Before one embarks on creating digital documents, it is worth defining at the outset how long the document is to be retained. If it is considered an Edinburgh University record, it may have to be accessible in 10 years time. In some cases it may have to be preserved 'forever'.

It is worth assigning one of 3 time-spans to the object:

Not preserved
Preserved for defined period
Preserved indefinitely

Once decided, these life spans should help decide to what file format the object is created in. If it is for the long term, then the simplest format should be chosen.

It might help to break records down into manageable groups- Records that share similar characteristics, such as minutes or exam results can be allocated similar retention schedules.

Object Creation

2. File Format

Consider the file format of the digital object.

The more simple the format, the longer it can be preserved.

Here is a list of suitable file formats:

- Rich text format, ASCII
- PDF
- Mark up language- if this is to be a web document then note the following: Dublin core should be added. Where possible, HTML should be either strict HTML or XHTML.

Technical details of the conditions one created should be recorded at this point to add to the metadata later.

Object Creation

3. Legal issues

What legal issues are to be considered?

At this stage it is worth considering the legal implications of archiving your digital object.

Consider the impact of the following legal acts-

- **Data Protection Act** 1998: The record may contain personal data that refers to individuals, thus access to the file may have to be restricted.
  In some cases, it may not be allowed to be stored longer than a certain length of time.
  What access controls will the object need after some length of time in the repository?
- **Freedom of Information Act** 2001: The public may have unrestricted access to certain records. Consider what categories of information may need to be viewed by the public - these records need to remain accessible at all times.
- **Electronic Communications Act** 2000: Does the record contain a *digital signature*? If so, it might be recognisable in a court of law as a legal document, as opposed to its paper counterpart.
  Digital objects should also be free of corruption before they are archived; the files must be readable.
- **Copyright, Design and Patents Act** 1988: Are permissions needed in order to copy this object for preservation purposes? Are you trying to archive an external website? This might not be possible due to the copyright law. Is the software to render the object being archived- permissions may be needed to obtain specifications of the software? Can one migrate the copy to a newer format?
- **Other laws and regulations** that might be applicable in your particular department, such as Health and Safety law, Contract Law, Disability Discrimination Act. It is worth bearing these in mind.

Action: Record all this in the administrative metadata

4. Metadata Creation	
This is the core process of digital preservation, and it is essential that all the above steps be recorded in the metadata. Any recording of metadata at this stage will make it much easier to preserve in the future. Break the metadata up into 3 groups • Administrative • Preservation metadata • Technical metadata Keep an *audit trail* of how the site is captured, any problems encountered, and any files that have been excluded. Document the operating system and application software. Ensure that the HTML specification is recorded. Metadata can be entered into an XML template, available from the digital preservation web site. www.lib.ed.ac.uk/lib/sites/digpres/metadataschema.shtml Web authors In the case of web authoring, pages should contain simple Dublin Core metadata - templates can be found at http://www.ed.ac.uk/help/templates/	**Object Creation**

At this stage you can:

- **Chose to send the file off to the library for permanent archiving,** *or*
- **Follow the guidelines below to preserve a digital object independently, but inform the university archives.**
- **These guidelines are purely** *recommendations* **for the safe preservation of data. In some circumstances it might be easier to send the file to the library for safekeeping.**

5. Pre Ingest	
Assemble all the relevant files on a space on your desktop. For back up procedures, keep a copy of these files separately on a CD-ROM. Carry out quality control and verification of files (e.g. are they all readable, and virus free?). Carry out a *checksum.* Allocate a unique ID. Consider what the 'significant properties' of the file are. Files present should be: 1. Content data object itself. 2. Technical metadata information in XML file. 3. XML file with administrative metadata- to be kept in a separate database. 4. Information to be added to the library OPAC if you want it searchable at a university level, if not then keep any bibliographic information in a separate database. 5. Text file with its unique ID and any *checksum* information.	**Preservation Process**

6. Ingest to archive	
Use ZIP or 'tar' [35] technology to group together the different files as one object. This will ensure the *byte stream* is preserved in perpetuity. If you are compressing the file by any other means, ensure that it is done in a 'lossless' way to ensure that the data is not altered. An account with the EUCS archive is needed for long-term storage. Storage on independent servers, with secure back up is also possible; however regular steps must be carried out to ensure files can be read and the storage medium does not become obsolete. The safest method is storage on EUCS *WORM* Storage. FTP file to EUCS archive. Keep a local database of the unique IDs, and their associated object titles, as well as a database of the *checksum*s.	**Preservation Process**

7. Access	
Access to the object If the object is at EUCS, access to the preserved file will be via your account only. The preserved file will not be online. If the object is stored via the library, it will be searchable via the OPAC; access for the time being however will be via authorised librarians only. In the future we will need to provide a user interface to the archive using a browser.	**Preservation Process**

[35] see http://www.gnu.org/manual/tar/html_mono/tar.html

4.2 Main recommendations

University level:

- A concerted effort should be made to apply Digital Preservation techniques to key Edinburgh University records.
- Digital archiving processes followed should correspond with the OAIS model.
- An electronic records management approach must be adopted – this will enable records to be managed, stored, retrieved, access-controlled and preserved throughout their life cycle.
- A trustworthy record-keeping environment must be created. Retaining their legal admissibility is crucial, and schemas have to be drawn up as to how long personal details should be retained.
- In order to ensure long term and continued access, a corporate policy should be agreed upon by stakeholders so that common standards of both file format and storage media are implemented from the very start.
- A Centralised Co-ordination in all the projects would ensure that the information systems and databases are interoperable. This will be more achievable if the course content information originates from a centralised database.

Legal issues:

- Recent legislative issues have made effective Electronic Records Management a pressing concern; the Data Protection Act of 1998 will clamp down on unnecessary duplication of data. The Freedom of Information Act, 2001 will also have a bearing on effectively managing and making accessible University corporate memory for the long-term.
- The Code of Practice for 'Evidential weight of information stored on electronic management systems' should be applied to any digital archive. The Integrity of data should be checked using *checksums*. Digital signatures could be used to this effect to prove they are the original documents. *Encryption* methods on the other hand, are more complex and are more for use for *exchange* of information over networks. Note that *compression* of files will also alter a file's authenticity and this should be noted in the metadata.
- Any digital archive has to make sure that there are provisions for frequent backup or duplication and stored *off site* disaster recovery. Access authorisations will have to be allocated and recorded.
- Disposal- to complete the electronic document life cycle, disposal must be adequately carried out ' reformatting or overwriting' N.B. *deletion of files* is not the same as destroying them.

File types:

- File Formats must be non-proprietary, rather than tied to a specific vendor.
- A good choice would be to store the information in structured XML files; this application-independent format enables any number of document types to then be re-created from the core data (e.g. MS Word, HTML, PDF), thus ensuring that future generations are not hardware or software dependent.
- Some of the problems of archiving dynamic database information can be overcome by preserving the fundamental data sets as well as a log of changes made to the database. This is in conjunction to recording metadata to provide precise details of how the database was set up.
- Image files must be TIFF if used for long-term storage and access.

For the creator of digital objects:

- Create files in as much as possible in non- proprietary format.
- If they are to be preserved, consider transferring them to alternative formats.
- Standard formats are preferable, although well-documented proprietary formats, such as Microsoft Word documents or Oracle databases, are better than non-documented formats.

Web authoring:

- Web document templates must be used where possible when authoring sites to be located on the university site.
- The university Web Editor has created web Document Templates. These are aimed at university departments and will be made available through the main website. If implemented, on-line university publications will be more customised, which will lead to easier management and preservation of web-based documents. Templates are available at: http://www.ed.ac.uk/help/templates/
- Use of the Dublin Core standard in web pages.
- Use of valid HTML or even more appropriate, XHTML, in order that files can be converted to XML later on.
- For corporate websites, which are to be kept as Edinburgh University records, use of complicated graphics, rollover images should be avoided.
- Check if online documents contain ‘ robot exclusion files’- legally one is not allowed to harvest these files.

Metadata:

- Metadata schemas must be flexible and interoperable with other internationally recognised schemas.
- As much metadata as possible must be recorded when files are created.
- Creators of digital material should be encouraged to include or record as much metadata as possible as this will benefit its long-term preservation.

4.3- Appendices

APPENDIX A

Explanation of metadata fields

Content Information	*The content information comprises the digital object itself (in a byte stream) plus the technical metadata needed to understand the byte stream.*		
Representation Information	*This section contains all the technical information needed to 'understand' the object. Often it will contain links to archival information classes*.* *Cedars notes that if there is not sufficient information, then free-text describing the resource will suffice.*		
Structure Information	*The structure information holds references to different descriptions of archival information class. Parts of this technical metadata will of course be duplicated in the different class descriptions.* *It will provide information as to how to transform the byte stream back to its original humanly readable format.*		
	Underlying abstract form Description	*This is a description of the file system that has been converted into a byte stream. For example, CD-ROM file tree or a web site hierarchical files.* *This description could be saved and applied to similar objects.*	
	Uaf Transformer	*Description of how the object can be transformed from its byte stream back into its UAF- e.g. unTar, unZip. This should also include the location of this transformer (URL, manual) (OCLC/RLG).*	
	Platform	*The specific hardware platform on which the object was created.*	
	Render Analyse Engines	*The software used to carry out the needed to carry out the transformation process.*	
	Input Format	*Describes the format of the object that the transformer acts on.*	
	Output Format	*Describes what will be produced by transforming the object.*	
	Render/Analyse Objects		
Semantic Information	*The semantic information pertains more to the individual object and, unlike the structural information doesn't apply to objects of the same archival class.*		

	Render/Analyse Objects	*This element will list the software and hardware platfoms that are needed to understand the specific data object. This has the same sub-elements as structural Render/Analyse Objects, however it differs in that policies of migration are listed here.*	
Primary Digital object			*This is where the object is logically stored; the unique ID of the byte stream links it within the information package*
Preservation Description Information	*This is information needed for the preservation processes of the object, and stores administrative and management information.*		
Reference	*This contains bibliographic information (although the object will be searched via the library OPAC with more extensive bibliographic fields) and information about other metadata and accompanying manuals.*		
Resource Description	*5 DC elements: Dctitle, Dcsubject, Dcdescription, Dccontributor, Dcrights, Dcdate* *N.B. It still needs to be verified what the Dcdate pertains to- date of creation, ingest. There will have to be qualifiers in this field.*		
Context	*This is needed to indicate what other objects are related to other objects in the archive- updates etc.*		
Related objects		**Relationship** **Reference**	
Provenance	*Crucial to the archive is knowledge of the provenance of the object, why and how it came to be ingested into the archive. This section will support the object's legal admissibility.*		
History		Reason for creation Custody History	
		Change History Before Archiving	
		Original Technical Environments	*This describes the software and hardware environment the object was created in (operating system details).*
		Prerequisites	*Precise soft/hard ware details.*
		Procedures	*How to run the software.*

		Documentation	*Any manuals etc that are associated with the object.*
		Size of file	
		Significant properties (OCLC/RLG)	
		Reason for preservation	
Management History	*This is crucial in describing changes to the object from pre-ingest onwards, under the responsibility of the archive.*		
		Ingest process history	
		Administration History	*What happened to the object after ingest.*
		Action History	*What was done to change the object, migration etc.*
		Retention Period (EUL)	*How long the object is to be retained in the archive.*
			What the files are comprised of (3 HTML files, 2 Gifs). This determines how much of the original object needs to be kept- just the intellectual content or its whole functional 'look and feel'.
Rights M/ment		Negotiation History	*Any negotiation regarding rights, before ingest.*
		Rights Information	
		Copyright Statement	
		Date Of Publication	
		Place Of Publication	
		Rights Warning	
		Actors	*Authorised access permissions names.*
		Actions	
		Permitted By Statute	*A note of standard fair dealing actions on the DO.*
		Permitted By License	*Terms of the license.*
Fixity			
Authentication Indicator	*To prove the authenticity of the object, a note of the checksum value will be made here.*		
Packaging	*This logically relates the 2 elements, CDO and PDI together e.g. file structure of a CD-ROM file.*		

Descriptive	Full set of fields available via library OPAC; MARC Record.

APPENDIX B

An XML template (and DTD) can be easily populated with the above metadata.
This template is courtesy of Cedars:

```
<?xml version="1.0"?>
<!DOCTYPE informationPackage SYSTEM "cedars.dtd">

<informationPackage>
<preservationDescriptionInformation>

<referenceInformation>
  <resourceDescription>
        <DCtitle> </DCtitle>
        <DCsubject> </DCsubject>
        <DCdescription> </DCdescription>
        <DCcontributor> </DCcontributor>
        <DCrights> </DCrights>
  </resourceDescription>
</referenceInformation>

<contextInformation>
  <relatedinformationobject>
        <relationship> </relationship>
        <reference> </reference>
  </relatedinformationobject>
</contextInformation>

<provenanceInformation>
  <historyOfOrigin>
        <reasonForCreation> </reasonForCreation>
        <custodyHistory> </custodyHistory>
        <changeHistoryBeforeArchiving>
        </changeHistoryBeforeArchiving>
   <originalTechnicalEnvironments>
        <prerequisites> </prerequisites>
        <procedures> </procedures>
        <documentation> </documentation>
        <SizeOfFile> </SizeOfFile>
```

```
<SignificantProperties>
</SignificantProperties>
  </originalTechnicalEnvironments>
  <reasonForPreservation>  </reasonForPreservation>
 </historyOfOrigin>


<managementHistory>
        <ingestProcessHistory> </ingestProcessHistory>
        <administrationHistory> </administrationHistory>
        <actionHistory> </actionHistory>
        <retentionPeriod> </retentionPeriod>
</managementHistory>

<rightsManagement>
        <negotiationHistory> </negotiationHistory>
 <rightsInformation>
        <copyrightStatement>
          <nameOfPublisher> </nameOfPublisher>
          <dateOfPublication> </dateOfPublication>
          <placeOfPublication> </placeOfPublication>
          <rightsWarning> </rightsWarning>
        <actors>
          <permittedByStatute> </permittedByStatute>
        </actors>
        <rightsWarning> </rightsWarning>
        </copyrightStatement>
 </rightsInformation>
</rightsManagement>
</provenanceInformation>


<fixityInformation>
        <authenticationIndicator> </authenticationIndicator>
</fixityInformation>

<contentInformation>
<representationInformation>
 <structureInformation>
        <uafDescription> </uafDescription>
        <uafTransformer> </uafTransformer>
        <platform> </platform>
        <parameters> </parameters>
```

```
        <renderAnalyseEngines> </renderAnalyseEngines>
        <outputFormat> </outputFormat>
        <inputFormat> </inputFormat>
        <racObject> </racObject>
  </structureInformation>
  <semanticInformation>
        <raoObject> </raoObject>
  </semanticInformation>
</representationInformation>
</contentInformation>

<primaryDigitalObject> </primaryDigitalObject>

</preservationDescriptionInformation>
</informationPackage>
```

APPENDIX C

Archiving the University of Edinburgh web site

Prior to deposit:

1. Deliver site to NS D-drive via CD Rom.

2. Harvest metadata via OCLC site – any additional admin & *audit trail*/ technical metadata to be added.

3. Create the 'Information Package' (OAIS Terminology)

    - Content data object (HTML, jpegs, gifs, CSS, DTD).
    - Representation Information (includes ref to unzip programme, index of file tree).
    - Preservation Description Information (bib, admin, tech metadata- XML files).
    - Unique ID- standard text file.
    - Checksum- MD5 string to produce unique no.

4. Compression of entire package- tar (to capture *byte stream* and directory structure).

5. FTP package to EUCS archiving service- 'daro' account.

Upon accession:

6. Back up copy kept on CD Rom - Special Collections.

7. Catalogue addition to Edinburgh Library OPAC, for searching bib info.

8. NS to keep local 'Index' database of index, checksum information, archive location, accession date, unique ID etc.

9. Access to archival copy only via NS.

10. Long term preservation: Monitor site's readability at yearly intervals, migrate to newer mediums.

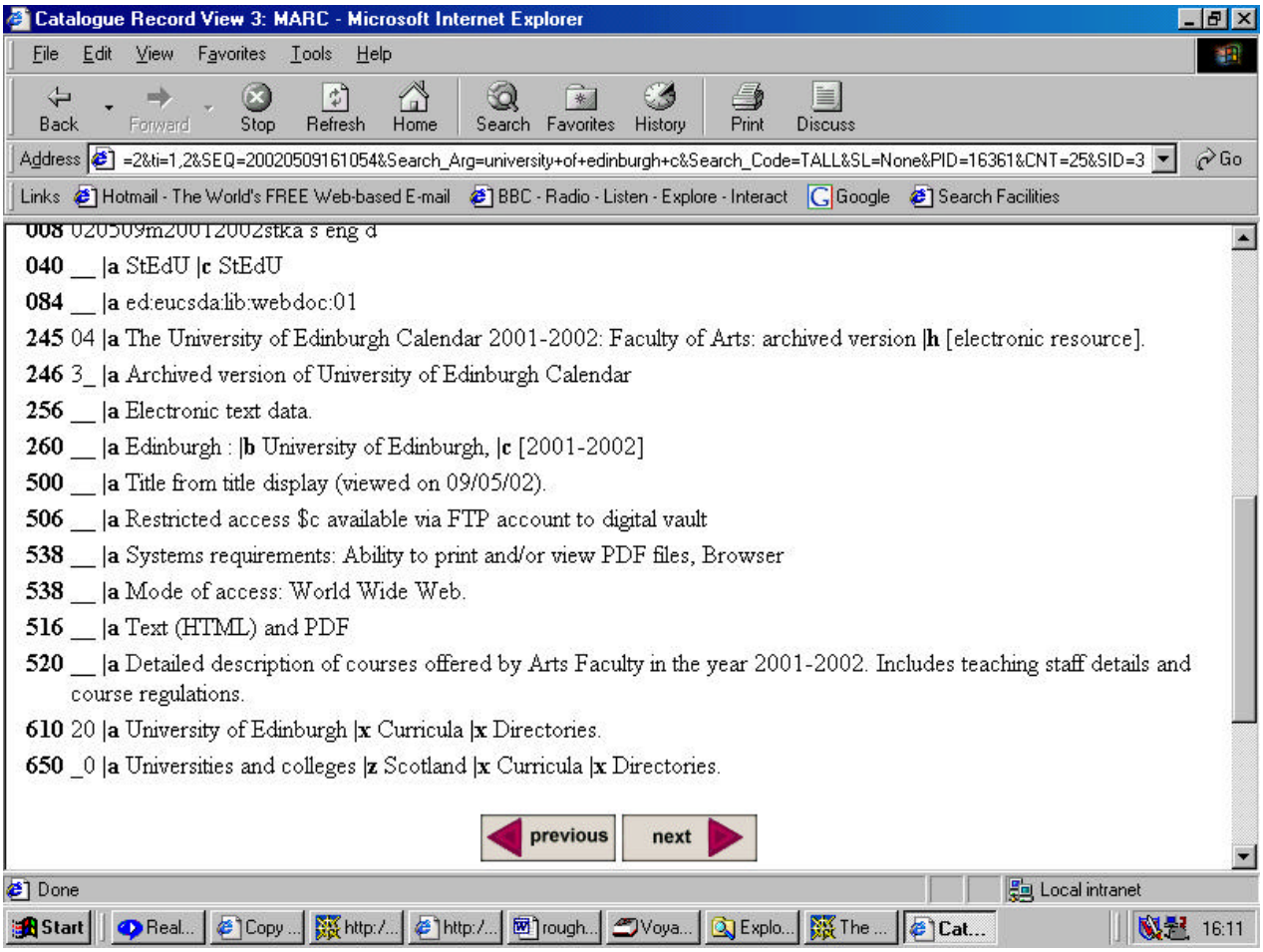


Fig. 4- Edinburgh OPAC Catalogue entry for Calendar

Bibliography

Aschenbrenner, A (2001) 'Long Term Preservation of Digital Material- Legal issues' Available at: **http://www.ifs.tuwien.ac.at/~aola/publications/thesis-ando/Legal_Issues.html** [11.04.02]

Barry, R (2001) 'Making a Difference – Revisiting the IU ER Project Five Years Later' Available from: http://www.rbarry.com/ [15.04.02]

Beagrie, N (2001) 'Towards a digital preservation coalition in the UK.'
Available from: http://www.cni.org/Hforums/ninch-announce/2001/0010.html. [19.4.01]

Beagrie , N and Greenstein, D (1998) 'A strategic policy framework for creating and preserving digital collections' London: British Library Research and Innovation Report: 107

Bearman, D (1999) 'Reality and Chimeras in the preservation of electronic records'
D-Lib Magazine, April 1999.
Available from: http://www.dlib.org/dlib/april99/bearman/04bearman.html. [19.5.01]

Bennett, J (1997) 'A framework of data types and formats and issues affecting the long term preservation of digital material' London: British Library Research and Innovation Report: 50

Bide, M (1999) 'Digital preservation: an introduction to the standards issues surrounding the deposit of non-print publications' London: Library and Information Commission, 1999.

BSI/Rob Allen et al. (1999) 'Evidential weight of information stored on electronic management systems' *DISC* PD 0008[Code of Practice] BSI; London

Cargille,K (2000) 'Digital archiving whose responsibility is it?' *Serials Review* 2000 v26 n 3

The CAMiLEON Project, http://www.si.umich.edu/CAMILEON/. [19.6.01]

The Cedars Project, www.leeds.ac.uk/Cedars [16.05.02]

'Code of practice on the management of records under the Freedom of Information Act', London, Public Records Office. Available at: http://www.pro.gov.uk/recordsmanagement/freedomofinformation.htm

Conway P (1996) 'Preservation in the digital world' *Microforming and Imaging Review* v 25 No.4

Crockett, M and Foster, J (2001) 'Training in Archive Skills Consultancy: an Ebastd handbook' available on CD ROM [16.5. 01]

Gatenby, P (2000) 'Digital archiving developing policy and best practice guidelines at the national library of Australia'
Available from: http://www.icsti.org/icsti/2000workshop/gatenby.html [16.4.01]

Granger, S (1999) 'Metadata and digital preservation a plea for cross interest collaboration' *Vine* 1999, Issue 117, Part 2, , pp.24-29
Also available at: http://dspace.dial.pipex.com/stewartg/metpres.html. [8.5.01]

Graham, P (1995) 'Preserving the Digital Library' p.7-19 in
Long term preservation of electronic materials : a JISC/British Library workshop as part of the Electronic Libraries Programme, 27th-28th November 1995 at the University of Warwick (1996) London : British Library Research and Development Department

Granger, S (2000) 'Emulation as a digital preservation strategy' *D-lib magazine* October 2000 Available from: www.dlib.org [8.5.01]

Guthrie, K (2001) 'Archiving in the Digital Age', Educause Review, November 2001 Available from: http://www.educause.edu/pub/er/erm01/erm016w.html [15.05.02]

Hedstrom, M (1995) 'Preserving Digital Information' in: *Long term preservation of electronic materials*: a JISC/British Library workshop as part of the Electronic Libraries Programme, 27th-28th November 1995 at the University of Warwick (1996) London : British Library Research and Development Department.

Hedstrom, Margaret and Lampe, Clifford (2001) 'Emulation vs. Migration- Do users care?' Available from; http://www.rlg.org/preserv/diginews/diginews5-6.html#feature1 [13.05.02]

Edstrom, M and Montgomery, S (1998) Digital Preservation Needs and Requirements in RLG Member Institutions: A study commissioned by the Research Libraries Group
Available from: www.rlg.org [13.6.01]

Hendley. T. 'Comparison of Methods and Costs of Digital Preservation.'
British Library Research and Innovation Report 106. London: United Kingdom.

Jones, M & Beagrie, N (2000) 'Preservation management of digital materials' Re:source 2000 workbook Available from: http://www.jisc.ac.uk/dner/preservation/workbook/workbook.pdf [8.5.01]

Lauder, S (2000) 'Digital Preservation Society' *Scottish libraries* 14.2

Lievesley, D and Jones, S (1998) 'An Investigation into the Digital Preservation Needs of Universities and Research Funders: the Future of Unpublished Research Materials'. London; British Library RIC Report no.109

Long term preservation of electronic materials': a JISC/British Library workshop as part of the Electronic Libraries Programme, 27th-28th November 1995 at the University of Warwick (1996) London : British Library Research and Development Department

Lynch, Clifford (1999) 'Canonicalisation: a fundamental tool to facilitate preservation and management of digital material.' *Dlib Magazine, 1999* Available from:
http://www.dlib.org/dlib/september99/09lynch.html [16.05.02]

The National Library of Australia's Preserving Access to Digital Information (PADI). Available from: http://www.nla.gov.au/nla/listserv/padi-l.html. [18.6.01]

Marcum, D (1997) 'Preservation in the digital age a moral and legal obligation' *International Information and Library Review* v.29 357-365.

Ockerbloom , J (2001) 'Archiving and Preserving PDF Files' *RLG Diginews February 2001*
Available from: http://www.rlg.org/preserv/diginews/diginews5-1.html#feature2 [05.12.01]

'OCLC/RLG Preservation Metadata Working Group Issues White paper' (2000)
Available from: http://www.oclc.org/digitalpreservation/ [8.5.01]

Pace, C (2000) 'A digital preservation; everything new is old again' C*omputers in Libraries* February 2000 v20 i2 p55

'Preserving Metadata for digital objects:*A* review of the state of the art' a white paper by the OCLC/RLG working group on preservation Metadata (2001)

Ross, S and Gow (1999), 'A Digital Archaeology: rescuing neglected and damage resources' Available at: http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2.pdf

Rothenburgh, J (1998) **'**Avoiding Technological Quicksand: finding a viable technical foundation for digital preservation' Council on Library and Information Resources

Available from: **http://www.clir.org/pubs/reports/rothenberg/research.html#summary** [8.5.01]

Russell, K (1999) 'Digital Preservation : ensuring access to materials in the future' Available at: http://www.leeds.ac.uk/cedars/Chapter.htm

UK JSTOR Mirror Service (2001) http://www.mimas.ac.uk/jstor/ [3.07.01]

Victorian Electronic Records Strategy, Available from www.prov.vic.gov.au/vers/welcome.htm [15.05.02]

Waters, D & Garret, J (1996) 'Preserving Digital Information: Final Report and Recommendations' Commission on Preservation and Access and the Research Libraries Group. 20 May 1996. Available from: http://www.rlg.ac.uk/ArchTF/ [16.5.01].

Wheatley, P (2000) 'Migration – a Camileon discussion paper' Available from: http://www.ariadne.ac.uk/issue29/camileon/ [15.05.02]