

The AHDS Archival Digital Repository

The Arts and Humanities Data Service (AHDS) was established in 1996 to collect, preserve and encouraging the reuse of digital resources created during scholarly research in the arts and humanities. The AHDS is now responsible for the preservation of over 3,000 digital resources and holds a wide range of data types, from plain text and image files to datasets (spreadsheets, databases, statistical data files) and digital recordings (audio or audio/video), as well as more complex resources such as Web sites and GIS (Geographical Information System) data.

The AHDS is organised as a distributed service, consisting of a managing executive, hosted by King's College London, and five subject centres (AHDS Archaeology, AHDS History, AHDS Literature, Languages and Linguistics, AHDS Performing Arts and AHDS Visual Arts) based at universities across England and Scotland. Each AHDS Centre takes responsibility for digital resources in its subject area. In the past, each centre has made separate arrangements for the preservation and delivery of its collections. One of the anticipated results of this was that each Centre adopted a different approach. AHDS History and AHDS Archaeology emphasised long-term preservation, while the other three Centres focused more on providing online access to their holdings. Now, as part of a move to a more centralised organisational model, the AHDS is developing an archival digital repository that will be used by all five AHDS Centres to preserve their collections, and will serve as the backend to a new range of common access methods provided to our users.

In recent years, there has been considerable interest in the role that digital repositories and archives (the two terms can usually be treated as synonymous) may be able to play in assisting the management and preservation of digital resources. Discussions about digital repositories can be found in a number of guises, including data archiving, digital libraries, e-print self-archiving and e-learning repositories. There is, however, considerable variation in what is meant by the term digital repository, and, to avoid confusion, it is important to appreciate the different emphasis given to the term by different groups. While there is general agreement that digital repositories are store places for digital resources, views vary on the extent to which their role extends beyond providing access to digital resources and towards supporting their long-term preservation, and on whether a digital repository is something operated by an organisation, or is itself an organisation.

These two criteria of the repository's goal, access or preservation, and the perspective taken on the scope of repository provide a useful categorisation of repository efforts (figure 1).

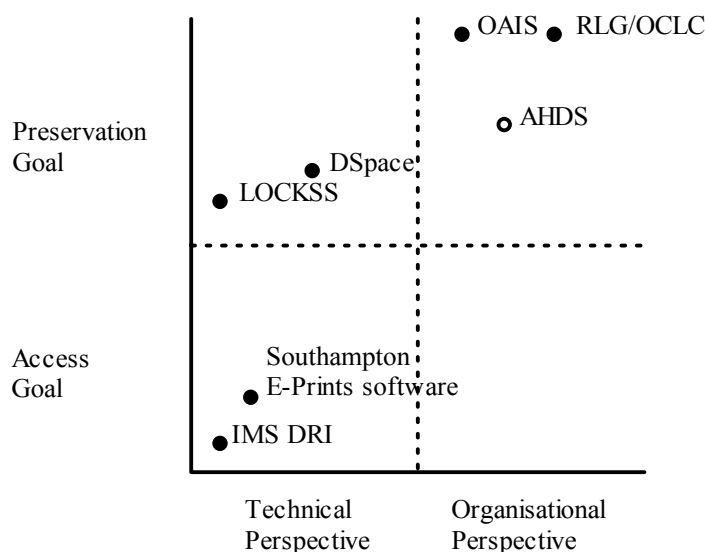


Figure 1: Different Concepts of the Digital Repository

At one extreme, the term digital repository is used to describe a networked system of hardware, software and metadata designed to store and deliver content. In this vein, the IMS Digital Repositories Specification (IMS DRI) defines a digital repository as:

Any collection of resources that is accessible via a network without prior knowledge of the structure of the collection. Repositories may hold actual assets or the metadata that describes assets.
(IMS, 2003a)

In contrast, the library and archive communities tend to have a much broader organisational view of the repository; where technical requirements are only one of a number of factors to be considered. For example, the joint OCLC and RLG working group on *Attributes of a Trusted Digital Repository* concluded that:

A trusted digital repository is one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future.
RLG (2002, p.5)

These two quotes are indicative of an important distinction that can be made between *archival digital repositories*, which emphasize long-term preservation of content, and the short-term access goals of *transitive digital archives*, such as e-print archives and learning object repositories (McClean & Lynch, 2003, James et al, 2003; Barker et al, 2004)

Access and preservation of digital content are of course closely linked. The conversion of binary data into meaningful information relies on a complex chain of hardware, software and formats, all of which are subject to on-going technological change. Consequently, providing long-term access to digital content inevitably involves the challenges of digital preservation. Content in a digital repository must be protected from the problems of data corruption and technological obsolescence, and the authenticity of the content must be ensured in some way.

These problems form the basis for work on developing the requirements for archival digital repositories. Internationally, the most influential work to date on the design of an archival digital repository has been the OAIS (Open Archival Information System) reference model (CCSDS, 2002), which describes:

an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community.
CCSDS (2002, p.1-1)

The OAIS model was developed to improve the preservation of data collected during space missions, but it has found its way into most discussions about digital preservation. This model describes in (at nearly 150 pages) considerable detail the six main activities an archival digital repository must undertake – ingest, archival storage, access, data management, administration and preservation planning. The OAIS model provides a wealth of valuable guidance for repository planners and managers. Indeed, it is expressly “designed as a conceptual framework in which to discuss and compare archives” (CCSDS, 2002, p.1-3). However, this does mean that actual implementation of the model is left to the individual repository planner.

The AHDS has used the OAIS reference model in this light, as a helpful summary of repository functionality against which plans for the new AHDS archival digital repository plans could be compared. Figure 2 gives a very simple overview of the new AHDS repository, highlighting the main flow of data and metadata through the AHDS. A simple indication is given of how each element in the repository maps to the OAIS model.

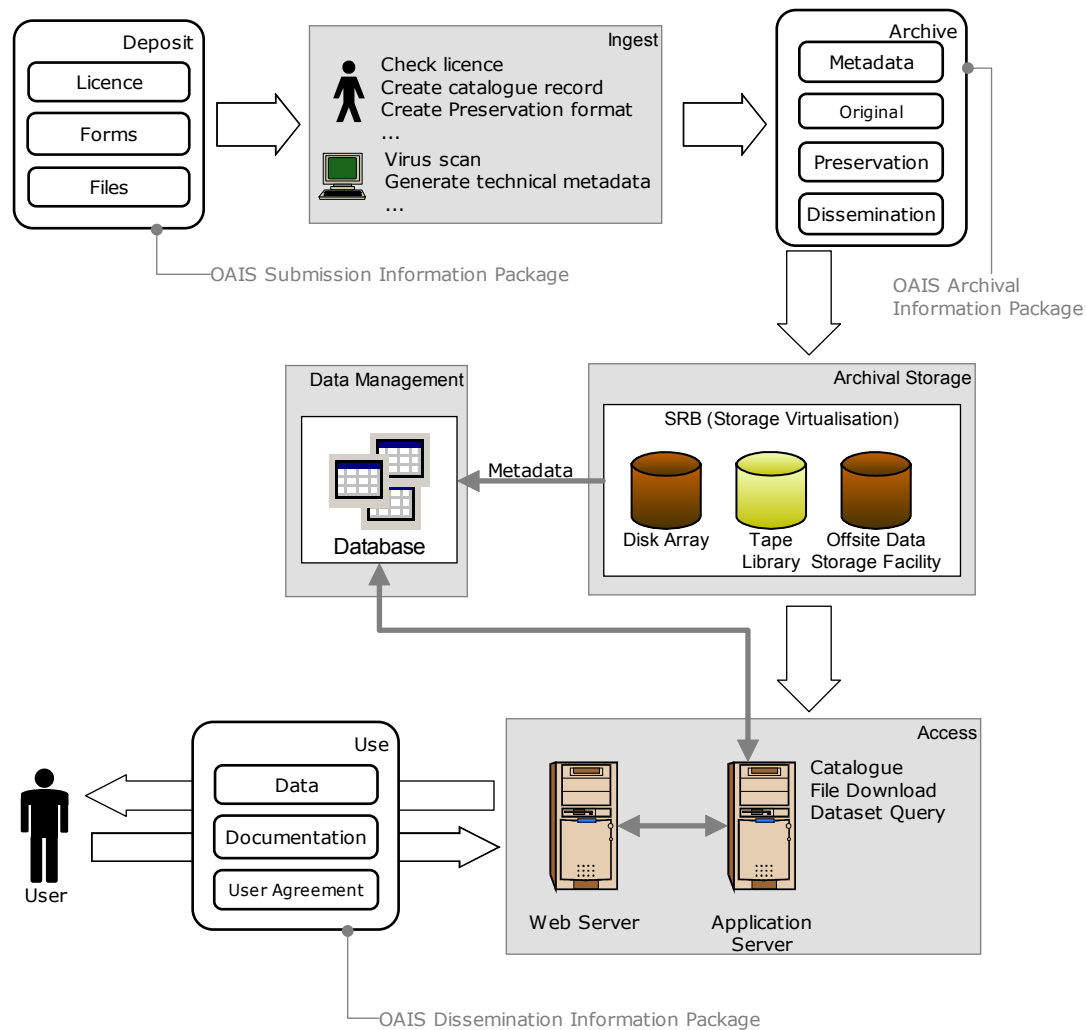


Figure 2: Outline of AHDS Repository

The AHDS began development of its new digital repository in 2001 with an internal review of collections management practices across the five AHDS Centres. This review identified differences in the relative effort assigned to ingest, preservation and delivery of collections at each of the AHDS Centres, reflecting differences in the focus of each Centre's work. In 2002, external consultants were commissioned to follow up this internal report by examining AHDS collections management and digital preservation practices. This work was completed in early 2003, and produced a series of seven reports. The most significant of these reports looked at current digital archiving and preservation practices across the AHDS, the range of data types and formats received by the AHDS, requirements for preservation metadata, and proposed general requirements for a standardised AHDS wide ingest procedure. These reports have served as the starting point for three internal working groups, focusing on technical systems, collections management, and metadata, which have been responsible for designing the repository.

Preservation Strategy

The great difficulty anyone trying to build an archival digital repository will encounter is the lack of practical advice, and of robust tools and mature techniques for digital preservation. A number of digital preservation strategies have been proposed, but there is no definitive approach to the problem of maintaining digital content across multiple generations of technology. Unfortunately, information on the likely costs, possible limitations and long-term sustainability of different strategies is far from complete – partly it must be said, for the very valid reason that no one has yet had the time to gain the experience needed to answer these questions.

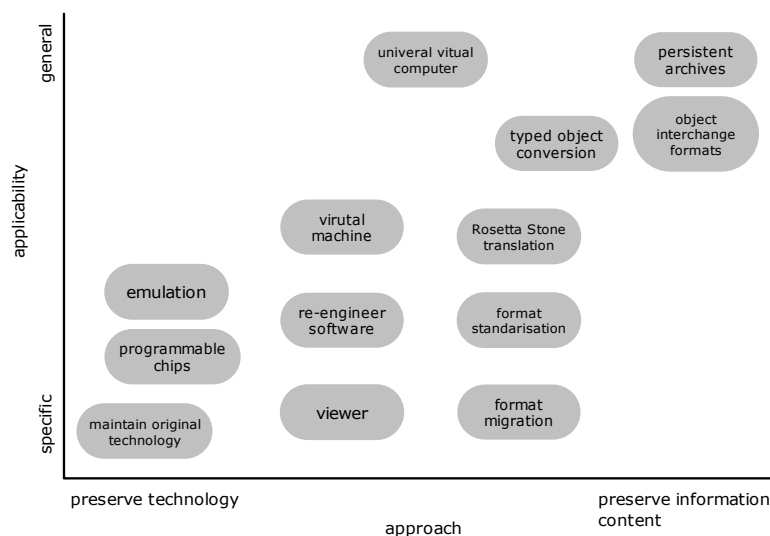
Few organisations with digital preservation responsibilities appear to have yet fully developed their policies in this area (ERPANET, 2003). It is unwise to commit to a course of action before its consequences are clear, nevertheless it is equally important that organisations which make some claim to preserve digital resources should declare to their stakeholders what they can do to achieve this goal at the present time.

Prompted particularly by the external consultancy report on AHDS collections management practices, the AHDS has developed a Collections Preservation Policy that defines the broad approach the AHDS takes to preserving digital collections. Issues to do with the selection and retention of collections are dealt with separately in the AHDS Collections Policy. The Collections Preservation policy is based on a three tiered understanding of digital preservation:

1. Preservation of the bit stream (basic sequences of binary digits) that ultimately represent the information stored in any digital resource
2. Preservation of the *information content* (words, images, sounds etc.) stored as bits and defined by a logical data model, embodied in a file or media format
3. Preservation of the *experience* (speed, layout, display device, input device characteristics etc.) of interacting with the information content

Techniques for achieving the first of these objectives are well understood and include environmentally controlled storage, data replication, backup, and media refreshment. In the OAI model, much of this activity falls into the archival storage function. The second and third objectives present a far greater challenge.¹

Binary data remains useful only for as long as it can be correctly *rendered* (displayed, played-back, interacted with) into meaningful content such as text, images and video clips. The process of rendering is performed by a complex mix of hardware and software, which is subject to rapid obsolescence. As a rule of thumb, it is reasonable to predict that current hardware and software will be able to correctly render a file for around ten years after its creation. By the end of this period, repositories need to have adopted a more active preservation strategy than simply preserving the bit stream of the file if they are to maintain access to information content held in the file. Either old data must be altered to operate in a new technical environment (migration, format standardisation) or the new environment must be modified so that it can render the old data (emulation, virtual computers).² Within these two broad approaches there are many different techniques (figure 3).



¹ The third objective may not be regarded as necessary in many circumstances.

² An important point to draw out here is that a digital repository typically only fully controls the data, and not the hardware and software needed to render that data. External commercial decisions about product lifecycles will often trigger the need to undertake a preservation action such as migration.

Figure 3: Digital Preservation Strategies (based on Thibodeau, 2002)

The AHDS has adopted an approach that relies on a mix of format standardisation, and format migration to preserve the information content of digital resources in an accessible state. Although migration is far from ideal and emulation should, in principle, provide a more efficient solution than migration, the approach is not yet widely used (Digital Preservation Testbed, 2003). Small organisations like the AHDS lack the resources to design and maintain emulators, so a purely emulation based approach to digital preservation would leave the AHDS dependent on outside developers to ensure access to AHDS holdings in the future. In contrast, most software supports several data formats and versions of data formats, allowing an archival digital repository to use non-specialist software to carry out migration..

The AHDS does follow the sensible practice of preserving the original bit stream alongside any migrated versions of the resource. This at least allows for the possibility of direct access to the original format of the resource in the future, although this will depend on collaborative or third party work to develop emulators and other tools. The AHDS relies on migration as the primary means of preserving its holdings.

Making these decisions about preservation strategy has helped shape subsequent decisions about the design of the new AHDS digital repository. The AHDS digital repository will actually store three logically distinct versions of each digital resource.

The *original version*, which consists of all the files given to the AHDS by a depositor, along with scanned copies of administrative paperwork including the AHDS licence form, catalogue form and data and documentation transfer form.

A *preservation version* consisting of the content encoded in the files of the original version, but periodically migrated so that there is always a version of the content held in a format that is currently accessible. At the time of deposit, the preservation version may consist of the same physical files as the original version, or these files may be migrated during ingest into file formats that are likely to have greater longevity. There are a range of factors relating to file formats that suggest greater longevity:

- The file format is based on an open or freely published standard
- Software that can correctly render the format is easily available, from a number of sources, and for a number of platforms
- The format is popular and widely used

The *dissemination version(s)* of the collection are again made up of the content encoded in the original version, but this time made available in file formats that are easy and convenient to use. In the new repository, this version may consist of nothing more than the metadata needed to describe a transformation from a preservation format into a format suitable for users. Where files are created, there is no intention that they will be kept in the long-term.

With the repository holding both files created by the depositor and the AHDS, and which may each serve several purposes, extensive metadata will be key to managing the contents of the repository.

Apart from the extensive work the information (metadata) working group has carried out to establish AHDS wide standards for descriptive metadata, requirements for structural, administrative and technical metadata have also proved difficult to establish.

The external consultancy produced a report providing a recommended set of elements for preservation metadata based on a review of existing proposals, but this has subsequently proved to be too ambitious. Quite a number of elements would be difficult to collect automatically, and therefore would consume limited staff time. The proposal also lacked vital detail such as methods for controlling the content of elements for describing file formats and software, without which many elements would contain free text content of limited value.

For the present, the collections working group is developing a much simplified version of the recommendations from the consultancy. The vital technical information to collect is the file format and exact version for each file being preserved; combined with external sources of information, such as The National Archive's PRONOM database, this information should make it possible to identify the software and hardware requirements of a resource and possible migration pathways. As another economical way of recording the technical requirements of each digital resource, basic details of the software and hardware used by the AHDS to access a resource during ingest will be recorded. Unlike determining the original or recommended hardware and software for a resource, recording the equipment actually used to open the resource is much easier.

The other main component of the preservation metadata we hope to capture is an audit trail of 'who' did 'what', 'where', and 'how' to each collection:

Element Name	Definition	Vocabulary
Process: Purpose	An indication of the nature of the process, e.g. "format version migration" or "export".	Controlled
Process: Agent	The name of the individual, organization or software process that carried out the process.	Standard layout
Process: Date	The date when the process was carried out	Standard layout
Process: Description	A short plain text description of the process.	Uncontrolled
Process: Software	Important software used in the process	Standard layout
Process: Hardware	<u>Critical</u> hardware used in the process.	Uncontrolled
Process: Results	Outcome of process (success/failure) plus information on how success was validated.	Uncontrolled

This set of elements is based, in particular, on experience at AHDS History and the UK Data Archive, where there is a long-standing practice of recording activity in a semi-structure note file held with each collection. This information is often very useful when working on older collections.

An as yet unresolved problem is how to deal with the occasional overlaps between descriptive metadata records, held in TEI, DDI or the AHDS in-house CMF (Common Metadata Framework) schemas and the preservation metadata schema we are writing. Some pieces of administrative and rights information, for example are stored differently depending on the schema currently being used, and while we want to be able to continue to support standards such as the DDI and TEI, we do not wish to store multiple, possibly conflicting, versions of vital information.

Archival Storage and Access

The archival store for the repository will be based on a file system, and metadata for each collection will be stored as an XML file within the directory structure for a collection (figure 4) The present intention is to use a METS (Metadata Encoding and Transmission Standard) document as a container for all the metadata related to a particular collection or item. METS allows other metadata schemas to be embedded in it, making it useful as a container for storing or sharing metadata. Its main native value comes from the element sets it provides for describing the logical and physical structure of complex resources. However, there is currently a lack of tools for creating this type of metadata easy. METS is also open to many different interpretations, and a number of specific issues may hamper interoperability.

Storing the metadata in the repository reduces the chances that the data and metadata will become separated. To make practical use of the metadata, though, it needs to be held in a database for quick access. Information in the database will be updated by trawling the repository's archival store for new or modified metadata files that will then be read into the database.

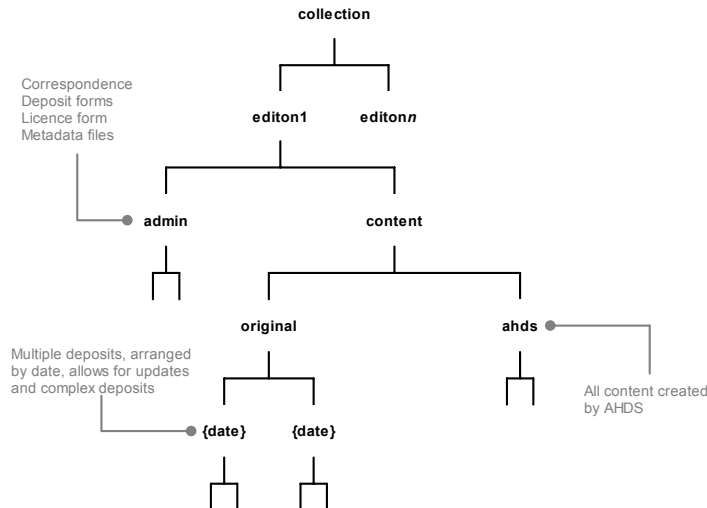


Figure 4: Directory Structure for Collections in the AHDS Repository

A disk array (two terabytes) and a tape library (with an attached capacity of ten terabytes) will provide the initial storage for the AHDS repository. Offsite data replication is being arranged with the CCLRC e-Science Centre Atlas Petabyte Data Store. There are now many affordable solutions for data storage of volumes into the terabytes. Storage volumes are increasing as rapidly as the cost of storage devices is decreasing, and there is some evidence to suggest that the cost of digital storage is similar to the cost of traditional storage of archival material (Chapman, 2003).

Given that the AHDS currently holds only approximately one terabyte of data, planning for a repository of over ten terabytes may seem excessive. While part of this large margin for growth is simply a result of economies of scale when purchasing equipment, it is also needed to allow for the possibility of rapid and unpredictable growth in our collection. Particular types of digital content, especially images, audio and audio/video, generate very large volumes of data. The main component of any significant increase in the volume of data (as opposed to numbers of digital resources) deposited with the AHDS in the next few years will probably be digitised images. At the moment, digitised images form about half of the total data volume held by the AHDS, although they constitute only a mere 20 or so of our 3,000 collections. The largest confirmed future deposit with the AHDS is a collection of digitised images that will total between two and three terabytes of data – a single collection that will be at least twice the size of our total holdings at present. In short, a small number of deposits may dramatically alter the total size of our holdings, and the repository needs to have sufficient space to handle a few unexpectedly large deposits.

The various storage devices will be managed through Storage Resource Broker, a storage virtualisation product developed by the San Diego Supercomputer Center (<http://www.npaci.edu/DICE/SRB/>), which is free for universities to use. SRB supports data replication and the ability to view data distributed across multiple storage devices as though it part of a single file system, which is a useful way of simplifying the management of a large repository.

SRB is a useful management tool that, well less polished than commercial alternatives, is much cheaper – free in fact – but it is not, in itself, a useful means of delivering content to users. Separate servers will run the archival store and the access services for users. This clear separation between storage and access distinguishes the AHDS repository from the approach taken by products such as DSpace, which allows users direct access into the archival store. The separate achieves two main aims. First, it allows access to files to be more closely controlled. Not all the material the AHDS holds should or can be delivered to users online. Some files, such as master TIFF images and some datasets are too large to deliver without pre-processing, while other files are held only for preservation due to rights issues. Second, the separation should act as an additional disincentive to altering content in the

repository to fit with the requirements of current delivery methods. Instead, the philosophy the AHDS has adopted is that content in the repository should be optimised for preservation, and transformed as necessary to make it easier to deliver to users.

An immediate practical consideration is that SRB may not yet be fast enough to support on-the-fly requests for data, so copies of data for delivery may need to be held outside the SRB managed archival store.

The AHDS repository will provide a range of access services, tailored to different types of material. A simple file download system, online display of text or images, and the possibility of a streaming service at a later date, will be offered alongside server based tools such as query tools for datasets and tools for analysing electronic texts. This can be done using a vast array of technologies, although they all tend to require customisation and on-going technical support to provide a reliable service. Having considered frameworks like FEDORA, the AHDS is using more generic Java and XML (Cocoon) technologies to integrate existing tools, or create new ones for online delivery of collections.

Ingest

Decisions made during the creation of a digital resource, especially the choice of software and format, can have a long-lasting impact on the feasibility and cost of retaining the digital resource (Jones & Beagrie, 2002). Through our work advocating data creation standards and best practices to researchers, the AHDS is perhaps unusually well placed to influence the design and construction of the digital resources that will eventually be deposited with us. Nevertheless, the first point at which any archival digital repository can definitely prepare a digital resource for preservation is during ingest into the repository.

Digital resources deposited with the AHDS are prepared for preservation using a semi-manual approach; an expert member of staff, following general procedures and best practice guidance, makes decisions about the ingest process that are then implemented using a variety of software tools. This approach is suitable for handling the diverse range of essentially 'one-off' digital resources received by the AHDS. Digital resources deposited with the AHDS vary considerably in nature and scale, ranging from collections encapsulated in a single one hundred kilobyte file, to collections containing thousands of files that add up to hundreds of gigabytes of data. However, neither the number of files nor the amount of data is necessarily a useful guide to the amount of work that may need to be done during ingest. The key factor is the complexity of the resource. Homogenous resources, where the same tools and techniques can be applied to all the objects in the collection are easier to ingest than heterogeneous resources. A thousand TIFF image files will be less difficult to preserve than a website made up of 100 HTML, GIF, JPG and script files. The number of distinct file types in a resource contributes to complexity, as does the number of connections between separate files. The most useful thing to judge is often the number of objects in a resource, where objects may be stored as several files, a single file, or parts of a file or files. A Microsoft Word file can, for example, be either relatively simple, containing only plain text, or very complex, including tables, equations, complex styles, and embedded objects such as spreadsheets and video clips.

Some repositories specialise in particular types of digital resource as a way of reducing this complexity. Although the AHDS cannot do this, we do seek to standardise aspects of resources when they are deposited. The AHDS provides a list of preferred, acceptable and problematic file formats for depositing a variety of resource types (<http://www.ahds.ac.uk/depositing/deposit-formats.htm>) as a first step towards doing this.

The second major step we take is to try and standardise the way digital resources are represented. This process is made much easier when there is a shared understanding between creators, curators and users of digital resources about what constitutes the *significant properties* (Cedars Project, 2002) of the digital resource. The significant properties of a digital resource are those aspects of how it is rendered that are regarded as crucial to its correct use and interpretation. It is important to press depositors to provide good documentation explaining the provenance, structure, and use of their resource. This type of

information helps repository staff identify the *significant properties* (Cedars Project, 2002) of a resource and decide how best to handle it. This information can then be used to identify particular file formats that will best preserve the information content of a resource. The AHDS is currently codifying its understanding of significant properties in a series of preservation handbooks that will provide guidance to staff responsible for preparing resources for preservation. These tackle the issue from two angles, firstly considering the set of significant properties are suggested by a particular data type – information such as the words, paragraphs and font styles in a textual document, for example – and then overlaying these with additional considerations from the way different data types are used across the arts and humanities. Historians, for example, often make use of relational database software when compiling information from historical documents. The order of records in historical documents is frequently very important and historians are apt to assume that the same applies to relational databases, which is not the case. So, when dealing with relational databases deposited by historians it is wise to assume that the order of the rows is significant, even though this is not a significant property of a relational database.

Conclusion

Scholarly digital resources are expensive to create and if their full value is to be exploited, they must be preserved and made available in a form suitable for informed reuse. This is the role of the new AHDS archival digital repository.

Developing an archival digital repository is a challenging task. Work must be carried out in a number of areas: deposit and ingest, digital preservation planning, data management, storage and access. In all these areas, policies, procedures must be developed to guide the work of repository staff, while hardware and software must be selected and installed to support their work.

The design of new AHDS repository is one way of addressing the need to preserve and deliver digital resources. It represents a balance struck between a series of factors, such as the flexibility of semi-manual processes against the efficiency of fully-automated processes, and the cost of taking preservation actions at the time of ingest against the danger of leaving them too late. The completed repository should secure the future of the AHDS's collection, and provide a solid basis for expanding that collection into larger and more complex resources in the future.

Hamish James
Collections Manager
Arts and Humanities Data Service

References

- Barker, E., James, H., Knight, G., Milligan, C., Polfreman, M. and Rist, R. *Long-Term Retention and Reuse of E-Learning Objects*, Report commissioned by the Joint Information Systems Committee (JISC), March 2004. Retrieved from http://www.jisc.ac.uk/index.cfm?name=project_elo on Sep 10, 2004.
- Cedars Project. *Cedars Guide to Digital Collection Management*. 2002. Retrieved from <http://www.leeds.ac.uk/cedars/guideto/collmanagement/guidetocolman.pdf> on Dec. 18, 2003.
- Chapman, S. 'Counting the Costs of Digital Preservation: Is Repository Storage Affordable?'. *Journal of Digital Information*, vol. 4 no. 2. 2003. Retrieved from <http://jodi.ecs.soton.ac.uk/Articles/v04/i02/Chapman/> on Sep. 09, 2004
- Consultative Committee for Space Data Systems [CCSDS] (2002). *Reference model for an open archival system*, CCSDS 650.0-B-1 Blue Book Retrieved from <http://www.ccsds.org/documents/650x0b1.pdf> on May 3, 2003.
- Digital Preservation Testbed. *Migration: Context and Current Status*, 2001. Retrieved from <http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf> on Jan 5, 2004
- Digital Preservation Testbed. *Emulation: Context and Current Status*. 2003. Retrieved from <http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=185&categorie=2> on Sep 10, 2004.
- ERPANET. 'Policies for Digital Preservation': *Seminar Report*. 2003. Retrieved from <http://www.erpanet.org/events/2003/paris/index.php> on 10 Sep, 2004.
- IMS. *IMS Digital Repositories Specification*. 2003a. Retrieved from <http://www.imsglobal.org/digitalrepositories/index.cfm> on 13 Mar, 2004.
- James, H., Ruusalepp, R., Anderson, S. and Pinfield, S., *Requirements and Feasibility Study on the Preservation of E-Prints*, Report commissioned by the Joint Information Systems Committee (JISC), May 2003. Retrieved from http://www.jisc.ac.uk/index.cfm?name=project_eprints_pres on Sep. 10, 2004.
- Jones. M & Beagrie. N, "Preservation Management of Digital Materials", 2002. Retrieved from <http://www.dpconline.org/graphics/handbook/> on Mar. 15, 2004.
- Lynch, C. A. *Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age*. ARL Bimonthly Report, no. 226. 2003. Retrieved from <http://www.arl.org/newsltr/226/ir.html> on Mar. 15, 2004.
- Mclean, N & Lynch, C. *Interoperability between information and learning environments – bridging the gaps, Draft version*, IMS Global Learning Consortium & Coalition for Networked Information, 28 June, 2003. Retrieved from http://www.imsglobal.org/DLims_white_paper_publicdraft_1.pdf on Mar. 15, 2004.
- RLG, Research Libraries Group [RLG]. *Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report*. 2002. Retrieved from <http://www.rlg.org/longterm/repositories.pdf> on Mar. 3, 2004.
- Thibodeau, K. 2002. 'Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years' in proceedings of *The State of Digital Preservation: An International Perspective*. Conference Proceedings. Washington. 2002. Retrieved from <http://www.clir.org/pubs/abstract/pub107abst.html> on Sep. 10 2004.