Digital Archiving at the Archaeology Data Service: A Quest for OAIS Compliance

Jenny MITCHAM[1] – Julian D. RICHARDS[2]

Archaeology Data Service, University of York
[1]jlm10@york.ac.uk
[2]jdr1@york.ac.uk

Abstract

The importance of secure digital archiving of archaeological data cannot be underestimated. This is particularly the case where the archaeological data in question cannot be recreated because a site has been lost through the very process of excavation, and where primary data are increasingly 'born-digital'. Since its inception in 1996, the Archaeology Data Service (ADS) has been archiving and preserving archaeological data for the benefit of current and future researchers. When ADS was established there were few standards for digital preservation. However, over the last decade a number of initiatives have emerged and the archaeological and heritage community needs to be aware of these. One of these is now an ISO standard known as the Open Archival Information System (OAIS). This is a framework and reference model that all archives can use to ensure that appropriate basic activities and data flows form part of their preservation work. The OAIS also provides a language and set of terms that archives can use to communicate with each other. At the ADS, we have spent some time looking at the OAIS model and mapping our activities to it in order to establish whether we can describe ADS as OAIS conformant. This paper describes the things that an archive needs to do in order to comply with the OAIS model and how staffing structures and day-to-day activities map to it.

Keywords

Digital archiving, Archaeology, case study, Open Archival Information System, OAIS

1. Introduction

The Archaeology Data Service (ADS) was founded in 1996 for the purpose of preserving digital data produced by archaeologists based in the UK, and making it available for scholarly re-use. The ADS was initially established as part of the Arts and Humanities Data Service (AHDS), with sister services covering other disciplines within the arts and humanities. Archaeologists based in Higher Education institutions were able to deposit data for long term preservation free of charge, on the basis of core funding from the Arts and Humanities Research Council and the Joint Information Systems Committee. However, the ADS also developed a Charging Policy (http://ads.ahds.ac.uk/project/userinfo/charging.html) based on the principle of a one-off charge levied at the point of deposit to cover the preservation of digital data derived from archaeological research funded by other bodies, such as governmental agencies including English Heritage, or as part of commercial development. Data are archived to ensure long term preservation, but they are also made available free of charge for download or via online interfaces to encourage reuse.

2. Why is digital archiving so important?

Over the last decade the archaeological profession has become more aware of the importance of digital archiving. Subject to control of extremes of temperature and humidity, the preservation of traditional paper records could largely be a passive process. Digital archiving, by contrast, requires continuous and active data management rather than static data storage.

2.1. Archaeological data is often irreplaceable

Data discovered through archaeological excavation can be collected only once. If that data is lost, it is lost forever — the excavations cannot be repeated in the future. With the high costs of paper publication, less and less of the raw data makes it through to print, and much of it may only ever exist in digital form. For the purposes of reuse, a digital format is far more appropriate for born-digital raw data, and a dedicated digital archive is needed in order to ensure longevity.

2.2. File formats can become obsolete

Information technology moves very quickly. Software companies bring out new and updated versions of their file formats every few years. Although reputable software companies are usually committed to maintaining backwards compatibility to ensure that new versions of software can read old formats, this will not go on indefinitely.

As a recent example of this, in September 2007, Microsoft released Service Pack 3 for Office 2003. Once installed, this update disabled the ability to open the previously supported legacy file formats of Word 97. One of the major problems with this was the fact that this change was not widely publicised so even if users read the release notes before deciding to install this security update, they may not have fully understood the implications of going ahead with the installation. For many, the first they would know of the issue would be an error message brought up when double clicking on a Word 97 file (Ashley 2008).

This and other issues like it are the sorts of events which might stimulate a flurry of activity in a digital archive to migrate files to newer formats, but in an ideal world, we should be doing this *before* these sorts of emergencies occur.

2.3. Media can become obsolete

The problems with file formats are compounded by problems of media redundancy. Floppy disks for example are fast becoming a thing of the past. There was a time when they were one of the most common ways of storing and sharing files, but now they are hardly used at all. As a result of this, many new computers are no longer equipped with floppy disk drives. Where data sit on old media like this, we need to ensure that they are copied to current media before it is too late.

2.4. Data can become corrupted

Data corruption can occur either accidentally or intentionally and it is important that measures are taken to guard against it. These measures should include a robust backup strategy, so that lost data can be recovered where necessary, and a procedure for periodically checking whether corruption has occurred. Some files in an archive may be accessed so infrequently that where a file has become corrupt it may not be discovered for years. It is important to ensure that the integrity and authenticity of all files is maintained on a regular basis so that any issues can be dealt with while 'good' backups still exist.

2.5. Media can become corrupted

The media on which archaeological data are stored does not have an infinite lifespan. CDs do not last forever despite what was once claimed. They could be rendered useless in a time span as short as 2–5 years! Media refreshment should be an essential part of any archiving strategy.

2.6. Expertise and knowledge can disappear

Finally, the loss of human expertise and knowledge is one point that is often forgotten when we talk about the need for digital archiving. Crucially, the staff who created, and therefore understand the files will not be around forever, and even if they were, they may not even understand their own data once they have moved on to a new project or another job. Unless files are very well-organised, named in an easily understandable way, and well-documented, it may be impossible for anyone to make sense of them in 20 years time even if none of the above problems apply. A digital archive needs to ensure that any problems such as these are ironed out at the earliest possible opportunity in order that the data are suitable for re-use well into the future. If this basic step isn't met, there would seem to be little point in storing the data at all.

3. Introduction to the Open Archival Information System (OAIS)

The Open Archival Information System (OAIS) provides a reference model and framework that archives can work with. It defines the basic functional components of an archive and provides a comprehensive framework for describing and analysing preservation issues. Initiated in 1995 by the Consultative Committee for Space Data Systems (CCSDS), and developed through an extensive process of external review and comment, it was approved in January 2002 as international ISO standard 14721. (For a full discussion of the background to the OAIS model see Lavoie 2004, 1–3).

The simple definition of an OAIS as taken from the ISO standard itself is as follows:

"An OAIS is 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, 1-1)

Many organisations using OAIS as a model for their archive find it most useful as a language and set of terms that they can use to describe their activities. This is particularly useful when collaborating with other external archives and organisations. The OAIS model itself is flexible and allows for a very wide range of implementations. By using the language of OAIS, disparate organisations with very different internal structures and procedures can still communicate effectively with each other when discussing their work. The most useful terms are defined in *Table 1*.

Fig. 1 shows a simplified view of an OAIS archive. You can see the producer (who has provided the data to the archive) on the left hand side and the consumer (who is using the data) on the right. The rest of the diagram represents the archive itself and the main activities that are carried out behind the scenes.

The flexibility of the OAIS model is one of its strengths, allowing many different types of archive (both digital and physical) to implement it successfully. There are however a set of mandatory

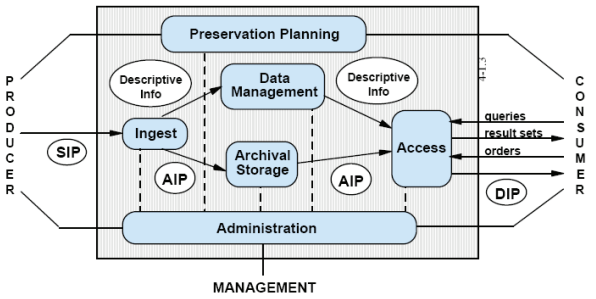

Fig. 1. A simple OAIS diagram illustrating the 6 functional entities (in boxes) the data packages (circled) and the three related interfaces (producers, consumers and management) (taken from CCDSD 2002, 4-1).

responsibilities that any archive must discharge in order to conform to the OAIS model:

- "Negotiate for and accept appropriate information from information Producers.
- Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation.
- Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided.
- Ensure that the information to be preserved is Independently Understandable to the Designated

OAIS term	Description
Producer	The individuals and organisations who create the data and deliver data to the archive
Consumer	The individuals and organisations who use data once it has been disseminated by the archive
Designated Community	A group of potential consumers who should be able to understand and use data which is archived and disseminated by a specific archive. Data may need packaging differently according to whether the designated community is a subject specific group or the wider public
Submission Information Package (SIP)	These are the data which are delivered to the archive by the producer. This 'package' will consist of both the data files themselves and any metadata supplied by the producer to help describe and document the data
Archival Information Package (AIP)	These are the data held in the archive once preservation work has been carried out. Again they will consist of both content and metadata
Dissemination Information Package (DIP)	These are the data (and documentation) that have been prepared for dissemination to consumers

Table 1. A selection of key OAIS terms defined (see CCSDS 2002 section 1.7.2 for definitive glossary).

Community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information.
– Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original.
– Make the preserved information available to the Designated Community."
(CCSDS 2002, 3-1)

To be OAIS conformant an archive also needs to support the model of information described in section 2.2 of the ISO standard (CCSDS 2002, 1-3). Section 2.2 details the 'Information Packages' that are stored in an OAIS (see *Table 1* for descriptions of the three different information packages), though it is stressed that individual implementations of this model again are flexible (CCSDS 2002, 2-3–2-7).

4. OAIS and the Archaeology Data Service

The digital archive at the Archaeology Data Service (ADS) was established in 1996 several years prior to the acceptance of the OAIS model as an ISO standard. ADS archival procedures and policies have evolved over time as the organisation itself, and the wider world of digital archiving, has grown and matured. When OAIS was brought to our attention it was thought to be a useful exercise to establish whether it was a model which ADS could successfully adopt and to which its activities could be mapped.

A key step in doing this was to examine the 6 mandatory responsibilities as defined within the OAIS standard and to establish how far these were carried out by the ADS.

4.1. Negotiates for and accepts information

The responsibility for negotiating deposit of data sets is fulfilled by the ADS Collections Development Manager who acts as the interface between the data producer and the ADS. The ADS has a Collections Policy defining the categories of data that will be considered for accessioning (available online at http://ads.ahds.ac.uk/project/collpol.html). At the most basic level, collections have to be related to archaeology in some way, but the Collections Policy specifies the geographical, chronological and thematic scope of our collections. If someone contacts ADS with a view to archiving their data, the first thing to establish is whether the subject matter complies with the Collections Policy. However, the negotiations go further than this. The Collections Development Manager will also give advice on file formats, file naming strategies, documentation and metadata to ensure that any data deposit is suitable for archiving.

4.2. Obtains sufficient control for preservation

The ADS needs to have sufficient control of the data to be able to effectively carry out its archival work. There would be no point in a Producer depositing a batch of data to preserve if they didn't also grant ADS permission to migrate their files into newer file formats in order to create digital objects more suitable for preservation.

Again it is the job of the Collections Development Manager to discharge this responsibility. With each new deposit to the archive, they must ensure that it is covered by a submission agreement signed by the data producer. A sample of the ADS submission agreement (or deposit licence) can be downloaded from the ADS website (http://ads.ahds.ac.uk/project/userinfo/deposit_guidelines/deposit_how.cfm). Section 4 is one of the key sections of this agreement, giving the ADS the appropriate rights to work on the data. By agreeing to this, the producers are allowing ADS to distribute the data in various formats, to catalogue, enhance and validate the data, to document it and most importantly to "electronically store, translate, copy, or re-arrange the Data Collection to ensure its future preservation and accessibility".

ADS would not attempt to carry out any archiving work on a Submission Information Package until a Submission Agreement is in place.

4.3. Determines designated consumer community

The concept of the 'designated community' is key to the OAIS model. This is another way of referring to the audience for an archive — the people who the archive is targeting.

Responsibility for this rests with the ADS User Services Manager. It is the job of the User Services

Manager to liaise with the data consumers. They manage the ADS helpdesk and are the first point of contact for those wishing to use ADS resources. Part of the job is to provide support for groups of existing and potential users and to give information and advice on the range of services and collections available. The User Services Manager must be familiar with the needs of the ADS user community.

4.4. Ensures information is independently understandable

To ensure that information is independently understandable the ADS needs to ensure that all the data ingested into the archive is clearly and fully documented. There is little point in preserving data that is badly organised, full of unidentified codes and abbreviations and undocumented. The chances of anyone being able to successfully reuse this data would be slim.

An obvious example of this problem is that of a coded database. Archaeologists frequently produce databases that contain coded or abbreviated data. A well structured database may contain the necessary code-breaking information in a linked lookup table, but in some cases this information may be omitted from the database and its documentation, rendering it impossible to reuse.

As an archive, the ADS needs to ensure that the data it is preserving and presenting to its designated community makes sense on its own. ADS does not want a situation where data consumers have to go back to the data producer in order to ask questions before they can begin to interpret and work with the archived data. In the long term, the data producer may be unavailable or unable to remember enough about a project to answer the questions. It is the job of the archive to ask all of these questions and ensure that all necessary information is delivered to the consumers.

This responsibility is carried out by the Curatorial and Technical Team at the ADS. As they start to ingest new material into the archive, examining the Submission Information Package and preparing it for archive and dissemination, it is their job to ensure all the data is independently understandable and to contact the data producer where further documentation or clarification is needed. ADS has also published a series of Guides to Good Practice which provide further advice on the metadata and documentation required for a range of specific archaeological data types in order to make them suitable for preservation and re-use. These are available as publications in hard copy or via the ADS website (http://ads.ahds.ac.uk/project/goodguides/g2gp.html).

4.5. Follows established preservation policies and procedures

The ADS also has a large number of internal procedures and policies that guide and inform its work in preserving data. These range from the generic 'Repository Operations' document, describing how and where to store all the elements that make up the Archival Information Package, to the more specific 'Data Procedure Documents' that go into the finer details of preservation policy for each type of file that might be received as part of an archive. The ADS currently has Data Procedure for a diverse range of files, from raster images, binary text, and databases to virtual reality, photogrammetry and lidar.

It is the team of Curatorial and Technical Officers who are responsible for writing these policies and keeping them up-to-date. As technology changes rapidly it is important to ensure they are reviewed on a yearly basis.

4.6. Makes the information available

The data held by ADS are disseminated to the designated community online through a variety of tailored web interfaces (http://ads.ahds.ac.uk/). These range from pages that simply list a series of downloadable files (with documentation and usage instructions), to searchable online interfaces into individual databases, and interactive maps for querying spatial datasets online.

All of these resources are available to consumers free of charge once the ADS Terms and Conditions of access are agreed (as defined in our Copyright and Liability Statement http://ads.ahds.ac.uk/copy.html and Common Access Agreement http://ads.ahds.ac.uk/cap.html).

The ADS believes that there is little point in preserving data unless they are reused and so part of its role is to maximise data reuse. Once the web interface for a new archive is ready to release, publicity work is carried out by the User Services Manager, who targets the designated community through e-mails, newsletters and newsfeeds and informs them about new resources as they become available.

5. Future challenges

One of the strengths of the OAIS model is its flexibility. Conversely this also becomes one of its weaknesses when one attempts to put it into practice. OAIS conformance is a difficult thing to measure. At the ADS we certainly carry out the 6 mandatory responsibilities of an OAIS, can apply the concept of Information Packages to our archive and can map our staff and activities to the OAIS model, but it is still hard to state absolute compliance.

One of the initial goals of CCSDS when they started to work on OAIS was to create a solid foundation for future standards-building activities and this is certainly another of its strengths. Published in February 2007 and building firmly on the footings of the OAIS comes an initiative called '*Trustworthy Repositories: Audit and Certification. Criteria and checklist*' (known as TRAC) (OCLC and CRL 2007). This initiative looks at the issue of certification of digital archives and lists 84 individual criteria that an archive has to demonstrate they meet in order to become a 'Trusted Digital Repository'. Unlike the OAIS model it is very prescriptive about what the responsibilities of an archive should be. Though still relatively new and lacking the necessary certification bodies in order to enable formal TRAC audits to take place it is certainly something that digital archives should be aware of. The ADS has carried out an initial self-audit and the results have been useful, highlighting current strengths, as well as the areas where more work is needed. Over the coming years ADS intends to address these issues in order to ensure that precious digital data are preserved into perpetuity.

References

Ashley, Kevin (2008). *The MS Office 2003 format debacle*. University of London Computing Centre digital archives blog (available online at http://dablog.ulcc.ac.uk/2008/01/11/the-ms-office-2003-format-debacle/)

Consultative Committee for Space Data Systems (2002). *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1 Blue Book.

Lavoie, Brian F. (2004). *The Open Archival Information System Reference Model: Introductory Guide*. DPC Technology Watch Series Report 04-01 (available online at http://www.dpconline.org/docs/lavoie_OAIS.pdf)

Online Computer Library Centre and the Centre for Research Libraries (2007). *Trustworthy Repositories: Audit and Certification. Criteria and checklist*. Version 1 February 2007.