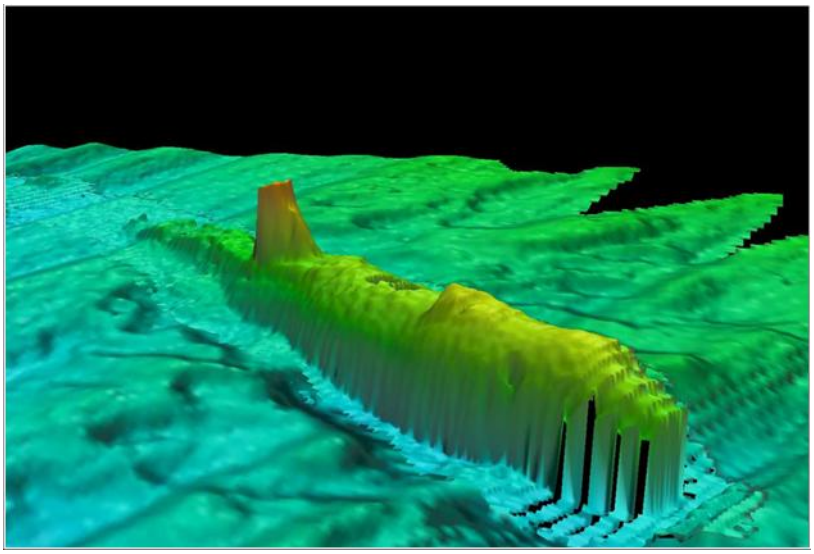
Preservation and Management Strategies for Exceptionally Large Data Formats: 'Big Data'

(EH Project No: 3984)



Tony Austin and Jenny Mitcham
Final 1.03
28 September 2007

Contents

List of figures

1 Introduction

It is a self-evident truth that archaeologists push the boundaries of available computing resources in the course of their work. Access to an exponential increase in computing power has allowed archaeologists to investigate and use a range of technologies that were developed in other disciplines such as the Earth Sciences.

Specifically, the ever increasing storage capacity of digital media allows archaeologists to work with larger and larger datasets. Not so long ago to talk of megabytes of data seemed awesome. Today the gigabyte (1000 megabytes) is becoming a relatively common term with some research teams even working with terabytes (1 tb = 1000 gigabytes) of data as is the case with the **North Sea Palaeolandscapes** project[1].

For the purpose of this report Big Data technologies include

- **Lidar** (Light Detection and Ranging or Laser Imaging Detection and Ranging) data which is generated by timing laser pulses from an aerial position (plane or satellite) over a selected area to produce a surface mapping. lidar data tends to be commercial but data centres of organisations like the Natural Environment Research Council (NERC) hold some datasets.

- **3D Laser Scanning** which is similar to lidar but within a terrestrial environment with the scanner local to the object of interest. A well known application of laser scanning technology is the survey undertaken of the stones at Stonehenge[2]

- **Maritime survey** covers an almost bewildering range of techniques many of which have been embraced by archaeologists specialising in this area. Techniques include sidescan sonar, sub bottom profiling, multi beam bathymetry, single beam bathymetry, single beam acoustic ground discrimination sonar (AGDS), acoustic tracking and various magnetic techniques.

- **Digital Video** where footage is increasingly created during archaeological projects.

[1] http://www.archant.bham.ac.uk/research/fieldwork_research_themes/projects/North_Sea_Palaeolandscapes/

[2] http://www.stonehengelaserscan.org/

That there are special problems associated with the curation and reuse of such datasets was discussed at a Heritage 3D[3] workshop in November 2004 and at a previously arranged 'Big Data day' a few days later in York. The latter, hosted by Dr Jon Kenny of the Archaeology Data Service (ADS), was in response to the growing concerns within the Commissions Team at English Heritage about very large datasets being generated by some projects; particularly those funded through the Aggregates Levy Sustainability Fund (ALSF)[4]. Subsequently English Heritage commissioned the ADS to investigate **Preservation and Management Strategies for Exceptionally Large Data Formats** or as it has become commonly known the '**Big Data**' project[5]

A range of approaches were developed targeting not only archaeological researchers but a wider community including practitioners and specialists from data services. Approaches included

- **Literature research** looked at current best practice amongst a wide range of archival specialists and data centres used to working with very large datasets (footnotes and references herein and in other Big Data outcomes)

- A **Questionnaire**[6] identified organisations generating or working with 'big data', categorising this data and investigating its potential for re-use.

- A **Workshop**[7] debated the results from the questionnaire and beyond and introduced the Big Data case studies.

- A **Formats review**[8] of software packages and technologies used to work with Big Data.

- **Case studies**[9] of archaeological projects producing 'big data' were used to examine practical problems arising from the transfer, curation and dissemination of large datasets. These included

[3] http://www.ceg.ncl.ac.uk/heritage3d/

[4] http://alsf.defra.gov.uk/

[5] http://ads.ahds.ac.uk/project/bigdata/

[6] http://ads.ahds.ac.uk/project/bigdata/survey.html

[7] http://ads.ahds.ac.uk/project/bigdata/workshop.html

[8] http://ads.ahds.ac.uk/project/bigdata/formats.html

[9] http://ads.ahds.ac.uk/project/bigdata/caseStudies.html

- **Breaking through Rock Art Recording**: An Arts and Humanities Research Board (AHRB) funded project undertaken by Durham University to investigate the potential of 3D laser scanning. The sites recorded included Castlerigg, Long Meg and her Daughters, the Copt Howe panel in Cumbria and Horseshoe Rock in Northumbria[10]

- **Wrecks on the Seabed**: This project led by Wessex Archaeology examined ways of assessing and evaluating wreck sites using a wide range of maritime survey techniques including video. The work will help understand the effects of marine aggregate dredging on shipwrecks. The project is funded by the Aggregate Levy Sustainability Fund (ALSF) administered by English Heritage[11]

- **Where Rivers Meet**: Another ALSF funded project being undertaken by the University of Birmingham investigating the landscape, ritual, settlement and the archaeology of river gravels at the confluence of the Trent and Tame Rivers in Staffordshire. The project utilised a wide range of data sources including lidar imagery, Geophysical survey and aerial photography[12]

[10] http://www.dur.ac.uk/prehistoric.art/btrar/btrar.htm

[11] http://www.wessexarch.co.uk/projects/marine/alsf/wrecks_seabed/

[12] http://www.iaa.bham.ac.uk/research/fieldwork_research_themes/projects/whereriversmeet/

2 Archival strategies at large: the context

Organisations with responsibility for the long term preservation and management of digital data will or should have well documented archival strategies and procedures in place. Other organisations act as advisory bodies. Documentation can range from generic policy statements such as the NERC Policy Handbook[13] through to the quite specific, for example, a series of Preservation Handbooks produced by the Arts and Humanities Data Service (AHDS) and its subject specific data centres[14]. Other organisations providing useful documentation in terms of strategies and procedures include the UK Data Archive (UKDA)[15], the British Library[16], the Library of Congress[17], the National Library of Australia[18], the United Kingdom Hydrographic Office (UKHO)[19], NASA's National Space Science Data Centre (NSSDC)[20], the Electronic Resource Preservation and Access Network (ERPANET)[21], The Digital Preservation Coalition (DPC)[22] and the Digital Curation Centre (DCC)[23]. Whilst often organisationally specific some generic themes emerge from the available information including the emergence of the International Organization for Standardization (ISO) standard Open Archival Information System (OAIS) and the increasing take up of Lifecycle Management as an archival strategy.

2.1 OAIS

The development of the OAIS reference model has been pioneered by NASA's Consultative Committee for Space Data Systems (CCSDS) It has recently been accepted as an ISO (14721:2003) standard[24]. A technical

[13] http://www.nerc.ac.uk/research/sites/data/documents/datahandbook.pdf

[14] http://ahds.ac.uk/preservation/ahds-preservation-documents.htm

[15] http://www.data-archive.ac.uk/

[16] http://www.bl.uk/about/collectioncare/digpresintro.html

[17] http://www.digitalpreservation.gov/

[18] http://www.nla.gov.au/padi/

[19] http://www.ukho.gov.uk/amd/ProvidingHydrographicSurveys.asp

[20] http://nssdc.gsfc.nasa.gov/

[21] http://www.erpanet.org/

[22] http://www.dpconline.org/

[23] http://www.dcc.ac.uk/

[24] http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=24683&ICS1=49&ICS2=140&ICS3

recommendation is also available for consultation on the CCSDS website[25]. As a reference model OAIS provides a conceptual framework within which to consider the functional requirements for an archival system suited to the long term management and preservation of digital data. Such consideration can be given to both proposed and to existing systems. The model is also seen as a way of comparing systems through mapping discipline specific jargon to OAIS terminology and that such terminology is clear and unambiguous enough to allow understanding by those beyond dedicated archival staff. The core entities and work flows within the model are shown in fig. 1

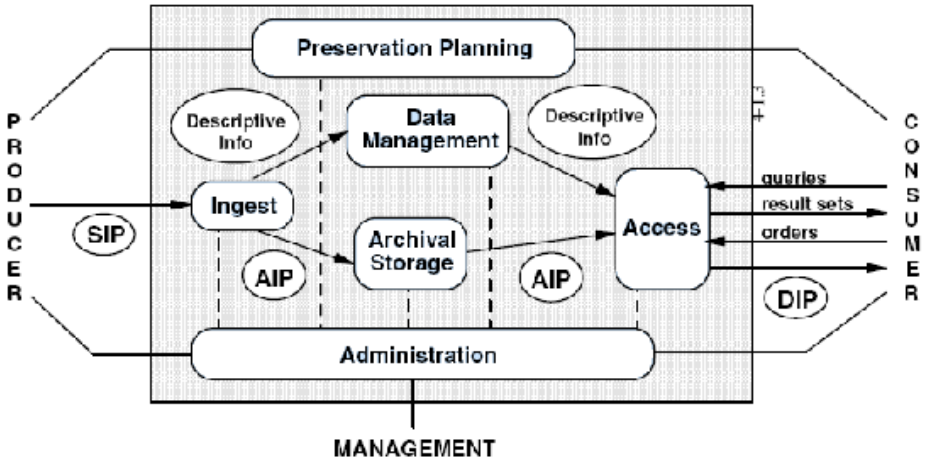


Fig. 1 OAIS Functional Entities (after CCSDS Fig.4.1[13])

Data producers create Submission Information Packages (SIP). A SIP equates to a deposit of digital data plus any documentation and metadata necessary for the archive to facilitate the long term preservation of the data and to provide access for consumers (i.e. reuse). The SIP provides a basis for the creation of an Archival Information Package (AIP) and a Dissemination Information Package (DIP) generated by the archive. The process involves generating preservation and dissemination versions of the deposited data where necessary. For example, a Microsoft® Word file might be migrated to an XML based format such as an Open Office text document for long term preservation and to PDF for dissemination. Metadata documenting this processing is added to the AIP as is any relevant information from the SIP. Similarly any resource discovery metadata and reuse documentation in the SIP is added to the DIP. Consequently metadata and documentation supplied as part of a SIP assume major importance in terms of data deposition. The OAIS standard notes of the SIP that 'Its form and detailed content are typically negotiated between the Producer and the OAIS'[26]. In practice most repositories offer guidelines to depositors about acceptable formats, delivery

[25] http://public.ccsds.org/publications/archive/650x0b1.pdf

[26] ibid, page 2-7

media, copyright issues and necessary documentation and metadata. Many existing guidelines will be relevant to Big Data but particular issues that have arisen are discussed more fully below.

The most recent development is the publication of a certification document **Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist**[27] by the US based Research Libraries Group (RLG) part of the Online Computer Library Center (OCLC), the Center for Research Libraries (CRL) and the National Archives and Records Administration (NARA). The purpose of the checklist is identifying repositories capable of reliably managing digital collections. The audit checklist is closely tied to the OAIS reference model in terms of a conceptual framework and terminology and considers organisational suitability, repository workflows, user communities and usability of data, and the underlying technical infrastructure including security. All of these areas must be openly documented. Organisations that can demonstrate that they meet the criteria within the checklist will be identified as Trusted Digital Repositories.

The CRL is currently undertaking a project to test the RLG-NARA metrics through actual audits of subject digital archives and one archiving system'[28]. A study exploring how the audit checklist can be applied to the management policies derived from a system based on DSpace digital asset management software in combination with the distributed data management software, Storage Resource Broker (SRB) has been undertaken[29]. The ADS was used as a case study to examine the relationship of an established repository to the OAIS model in a series of workshops undertaken by the Joint Information Systems Committee (JISC) funded Digital Preservation Training Programme (DPTP)[30].

In general the archival community are rushing to claim or are actively seeking compliance with the reference model through this process of certification. It should; however, be noted that the audit checklist is very recent development. As such for the time being a state of trust needs to exist between creator and archive.

2.2 Lifecycle management

Whilst there are other archival strategies OAIS conformance with its emphasis on ongoing management and administration of a digital resource implies an object lifecycle. At a recent (2006) conference **The LIFE Project: Bringing**

[27] http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=9

[28] http://www.crl.edu/content.asp?l1=13&l2=58&l3=142

[29] http://sils.unc.edu/events/2006jcdl/digitalcuration/Moore_Smith-JCDLWorkshop2006.pdf

[30] http://www.ulcc.ac.uk/dptp/about-dptp.html

digital preservation to life[31] Neil Beagrie in a paper entitled 'The LIFEcycle model, from paper to digital' discussed the evolution of lifecycle management from its beginnings in publications such as the Terotechnology Handbook (1978)[32] which considered lifecycle costing and the idea of 'total cost of ownership' for physical objects. Subsequently during the 1990s the AHDS and the British Library and others built on this approach for digital assets. He noted how the early involvement of the JISC and the AHDS with project proposals through the provision of guidance and advice helped to reduce costs downstream. One manifestation of this was noted as the publication of a number of AHDS Guides to Good Practice[33].

By 1998 lifecycle frameworks for managing digital resources had become well defined as described, for example, by Beagrie and Dan Greenstein in **A Strategic Policy Framework for Creating and Preserving Digital Collections**[34] and the subsequent development of this framework into a cost model by Tony Hendley in a British Library Research and Innovation Report (106)[35]. The Life Project final report provides a more recent and detailed methodology for calculating 'the long-term costs and future requirements of the preservation of digital assets'[36]. The report will undoubtedly feed into many archival policies.

The generally recognised categories of the lifecycle of digital assets are

- Data creation
- Acquisition, retention or disposal
- Preservation and management
- Access and use

These categories and elements within them will provide the framework for the rest of this report.

2.3 Other strategies

[31] http://www.dpconline.org/graphics/join/lifeconfrep.html

[32] Terotechnology Handbook (1978) HMSO

[33] http://www.ahds.ac.uk/archaeology/creating/guides/index.htm

[34] http://www.ukoln.ac.uk/services/papers/bl/framework/framework.html

[35] http://www.ukoln.ac.uk/services/elib/papers/tavistock/hendley/hendley.html

[36] http://eprints.ucl.ac.uk/archive/00001854/01/LifeProjMaster.pdf

The OAIS model described above implies a preservation strategy based on migration. An ideal is to move data to a software-independent format and subsequently migrate this through successive technical infrastructures over time (known as refreshment). There is without doubt a preference within the archival community to migrate to the most stable of all formats; ASCII text which is an international standard of long standing; however, this is often not an option as with images for example. In such cases version and format migration is practiced. Files in such formats are also subject to periodic refreshment. It should be noted that this is not the only preservation strategy. Alternatives include technology preservation and emulation.

2.3.1 Technology preservation

Here the data is preserved unchanged along with the technology (hardware and/or software) upon which it depends. Clearly there are problems with such a strategy as technology will fail over time and replacement becomes increasingly difficult and more costly. Jeff Rothenberg (1999) notes the problems associated with this reliance on 'computer museums'[37]. The ADS attempts to maintain a 'computer museum' but not to effect technology preservation, rather in a probably vain hope of facilitating data recovery from outdated media[38] although some of the 'exhibits' have been used in earnest!

2.3.2 Emulation

Rothenberg favours emulation as an alternative preservation strategy[39]. It is seen to have particular relevance where the look, feel, and behaviour of a data resource is of importance. Critiques of emulation include that it is still in its infancy in terms of development, that it is likely to be more costly than the implementation of a migration strategy, that there are likely to be software copyright issues and that (the original) software and hardware is rarely documented to a high enough level to allow subsequent emulation[40]. An interesting and confusing development came about during the CAMiLEON project which developed a strategy called 'Migration on Request'[41] which in fact is emulation with a tool being built to process the original byte stream of a digital object on request.

---

[37] http://www.clir.org/PUBS/reports/rothenberg/pub77.pdf (section 6.3)

[38] http://ads.ahds.ac.uk/project/museum/

[39] ibid (section 8)

[40] http://www.dpconline.org/graphics/orgact/storage.html

[41] http://www.si.umich.edu/CAMILEON/reports/mor/

Interestingly it was recently decided to move the interactive video created in 1986 by the BBC to celebrate the 900th anniversary of the Domesday Book from its dependence on outmoded media and computer hardware. Numbers of experts were approached including the CAMiLEON project who 'argued that the slight faults in images as displayed from the <original> analogue discs were a part of that experience, and should not be cleaned up' but the National Archive 'wanted to preserve the data with the highest quality available consistent with longevity' and hence opted for migration[42].

Comments and recommendations

The long term preservation and dissemination of Big Data (indeed any data) should ideally be within an OAIS compliant framework (ISO 14721:2003 standard) [p 9 -11]

Because the certification metrics are very new many archives are currently working towards OAIS compliance. As such trust must exist between creator and archive [p11]

The Submission Information Package or SIP assumes major importance in the relationship between data producer and an OAIS compliant archive where as well as the data; documentation and metadata inform on preservation and reuse [p 10]

[42] http://www.ariadne.ac.uk/issue36/tna/

3 Data creation

This will normally involve a design phase followed by an implementation phase in which the data is created or acquired.

During the design phase the future of the data to be created should be given ample consideration. Where the potential for reuse is considered worthwhile data must be in, or have migration paths to, formats suitable for long term preservation and dissemination. Also it will be essential to develop documentation including metadata to facilitate this. This would be considered good practice even when reuse is not an issue. In short the Submission Information Package or SIP is a meaningful concept even before the lifecycle of a digital resource begins.

3.1 The Big Data community (questionnaire)

There were a total of 48 respondents to the online questionnaire. This might appear small in number but Big Data technologies are highly specialised.

As indicated by organisational affiliation, interest in the Big Data project is not restricted to one discipline. As well as Archaeology the wider Humanities and Earth Sciences were represented. Amongst respondents 73% expressed an interest in joining a representative directory and interest group and have been contacted on occasion via a Big Data email list; bigdata_list@ads.ahds.ac.uk

3.2 What is Big Data? (workshop)

Clearly this relates to the resources available to organizations involved with the data. The threshold of what is problematic because it is big is going to be much lower within Archaeology than, for example, the Earth Sciences which is much better resourced. Whilst memory gets cheaper long term curation does not. Currently tens of gigabytes are probably problematic for archaeologists in terms of accessibility and long term storage (availability) but the goal posts are always moving to the larger both in terms of storage availability and datasets in use.

Development at the ADS reflects this. Back in the late 1990s the size of resources were thought of in kilobytes (= 1000 bytes). Today megabytes (= 1000 kilobytes) are the norm with the occasional resources in gigabytes (= 1000 megabytes) but as noted in the introduction to this study archaeologists are already working with terabytes (=1000 gigabytes). To put this into a wider context companies such as Google are already working with hundreds of

petabytes[43]. A petabyte =1000 terabytes of data which is 10^{15} or 1,000,000,000,000,000 bytes[44].

Respondents to the Big Data questionnaire were asked how big a typical project might be (Q2). Some 75% of respondents are typically concerned with up to 100 GB of data with 50% below 50 GB. Frightening and reassuring at the same time in that most projects are within the lower range of what might be considered 'big data'. It is daunting that nearly 20% of respondents are handling datasets over 200 GB in size; however, on closer examination these appear to be mostly archives.

3.3 Big Data technologies

As abstracted from the project case studies what were thought likely to be the core 'big data' technologies was largely confirmed by responses to the online questionnaire. Note that data capture and pre-processing of data can be within a project or as supplied by external organisations including commercial companies.

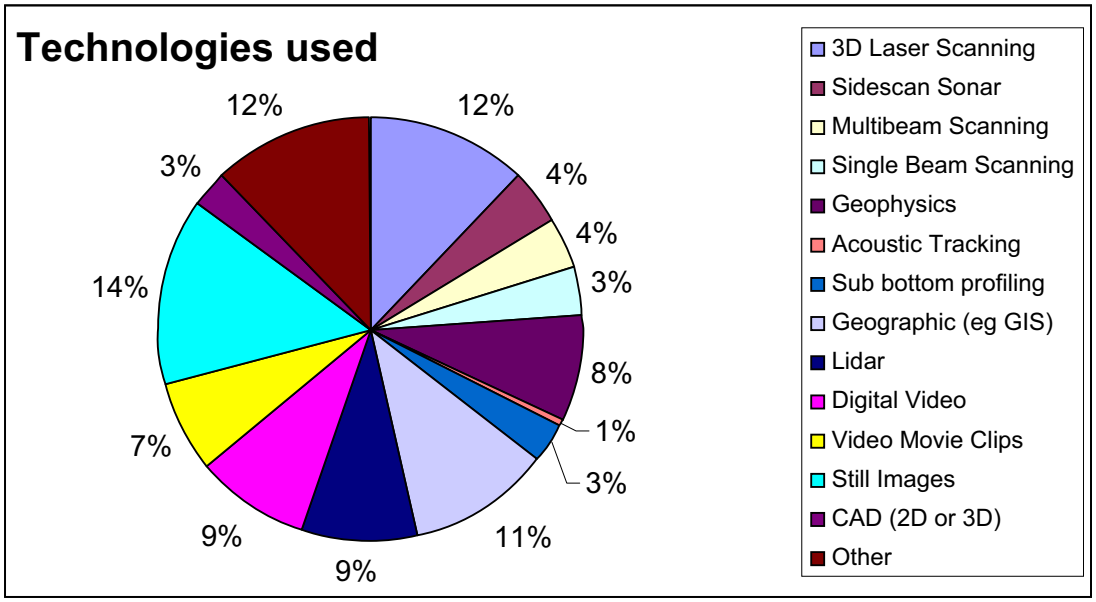


Fig. 2 Data capture techniques (Questionnaire Q1)

Respondents to the questionnaire were also asked what software packages were used when working with Big Data. Of the 101 software packages noted a staggering 52 are unique (that is after editing for things like lower and upper

[43] http://www.wired.com/wired/archive/14.10/cloudware_pr.html

[44] http://en.wikipedia.org/wiki/Petabyte

case character differences). It seems the world of 'big data' is very fragmented. Fig. 3 notes packages mentioned more than once.

Fortunately software packages tend to support multiple formats both in terms of import and export. For example, graphics packages usually support multiple image formats. A JPEG (.jpg) file can be opened and saved as a TIFF (.tif) file which is widely recognised as an archival standard for images.

Clearly, some of the packages noted below may rarely produce Big Data but are used during the analysis and presentation stages of a project. Examples include video clips.

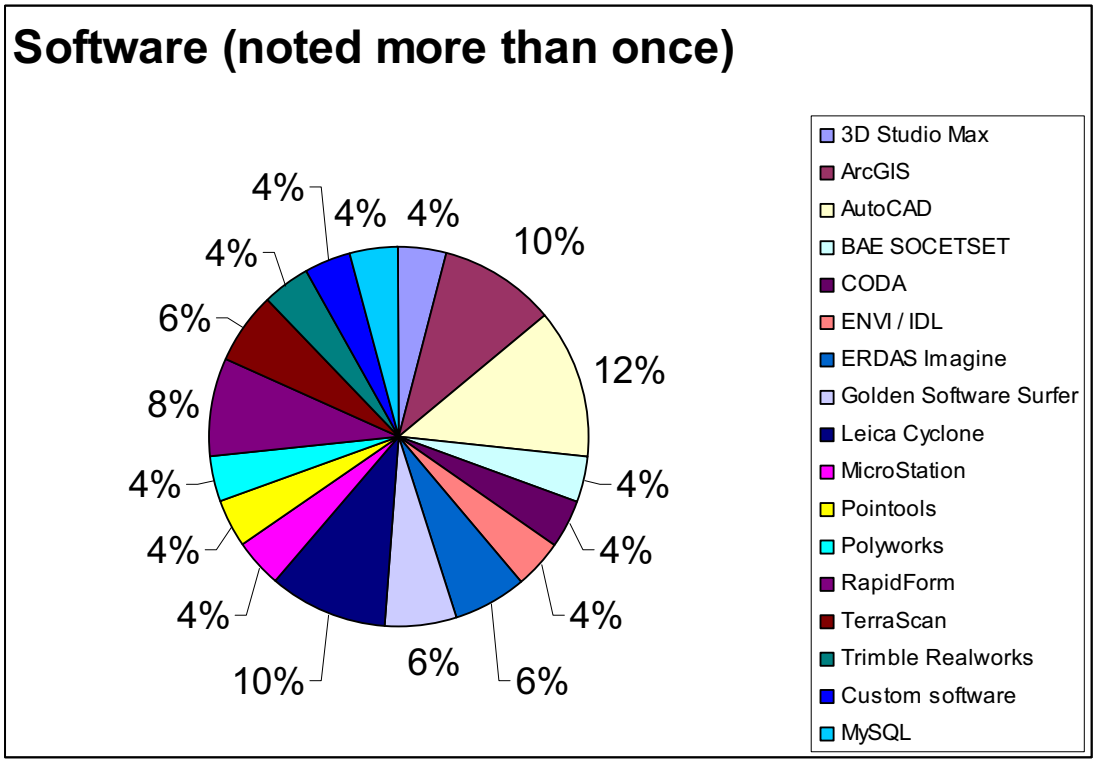


Fig. 3 Software packages used for data capture and analysis

Along with the case studies the above analysis of technologies and the software packages supporting them has helped to identify some of the formats being used for Big Data. Just as the preceding chapter on archival strategies applies equally to all resources considered for archiving many formats have the potential to be Big Data, for example, a digital image library could easily be gigabytes in size. Whilst many of the conclusions reached here would apply equally to such resources this study is particularly concerned with Big Data formats in use with technologies such as lidar surveys, laser scanning and maritime surveys.

Many of the software packages associated with Big Data are both proprietary and produce binary (executable) files. Binary files are generally not seen as the best solution for long term preservation except where such a format is a

well established standard such as TIFF. Over 80% of the packages being used by respondents to the Big Data questionnaire create binary files (Q3). Fortunately nearly 50 % either use or can export data as ASCII text.

During the Big Data project, especially the workshop, it became very apparent that a contradiction exists between users and archivists or curators of large datasets. Users expressed a preference for binary data in openly published formats because file sizes are significantly smaller which makes handling and exchanging data easier. It was clear that representatives from data centres preferred data as ASCII text, generally seen as the most stable of standards, for preservation purposes within a long term archival strategy. This is resolvable in many cases through normal archival practice where dissemination or data exchange versions of a file can differ from the preservation version. For example, standard ADS practice is to migrate a Microsoft® Word document to an XML based Open Office document for preservation and to binary PDF for dissemination.

A very interesting and recent development is the move by many software producers towards XML (eXtensible Markup Language) based formats or at least an XML format export facility. Beyond packages such as Open Office[45] and Microsoft Office 2007 (Office Open XML)[46] the Geospatial Data Abstraction Library (GDAL/OGR)[47] is a cross platform C++ translator library for raster and vector geospatial data formats that is released under an X/MIT style Open Source license by the Open Source Geospatial Foundation[48]. In short GIS files such as ESRI Shape files and MapInfo files can be migrated to an alternative supported format such as the XML-based Geography Markup Language (GML) [49]. Following testing including reverse engineering this is close to adoption by the ADS as a preservation strategy for GIS data such as ESRI shape and MapInfo files.

GIS data is generally in the form of a vector graphic which is essentially a series of XYZ coordinates defining an image. Other data can be associated with each coordinate. Other formats associated with Big Data share a commonality in being vector graphics. This commonality may provide an archival solution for some Big Data formats in the future in that it should be possible to build on software such as the open source GDAL GIS libraries to support similar vector-based formats and exports to GML. For a recent discussion of raster and vector graphics see the **Digital Image Archiving Study**[50] undertaken by the AHDS for the JISC.

---

[45] http://xml.openoffice.org/

[46] http://office.microsoft.com/en-us/help/HA100069351033.aspx

[47] http://www.gdal.org/index.html

[48] http://www.osgeo.org/

[49] http://www.opengis.net/gml/

[50] http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf

There are a number of reasons why a format recognised as an open standard might be unsuitable for archiving. Formats using lossy compression (where data is lost as part of the compression process) are generally seen as unsuited[51]. An open standard needs to be well and widely supported before it can be considered as a reliable preservation format. Even if a format is an open standard the available software to read it might be proprietary and expensive which can inhibit the potential for reuse.

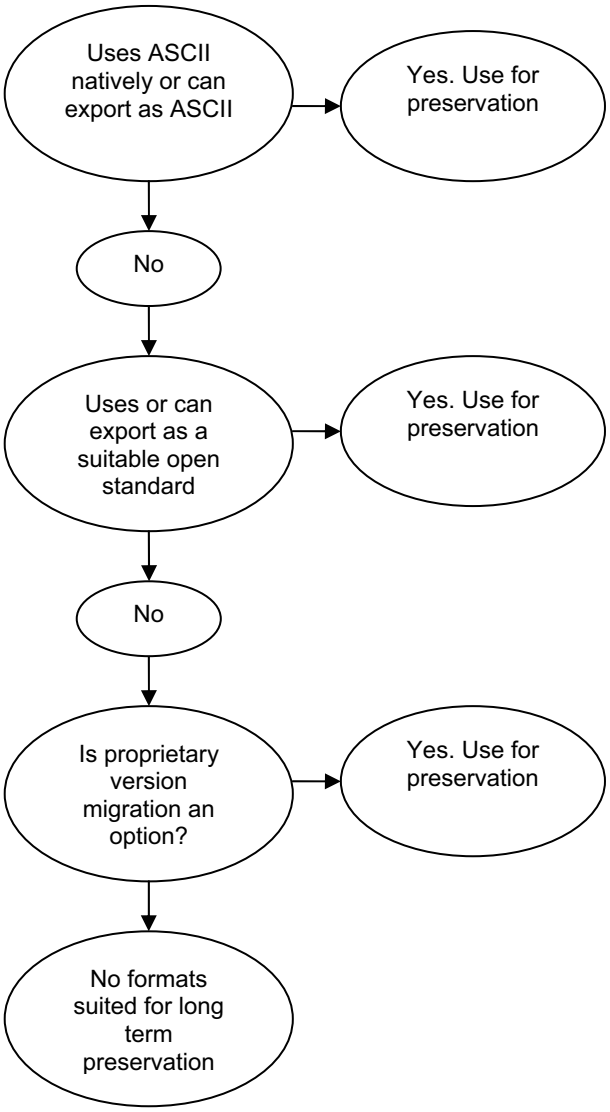


Fig. 4 A simplified software package preservation formats decision tree

[51] http://www.nationalarchives.gov.uk/documents/image_compression.pdf (currently a draft)

Published proprietary formats may not always be as open as they seem. For example, the Drawing eXchange Format (DXF) was developed to facilitate the movement of Computer Aided Design (CAD) drawings between packages. AutoDesk the vendors of AutoCAD® and the maintainers of the DXF specification consistently failed over a long period to keep it publicly up to date which was problematic for other CAD vendors trying to provide support. This has recently been rectified with the DXF specifications for recent versions including AutoCAD® 2008 available for download[52]. It should also be noted that some proprietary formats do develop into open standards. For example, Adobe® recently announced that they have begun the process of turning their very popular Portable Document Format (PDF) into an ISO standard[53]. Migrating through newer versions of a proprietary software package is the least preferred preservation strategy because it is an ongoing resource hungry process; especially so where the software in question is expensive to purchase.

If long term preservation and reuse are implicit goals data creators need to establish that the software to be used or toolsets exist to support format migration where necessary. A lengthy review of formats was undertaken as part of the Big Data project. Those with potential for preservation or data sharing are discussed below (sections 5 and 6).

3.4 Documentation

As already noted data along with documentation including metadata make up a Submission Information Package (SIP). Documentation is one of the cornerstones of archival practice and should exist even in-house within a project in order to facilitate management of associated data. The process of documentation should be actively pursued from the outset of a project as it is often difficult to create retrospectively. The relevance of documentation has been questioned as information is often implicit within the files themselves; however, this does not facilitate resource discovery, reuse and data management. Two of the Big Data case studies did not have adequate documentation or the resources to create it retrospectively with the consequence that their archives cannot be adequately managed. It is reasonable to suggest that to some degree this can be put down to a lack of existing guidance and documentation standards for the technologies being used by these projects which, of course, the Big Data project is trying to address.

[52] http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=8446698

[53] http://www.adobe.com/aboutadobe/pressroom/pressreleases/200701/012907OpenPDFAIIM.html

3.4.1 Metadata

Metadata can be used to document different aspects of a project at different levels. The process and the reasons for creating metadata are well documented in, for example, the AHDS Guides to Good Practice[54]. Although quite dated today they still have much of relevance. The recent AHDS Digital Image Archiving Study notes ISO (19115) Standard for Geographic Information – Metadata[55] and ISO 19139, the XML schema implementation, is the 'ultimate metadata' for GIS data[56]. The relevance here is wider in that the Standard can encompass any geospatially referenced dataset.

ISO 19115:2003 defines:

- mandatory and conditional metadata sections, metadata entities, and metadata elements
- the minimum set of metadata required to serve the full range of metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data)
- optional metadata elements - to allow for a more extensive standard description of geographic data, if required
- a method for extending metadata to fit specialized needs.

The UK GEMINI geospatial metadata standard[57] developed jointly by the Association for Geographic Information (AGI) and the Cabinet Office e-Government Unit is compliant with ISO 19115 and should be a suitable container for describing Big Data most of which is spatially referenced vector survey data. Each discrete dataset within a project should have a corresponding metadata record; a lidar survey, magnetometer survey, and so on.

UK GEMINI has already been adopted by major organisations. The Marine Data and Information Partnership (MDIP), for example, note that their standards are based on the UK GEMINI Standard[58]. Some examples of MDIP GEMINI metadata can be found on the SEEGrid website[59]. The Multi-Agency Geographic Information for the Countryside (MAGIC) project notes its

[54] http://ads.ahds.ac.uk/project/goodguides/g2gp.html

[55] http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020&ICS1=35

[56] http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf (section 7.6)

[57] http://www.gigateway.org.uk/metadata/standards.html

[58] http://www.oceannet.org/mdip/working_groups/interop_wg.html

[59] https://www.seegrid.csiro.au/twiki/bin/view/Marineweb/MarineDataandInformationPartnership

intention to adopt UK GEMINI as 'its baseline standard'[60]. Thinking on harmonisation of UK GEMINI with the Infrastructure for Spatial Information in Europe (INSPIRE) project's Draft Implementing Rules for Metadata is already underway. The INSPIRE draft standard[61] maps to ISO 19115 so this should be unproblematic.

Information that appears important to the successful management and reuse of Big Data that does not obviously fit into a UK GEMINI structure was identified during the technologies and formats review. This includes metadata about the equipment used and any settings, software used, methodology employed and an assessment of the accuracy of the data. Much of this may fit into the UK GEMINI Abstract element of which the specification notes in terms of usage

- State what the 'things' are that are recorded
- State the key aspects recorded about these things
- State what form the data takes
- State any other limiting information, such as time period of validity of the data
- Add purpose of data resource where relevant (e.g. for survey data)
- Aim to be understood by non-experts
- Do not include general background information
- Avoid jargon and unexplained abbreviations.

Alternatively the Additional Information Source element could be used to point to associated documentation (below) such as a brief survey overview. The lack of a relation element in the UK GEMINI metadata set could be seen as a shortcoming. Such information could also be recorded in the associated documentation pointed to in the Additional Information Source element. UK GEMINI does support a Lineage element which can be used to record 'information about the events or source data used in the construction of the dataset'. The latter is of particularly importance in the case of distributed archives where source data and derived datasets might be archived with different organisation. Lineage; however, is only one of number of possible relations a digital object or dataset might have.

[60] http://www.magic.gov.uk/progress.html

[61] http://www.agi.org.uk/SITE/UPLOAD/DOCUMENT/policy/draftINSPIREMetadataIRv2_20070202.pdf

Other forms of metadata are associated with good archival practice as, for example, indicated in the OAIS Reference Model[62]. The model describes the Preservation Description Information (DPI) package which consists of 'Content Information and which can be categorized as Provenance, Reference, Fixity, and Context information'. Areas within this are covered through the adoption of metadata standards such as UK GEMINI but file level metadata such as fixity values and provenance which includes 'processing history' need addressing.

Provenance information is concerned with 'history' and records, for example, 'the principal investigator who recorded the data, and the information concerning its storage, handling, and migration'. Reference information is concerned with unambiguously identifying content information through, for example, the provision of an ISBN number for a publication. Context information in terms of OAIS is concerned with environment. Examples include 'why the Content Information was created and how it relates to other Content Information objects'.

A fixity value or checksum 'is a form of redundancy check, a simple way to protect the integrity of data by detecting errors in data that are sent through space (telecommunications) or time (storage)'[63]. The MD5 (Message-Digest algorithm 5) and the SHA (Secure Hash Algorithm) are widely used cryptographic hash functions. Applying these algorithms to a file produces an (almost certainly) unique hash or checksum value and will consistently produce this value if a file is unchanged. Thus it provides a mechanism for validating and auditing data. Security weaknesses have been identified in MD5 but this is unlikely to be a problem unless data is sensitive. Utilities such as FastSum[64] which generates MD5 hashes and File Checksum Integrity Verifier (FCIV)[65] which supports both MD5 and SHA-1 are freely downloadable (note these are Windows DOS utilities but similar exist for Unix based systems including Linux and in many cases are pre-installed). Both these examples support batch processing.

An isolated checksum is of course of no use on its own. It has to be associated with a file, a location, a project and a survey as structured data

File Metadata	Comments	Example data
UNIQUE_ID	Auto-generate – unique	1234567
FILE_LOCATION	Directory + filename	/adsdata/cottam_ba/jpg/fwking_plan.jpg
CHECKSUM_TYPE	MD5, SHA-1, etc	MD5
CHECKSUM_VALUE	Generated by algorithm	578cbb18f73a885988426797bcab8770
PROJECT_ID	Unique project ID	ADS-123

[62] http://public.ccsds.org/publications/archive/650x0b1.pdf

[63] http://en.wikipedia.org/wiki/Checksums

[64] http://www.fastsum.com/

[65] http://support.microsoft.com/kb/841290

SURVEY_ID		Laser_05-Jun-2003
GENERATED		16-May-2006
GENERATED_BY		Austin, T
LAST_AUDITED		16-May-2007

This is suggested as a minimum. The ADS, for example, generate file size, file last modified date, format (file extension), file version and other data for management purposes. It obviously needs to be maintained rigorously to be useful.

Maintaining a process history is an essential if tedious part of archival practice. An example would be importing XYZ data into a GIS. Again this can be recorded as simple structured data. The same structure can hold both file level and batch processing information. The following example is based on AHDS practice

Process metadata	Comments	Example data
PROCESS_ID	Auto-generate – unique	1234567
PROJECT_ID	For example a survey ID	PRO-453
SOURCE_FORMAT		xyz
DESTINATION_FORMAT		shp
PROCESS_AGENT	Who did the processing	Mitcham, J
PROCESS_COMMENTS		Referenced to WGS84
PROCESS_START_DATE		17-May-2007
PROCESS_COMPLETION_DATE		17-May-2007
PROCESS_DESCRIPTION		Import of XYZ data into ArcView for analytical purposes and dissemination as research outcome
PROCESS_GUIDELINES		None
PROCESS_HARDWARE_USED		Viglen Genie Intel Pentium 4
PROCESS_SOFTWARE_USED		ESRI Arcview 9.1
PROCESS_INPUT		/adsdata/pro-453/xyz/file.xyz
PROCESS_OUTPUT		/adsdata/pro-453/shp/file.shp
PROCESS_RESULT		Success
PROCESS_TYPE	See below	Conversion - dissemination
ADDED		18-May-2007
ADDED_BY		Austin, T

The AHDS model restricts process types to a defined list (a lookup table) which should work within a wider setting.

Process Type
Capture
Conversion - preservation
Conversion - dissemination
Editing - Corrective
Editing - Aesthetic
Creation - documentation
Creation - metadata
Other Event

3.4.2 Other documentation

This consists of anything that will facilitate preservation and reuse of a dataset. It could, for example, be published reports, brief grey literature reports or even a few scanned pages from a notebook. These might provide information missing from, supportive of, or more detailed than metadata records. They can often provide further contextual information about how a dataset fits together. A good example of this is a standard practice for preserving databases where data is exported to delimited ASCII text. This would become very difficult to reuse at a later date without supporting documentation describing the structure of the database in the form of an Entity Relationship Model (ERM) and the structure of each table in the form of a Data Dictionary.

Documentation may have particular relevance to Big Data where a number of survey techniques involve a series of traverses over a spatially defined area (see formats review). Composite mosaics can be produced as either part of acquisition or as part of post processing. In the latter case it clearly critical how data from each traverse relates to the others. The possibility exists to use the Additional Information Source element in the UK GEMINI to point to such information. A robust and adhered to file naming convention can also reinforce this.

3.5 Archival strategy

Less than half of the respondents to the Big Data questionnaire had any sort of archival strategy in place (Q5). A number noted that they were awaiting the outcomes of the Big Data project for guidance.

The preceding sections attempt to define a minimal basis for two of the cornerstones of an archival strategy for projects where data is seen to have a post-project relevance

- Use of software supporting formats with clear migration paths for both preservation and reuse
- The creation of adequate documentation to facilitate this as well as supporting in house administration and management during the project

The other cornerstones are

- Access to an adequate hardware system
- A robust backup strategy in place

Data storage is largely unproblematic for most projects with terabyte external hard drives available for under £300. Back up strategies have been well discussed elsewhere. The FISH (Forum on Information Standards in Heritage)

Fact Sheet no. 1 'A Six Step Guide to Digital Preservation' provides a brief overview[66]. Archival organisations invest heavily in backing up data. For example the ADS subscribes to the University of York backup service which uses Legato Networker and an Adic Scalar Tape Library[67] and also maintains copies of data in the AHDS central repository which in turn is backed up to tape. A basic strategy for a project could; however, be as simple as a couple of high capacity hard drives with one stored off site in a fairly inert environment. These would need synchronising on a regular basis with the master data.

Comments and recommendations

In order to effectively undertake the long term preservation and dissemination of Big Data (indeed any data) archival organisations need a well formed Submission Information Package (SIP) [p 17 -22]

Consideration must be given to software and the formats it supports during data creation. Where long term reuse is a goal there must be clear migration paths for both preservation and reuse [p 13 -17]

In general ASCII text is seen as the most stable format for data preservation whilst open binary formats suit the dissemination of Big Data because of a dramatic reduction in file size [p 15]

Inadequate documentation during data creation is the single biggest barrier to the future reuse of data. Documentation including metadata facilitates reuse as well as supporting in-house administration and management during a project [p 17]

It is recommended that the UK GEMINI metadata standard which is compliant with the ISO (19115) Standard for Geographic Information is used to describe survey data. Further, maintenance of provenance and fixity metadata is identified as a crucial part of data creation [p 18 -19]

Any other documentation that may facilitate reuse should also be included in the SIP [p 22]

[66] http://ads.ahds.ac.uk/newsletter/issue19/fishsheet1.pdf

[67] http://www.york.ac.uk/services/cserv/offdocs/keynotes/oct01.pdf

4 Acquisition, retention or disposal

Once the data creation and analysis phases are complete a final decision as to whether a dataset is suitable for long term preservation and dissemination needs to be made be it in house or with an external archive. Agreement has to be reached between the data creator and archive on a number of issues

- Does the data fit into an archive's collection policy?
- Is it fit for purpose?
- Is it sufficiently documented?
- What to archive?
- What will it cost?

The process of ingest is generally well documented with archival organisations or data centres providing, for example, collections and charging policies, guidelines and FAQs. A well formed Submission Information Package will aid the actual process of ingest but there are a number potential problem areas pertaining to Big Data.

4.1 Guidelines

The Big Data project was commissioned because of a lack of guidance about Big Data technologies for the archaeological community. Existing archival guidance may not cover such technologies. Much can be learned from non-archaeological organisations that have experience of these technologies as detailed in the Big Data formats review. Thus consideration should be given to producing a Guide to Good Practice for the archaeological community based on this report and related documents.

4.2 Retention and disposal

The question of what to preserve was discussed in depth at the Big Data workshop. This is relevant to all data but particularly so for Big Data because of the file sizes involved. Raw (or the rawest available – acquired data has often been pre-processed) data is deemed important. As long as processing history is fully documented (see section 3) and repeatable it seems unnecessary to keep intermediate data. The fully processed data is the archaeological outcome that can be manipulated and re-examined within suitable software. It thus also has reuse value. Decisions about retention or deletion will be ongoing throughout the lifecycle of a resource. For example, datasets may be superseded or no longer have reuse value.

4.3 Cost of archiving

So long as (1) data are deposited in recommended formats (or alternatively that there are migration paths to convert them to such formats), (2) sufficient documentation is provided, and (3) ingest procedures are established, the processing of the Big Data case studies suggested that their archiving was no more consuming in terms of human resource than the archiving of any other data. Clearly, because of the physical size of the data, it takes much longer to move files around, for example, when moving from delivery media into an archival environment. Similarly, confirming the success of the transfer through generating fixity or checksum values is a much longer process because each byte in a file is referenced. Both of these processes can; however, be run as background tasks. By definition the physical storage requirements of Big Data will be greater than a more conventional dataset. Consequently, Big Data archives would, for example, fit comfortably within the current ADS charging policy[68] where storage is charged by the megabyte and ingest costs are based on the number of files that make up a resource.

It should be noted that 'storage' encompasses the ongoing periodic process of data refreshment (see 2.3). In order to take advantage of technological advances and decreasing costs in certain areas archives have to periodically upgrade systems or parts thereof. As an example, in its 10 year history the ADS recently moved to its third generation of equipment. Thus it is operating on a five year upgrade cycle. This is expensive both in terms of equipment and staff time. The long term cost of storage is often difficult to conceptualize but a dataset maintained for 100 years would go through 20 refreshments based on the five year cycle noted above. There is no reason why certain digital datasets should not be maintained for such a period. After all many of our most valued paper archives are of considerable antiquity.

It is noted above that Big Data would fit within the current ADS charging policy. Recent developments have, however, initiated a review of this policy. The first development is the rise of interest in lifecycle modelling with its emphasis on retention and discard policies. This is a clear break with the tradition of trying to preserve everything for as long as possible. The second and related development pertains to Big Data itself in that, whilst it can be accommodated within existing models, it would appear astonishingly expensive to archive. The requirement within the Big Data project was to retain the archives derived from the case studies for five years only (project design 3.2.6). On a five year refreshment cycle data from these archives might only be refreshed once before being discarded. Clearly this needs to be reflected in any charges for archival services. This could provide the basis of a cost model reflecting the retention period of a resource in that archiving a resource for 10 years would require refreshing twice, a period of 15 years would see three refreshments and so on.

[68] http://ads.ahds.ac.uk/project/userinfo/charging.html

From experience the cost of refreshment for a given resource decreases with time as archival systems become more sophisticated and a given archive becomes an increasingly smaller part (presuming archival growth) of a periodic refreshment. Thus there is a gradual decrease in the cost of refreshing a given resource although this is partially offset by the increasing cost in terms of human resource (i.e. increasing wages). Between refreshments the ongoing management and administration within an OAIS framework is proactive and similarly subject to increasing costs in terms of human resource.

In contrast the cost of physical disc storage and back up media such as tape decreases rapidly. Currently the cost of a gigabyte of disc storage can be as low as 7p. Analysis of past and current trends suggests this will be 1p in five years time and so negligible not long after that to be considered as zero cost[69]. However, the capital cost of the systems associated with such storage can be substantial as can ongoing maintenance, backup and insurance costs. Like disc storage systems they consistently fall in price but still remain a significant cost over time.

The test of time suggests that so far the one off 50p per megabyte charge in the current ADS charging policy is near the mark for an earlier archival tradition. Recent developments, however, in terms of systems upgrades suggests the 50p charge can be reduced significantly. The 'per megabyte' charge is shorthand for what has been described above which might be better described today as 'ongoing management and refreshment'. The following is simplistic but attempts represent more accurately the current situation of lifecycle management with its associated retention and discard policies

Retention period	Cost for refreshment
5 years	R + E
10 years	R – DR + E - DE
15 years	R – 2DR + E - 2DE
20 years	R – 3DR + E – 3DE
25 years	R – 4DR + E – 4DE

Where R = refreshment cost
DR = decreasing cost of refreshment
E = cost of physical equipment
DE = decreasing cost of equipment

As an example, if R = 9p, DR = 3p, E = 4p and DE = 1p (all pence per megabyte charges - please note these figures should be close to a final policy but are subject to an ongoing examination of past processes) then

[69] http://www.berghell.com/whitepapers/Storage%20Costs.pdf

Retention period	Cost for refreshment	Cumulative total (pence)
5 years	9 + 4 = 13	13
10 years	9 - 3 + 4 – 1 = 9	22
15 years	9 – 6 + 4 – 2 = 5	27
20 years	9 – 9 + 4 – 3 = 4	28
ongoing		30

The above one off costs suggests that preservation costs become negligible after 20 years. This is, to a degree, a product of the simplicity of the model as clearly there will be ongoing costs beyond this point in terms of the refreshment, management and administration of a resource should a retention policy dictate it. Thus a one off charge of 30p per megabyte would cover ongoing preservation beyond 20 years. ADS policy is currently based on the assumption that 'best efforts' will be used to preserve all data deposited with ADS into perpetuity (i.e., following the 20-year cost-model above). However, in some cases it is possible that funding agencies may no longer require preservation beyond a specified period, which might be subject to review at regular intervals. A number of possible reasons to discard a dataset exist including that only a specified period of preservation was required, that it has been superseded or included in another resource, that it is no longer considered to have value and that there is no practical way to continue its preservation. It is envisaged that any potential discard will need to be confirmed by the ADS Advisory Committee (or any body that might succeed it). The committee comprises members drawn from the archaeological community at large[70].

As well as ongoing management and refreshment accessioning an archive involves a significant investment during ingest; the process of structuring and moving files. This process also requires extensive documentation to facilitate ongoing preservation and reuse. Ingest may also require the transfer of files into suitable preservation and dissemination formats if they have not been delivered as such. The cost of ingest is estimated separately from ongoing management and refreshment with current thinking suggesting standard charges for small, medium and large archives for which definitions are currently being refined. A further charge may apply where significant numbers of files need to be migrated from delivery formats.

The above describes a possible ADS approach to lifecycle management costs. Alternatives no doubt exist. The model is agreed internally but precise figures may change as part of the ongoing charging policy review. Its adoption will also necessitate updating general preservation policies. An example of costs under both the current charging policy and that currently under revision are detailed in Appendix A.

[70] http://ads.ahds.ac.uk/project/advisory.html

4.4 Copyright

That copyright was a major issue first became apparent at the Big Data workshop. Survey data is often owned by third parties. For example, the Where Rivers Meet case study acquired survey data from the satellite aerial data vendor, Infoterra[71] who retain copyright. Clearly an external organisation cannot archive this without the permission of the copyright holder which is unlikely to be forthcoming if the data still has a commercial value. Optimistically this could be seen as a distributed archive with the third party archiving the raw data and organisations such as ADS holding the derived data. The problem with this; however, is that reuse by anyone other than Birmingham University who undertook the Where Rivers Meet project would require purchasing the data. Interestingly, Infoterra did offer alternatives to the above scenario[72]. Firstly

> 'Should anyone such as a local archaeological trust wish to purchase the data for further work, Infoterra would offer the data at x0.5 of the original sale price. This data would then be available for use by any members of the trust'

And secondly

> 'should someone such as English Heritage wish to purchase the data on behalf of the wider archaelogical community, and therefore be able to disseminate the data free of charge (seek no commercial gain) for 'Academic' work, we would be willing to provide the data at x1 the cost of the original data. NB copyright of the data would still remain with Infoterra Limited, but EH (or managers of the data) would be granted a license to hold the data and disseminate on request'

The latter suggests that rather than funding the purchase of data by a project a better model would be for English Heritage (and presumably similar organisations) to purchase the data and to reduce project funding accordingly. The data could then also be made available to other research projects. There are of course cost implications for the archiving and dissemination of this raw data. According to the data audit of the Where Rivers Meet project it had 6565 MB (6.5 GB) of lidar data as ASCII text. As an example, under the current ADS charging policy this would cost

Item	Cost	Comment
20 ASCII files at £25	£500.00	
30 ASCII files at £2.50	£75.00	(bulk discount after the first 20 files)

[71] http://www.infoterra.co.uk/

[72] email correspondence with infoterra global

6565 MB at 0.50 per MB	£3282.50	
Total cost	**£3857.50**	**VAT not included**

If English Heritage had purchased the Where Rivers Meet lidar data under the model described above at Infoterra's quoted price of £5,825 it would take just one project to reuse the data for it to be a cost effective model. Other vendors of survey data may offer similar models.

4.5 Data transfer

The transfer of Big Data has been suggested as problematic. Without doubt it is going to be more involved than burning CDs or DVDs as for a conventional resource but the cost of high capacity external hard drives has been dropping dramatically with, for example, one terabyte drives available for under £300. Delivery media can of course be supplied or returned. Network transfers are discussed below (section 6).

Comments and recommendations

Consideration should be given to producing a Guide to Good Practice for the archaeological community based on this report [p 24]

Only raw (or the rawest available) data and project outcomes should be considered for archiving [p 24]

Big Data can be accommodated in existing lifecycle cost models but concepts of retention and discard suggest revision may be necessary. An ongoing revision of the ADS charging policy is introduced [p 25 - 27]

Third party copyright is problematic for reuse; however, consideration should be given be alternative models offered by vendors such as Infoterra [p 27 -28]

5 Preservation and ongoing management

As noted already data in the Submission Information Package (SIP) should be in, or have migration paths to, suitable preservation formats and the associated documentation be sufficient to support the creation of an Archival Information Package (AIP) 'consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS' and the content information defined as the 'set of information that is the original target of preservation. It is an Information Object comprised of its Content Data Object and its Representation Information. An example of Content Information could be a single table of numbers representing, and understandable as, temperatures, but excluding the documentation that would explain its history and origin, how it relates to other observations, etc' and the PDI as the 'information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, and Context information'[73]. That some of this information needs to be supplied by the data creator has been discussed above (section 3.4).

With the provision of a well formed SIP an archive will have minimal problems in generating the AIP. It is the rich metadata that provides for the ongoing management of the data it references through, for example, the automated audit of data using fixity or checksum values or through migration as a batch process.

The following table summarises a sample of Big Data formats that are considered to have applicability for long term preservation. It was abstracted from the Big Data formats review. The review contains information about a wide range of formats including toolkits and software libraries that can help with format migration. Neither the table or the review are exclusive in that there are undoubtedly other formats suited to preservation, other formats associated with the technologies under consideration and other Big Data technologies not considered.

Format/Properties/ Technologies	Description	Comment
ASCII text (.txt, .dat, .xyz, etc) Published standard ASCII Raw (logger)	In an example of this provided by Wessex Archaeology the raw data was collected using a data logger as structured ASCII text and incorporated into a database. There are well established archival procedures for databases in exporting tables as delimited ASCII text and documenting through an Entity Relationship Model (ERM) and a Data Dictionary.	Preserve as ASCII text with support documentation.

[73] http://public.ccsds.org/publications/archive/650x0b1.pdf (1.7.2 TERMINOLOGY)

DXF: Drawing eXchange Format (.dxf) Proprietary published (currently) ASCII and binary Processed usually 3D including Point cloud CAD Mesh	Published and maintained by AutoDesk vendors of AutoCAD. Was seen for a long time as a *de facto* standard for the exchange of CAD files[74] but then Autodesk stopped publishing (after v. 12) for DXF associated with new versions of AutoCAD. They have; however, recently published the standard for AutoCAD 2008 and several previous versions[75]. Version migration has been seen as the only real way of securing the long term preservation of CAD material; however, use of GDAL/OGR is a possible (as yet untested) strategy (see GML below). Also see OpenDWG, IGES and STEP as described in the recent Digital Image Archiving Study[76]. These emerging standards are not well supported in terms of tools as yet and are thus not recommended here.	ASCII DXF and version migration still seem to be the best preservation option but other options are emerging.
GML: Geography Markup Language (.gml) Published standard[77] ASCII Processed Geospatial data Including GIS CAD	XML (and hence ASCII) based standard for geospatially referenced data. This encoding specification was developed and is maintained by the Open Geospatial Consortium (OGC). The Ordnance Survey (OS) supply MasterMap® mapping data as GML[78]. Many GIS packages including ESRI and MapInfo products now support GML. The emergence of the Geospatial Data Abstraction Library (GDAL/OGR) is starting to provide the means to easily migrate geospatial data into formats such as GML for preservation and data exchange[79].	GML is ideally suited for preservation and data exchange of geospatial data.
MGD77 (.mgd77) Published ASCII Raw or can be Geophysical data including	Developed by the US National Geophysical Data Center (NGDC) following an international workshop in 1977[80]. Revised relatively recently. Described by UNESCO thus 'It has been sanctioned by the Intergovernmental Oceanographic Commission (IOC) as an accepted standard for international data exchange'[81]. The MGD77CONVERT toolset	In being ASCII based and published could act as a preservation format. Has support as a data exchange format.

[74] Walker, R. (ed.) 1993. *AGI Standards Committee GIS Dictionary*. Association for Geographic Information

[75] http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=8446698

[76] http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf

[77] http://www.opengis.net/gml/

[78] http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/information/technical/gml2.html

[79] http://www.gdal.org/index.html

[80] http://www.ngdc.noaa.gov/seg/gravity/document/html/mgd77.shtml#general

[81] http://ioc.unesco.org/iocweb/iocpub/iocpdf/tc045.pdf

[82] http://www.soest.hawaii.edu/GMT/gmt/doc/html/mgd77convert.html

Bathymetric Magnetic Gravity	allows conversion to the binary NetCDF format[82] which offers an alternative and smaller means of dissemination.	
MPEG 1 (.mpg, .mpeg) Published open standard[83] Binary Processed usually Video Audio	An International ISO/IEC (11172) developed by the Moving Picture Experts Group (MPEG) for Video CD (VCD) and less commonly DVD-Video. Provides reasonable quality audio/video playback comparable to VHS tape. The MPEG-1 Audio Layer III equates to MP3 audio. Many tools exist for working with this sort of data exist including the open source MediaCoder which is described as 'universal audio/video batch transcoder distributed under GPL license, which puts together lots of excellent audio/video codecs'[84].	Suitable for preservation and data exchange.
MPEG 2 (.mpg, .mpeg) Published open standard[85] Binary Processed usually Video Audio	As MPEG-1, an ISO/IEC (13818) standard but for DVD as well as various flavours of TV. 'MPEG-2 video is not optimized for low bit-rates (less than 1 Mbit/s), but outperforms MPEG-1 at 3 Mbit/s and above'[86] and hence much higher quality.	Suitable for preservation and data exchange.
MPEG 4 (.mp4) Published open standard[87] Binary Processed Video Audio	Another MPEG ISO/IEC (14496) standard concerned with 'web (streaming media) and CD distribution, conversation (videophone), and broadcast television, all of which benefit from compressing the AV stream'[88].	In being an online streaming standard could be used for data sharing.
NTF: National Transfer Format (.ntf) Published standard ASCII Raw and	Complex ASCII based storage and transfer format for vector and raster images (same extension). Largely used by the OS for distributing pre-MasterMap data (see GML). It is a British Standard BS 7567 'Electronic Transfer of Geographic Information'[89]. A wide range of NTF converters are available to, for example, popular GIS formats. Lidar data as supplied has often	In being ASCII based and published it should be suited for both transfer and preservation. Unclear; however, as to

[83] http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=25371

[84] http://mediacoder.sourceforge.net/

[85] http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=37679&ICS1=35&ICS2=40&ICS3=

[86] http://en.wikipedia.org/wiki/MPEG-2

[87] http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=38559

[88] http://en.wikipedia.org/wiki/MPEG-4

[89] http://www.bsistandards.co.uk/shop/products_view.php?prod=6536

processed Geospatial data including Point cloud CAD Digital Elevation Models (DEM) Lidar	been processed in terms of coordinate transformation and decimation.	how wide its usage is outside of the OS where it is being superseded by GML.
NetCDF: Network Common Data Form (.nc) Published Binary Raw or can be Scientific data including Bathymetric Lidar and others?	NetCDF 'is a set of software libraries and machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data'[90].Openly published[91]. Libraries freely available under licence. Tools include ncgen and ncdump which respectively generate from and dump to ASCII. Also supports the sub-setting of datasets. Appears widely used for scientific including bathymetric data, for example, the NERC British Oceanographic Data Centre (BODC)[92].	This could provide an ideal mechanism for preservation and data sharing through storing once and generating binary or ASCII as requested.
OBJ (.obj) Published ASCII Raw data or can be 3D including Laser scanning Mesh Point cloud Photogrammetry	A simple ASCII based format for representing 3D geometry. Initially developed by Wavefront Technologies. The format is apparently open and has wide support amongst both software vendors and open source community. Whilst the format specification is available on numerous websites[93] we were unable to identify a format maintainer. There are numerous converters available for OBJ files.	Wide support suggests a possible data exchange format. In being ASCII based it could act as a preservation format.
TFW: TIFF World file (.tfw) Proprietary ASCII (but associated image will be binary) Processed ESRI GIS products (others?)	A mechanism for geo-referencing images developed by ESRI (GIS software vendor). As such similar to GEOTIFF (see above) but in this case the metadata is held in a separate ASCII text file[94]. TIFF World files will be small in themselves but may be associated with large images.	That the metadata (spatial information is in ASCII could be seen as good for preservation.

[90] http://www.unidata.ucar.edu/software/netcdf/

[91] http://www.unidata.ucar.edu/software/netcdf/docs/netcdf/File-Format-Specification.html#File-Format-Specification

[92] http://www.bodc.ac.uk/data/online_delivery/gebco/

[93] http://people.scs.fsu.edu/~burkardt/txt/obj_format.txt

[94] http://support.esri.com/index.cfm?fa=knowledgebase.techArticles.articleShow&d=17489

GEOTIFF (.tiff) Public domain[95] Binary Processed GIS and other image processing packages	The GEOTIFF standard is in the public domain. It allows metadata, specifically georeferencing to be embedded within a TIFF image. There is complete conformance to the current TIFF 6.0 specification. As the recent **Digital Image Archiving Study** notes 'The use of uncompressed TIFF version 6 <as preservation format> is the best strategy at the current time, but a watching brief should be maintained on JPEG2000 as an emerging preservation format'[96]. TIFF is also a public domain format currently maintained by Adobe®[97]. It should be noted that the size of a TIFF file is limited to 4GB[98].	Despite being a binary format TIFF has long been recognised as a *de facto* preservation standard for raster images. Binary is currently the only real option for the bitstream encodings of raster images.
VRML (.wrl) Published open standard[99] ASCII Processed 3D graphics	Virtual Reality Modelling Language. As VRML 97 a published ISO (14772-1) standard for 3D vector graphics. Designed with the internet in mind. As such requires a plug-in or viewer[100]. Apparently still popular especially for the exchange of CAD drawings but is slowly being superseded by other standards such as X3D (below)	Possible exchange format. In being ASCII based has the potential to act as a preservation format but aging.
X3D (various) Published open standard[101] ASCII and binary flavours Processed usually 3D graphics	Developed as a replacement for VRML (above) by the web3D consortium[102] this ISO (19775) standard is XML based although a binary specification has been more recently released as an ISO (19776-3) standard. It is backwardly compatible with VRML. It is noted as being compatible with the MPEG-4 (above) specification. Like VRML requires a plug-in or viewer.	With XML being ASCII based this has archival possibilities.

[95] http://remotesensing.org/geotiff/spec/geotiffhome.html

[96] http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf 1.4.i

[97] http://partners.adobe.com/public/developer/tiff/index.html

[98] http://www.awaresystems.be/imaging/tiff/faq.html#q8

[99] http://www.web3d.org/x3d/specifications/vrml/

[100] http://vads.ahds.ac.uk/guides/vr_guide/sect37.html

[101] http://www.web3d.org/x3d/specifications

[102] http://www.web3d.org/x3d/

XML: eXtensible Markup Language (.xml or can be) Published open standard[103] ASCII RAW or processed Increasing range of technologies	XML[104] is a general-purpose markup language geared towards facilitating the sharing of data. An XML document is said to be 'well formed' when it conforms to XML's syntactical rules. It is described as valid when it conforms to semantic rules defined in a published schema. Many XML documents use a different file extension, for example .gml (see above). Others such as MIDAS XML developed by the Forum on Information Standards in Heritage (FISH)[105] are explicit in having the .xml extension.	Ideal for exchange and preservation if an established schema exists.
XYZ (.xyz .xyzrgb) ASCII (can be binary) Raw(ish) Laser scanning Lidar	Point cloud data - simply the X, Y and Z coordinates of each scanned point, sometimes with Red, Green and Blue colour values also. Lidar data may also have intensity values. XYZ data is sometimes decimated to make dataset more manageable. Depending on purpose this can often be done without discernable loss of detail. Lidar data as supplied has often been processed in terms of coordinate transformation and decimation.	ASCII text is seen as the best option for long term preservation along with suitable metadata.

Comments and recommendations

The provision of a well formed Submission Information Package or SIP is essential for the successful long term preservation of data [p 30]

That the data in the SIP is in, or has migration paths to, suitable Big Data formats for preservation is essential for the creation of the Archival Information Package or AIP [p 30]

The documentation including metadata in the SIP provides the basis of the framework for the successful ongoing management of the data [p 30]

[103] http://www.w3.org/XML/

[104] http://en.whttp://www.web3d.org/x3d/specificationsikipedia.org/wiki/Xml

[105] http://www.heritage-standards.org/

6 Access and use

Nearly 71% of respondents to the Big Data questionnaire reuse data at least once a year (Q6).

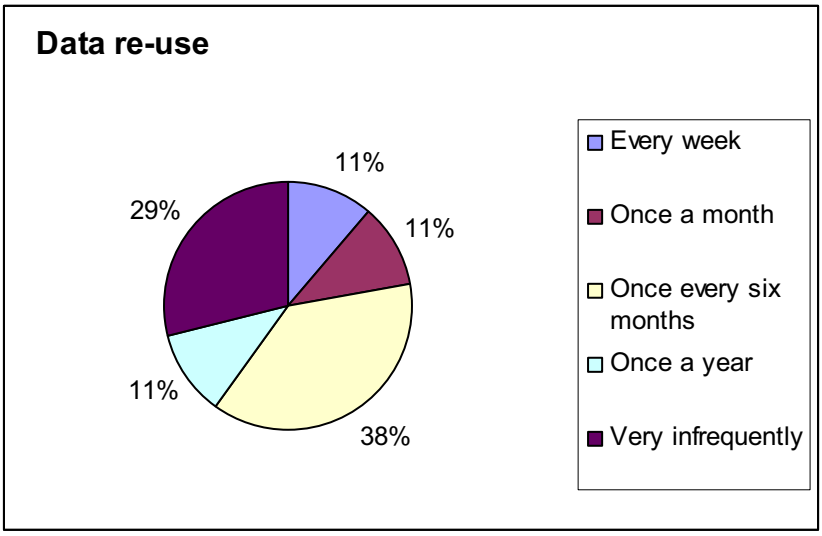


Fig. 5 Data reuse

Nearly 80% noted that they would allow access by others to their data (Q8) and over 80% stated that they had received large datasets from other organisations (Q9). 100% of respondents stated that they consider using existing datasets for a new project (Q10) and gave reasons such as

'A prerequisite to our work is to ensure duplication of data does not occur unless there is a justifiable reason for doing so'

'Yes, hoping to save time an money doing it. but the data often has to be very new and up-to-date, so it isn't always possible'

'…having such data available will assist any longer-term monitoring projects or even cast new light on a previously recorded subject'

'It is usual to refer to archive material of all kinds before embarking on new surveys'

Clearly there is both a strong desire to, and sound reasoning for, reuse of data.

6.1 Dissemination Information Packages

As described under preservation (section 5) data in the Submission Information Package (SIP) should be in, or have migration paths to, formats suitable for dissemination for reuse. The submitted format can in many cases be the same for both preservation and dissemination. The SIP needs to

contain any documentation that facilitates reuse including metadata relating to resource discovery, fitness for use, access, transfer and use (see section 3.4.1). A well formed SIP will facilitate the generation of the Dissemination Information Package (DIP)[106].

Many of the formats noted as suitable for preservation are also suitable for dissemination. This is the ideal situation; especially so for Big Data, as datasets need only be stored once; however, there is an already noted problem here in that archivists prefer ASCII whilst users prefer the smaller file sizes of binary files. Some formats have associated tools that would allow a file to be stored as ASCII and for a binary file to be automatically generated from it on demand. For example, the NetCDF format appears to support this scenario. The development of LAStoASCII and ASCIItoLAS tools would provide am ideal environment for this increasingly popular format

The following table notes formats considered to be suitable for disseminating data. These are additional to formats already noted as having suitability for preservation and dissemination. Again it is not exclusive

Format/Properties/ Technologies	Description	Comment
Generic Sensor Format (.gsf) Published[107] Binary Raw data Bathymetric	The Generic Sensor Format (GSF) is described as 'for use as an exchange format in the Department of Defense Bathymetric Library (DoDBL)'. The specification is currently openly published. As well as the generic it allows for attributes specific to a wide range of bathymetric surveying systems to be included.	Possible use as an exchange format if widely supported.
LAS (.las) Published[108] Binary Raw data or can be. Lidar Laser scanning	The LAS format is described as 'a public file format for the interchange of LIDAR data between vendors and customers. This binary file format is an alternative to proprietary systems or a generic ASCII file interchange system used by many companies'[109]. The American Society for Photogrammetry & Remote Sensing (ASPRS) endorses and supports the use of LAS along with industry stakeholders[110]. Discussions of extending LAS to additionally handle terrestrial laser scanning data are actively taking place[111]. A recent addendum to the English Heritage Metric Survey Specification covering laser scanning supports LAS as a data exchange and archival format for laser scanning[112] [113]. Such usage has not been formalised as yet.	Specifically designed for the exchange of data; a role for which it has strong support. In being a binary format would not be seen as ideal for a long term preservation role as ASCII text alternatives exist.

[106] see footnote 70

[107] http://www.ldeo.columbia.edu/res/pi/MB-System/formatdoc/gsf_spec.pdf

[108] http://www.lasformat.org/documents/ASPRS%20LAS%20Format%20Documentation%20-%20V1.1%20-%2003.07.05.pdf

[109] http://www.lasformat.org/

SDTS: Spatial Data Transfer Standard (various including .ddf) Published standard[114] Binary Raw data or can be Geospatial data DEM Terrain Image	An Earth Science standard developed by the USGS for data exchange. Downloaded files are a tarred (zipped) directory which as well as data contains numbers of DDF or data description files. Compliance with SDTS is a requirement for federal agencies in the US. Supports Raster and Vector data. There are large numbers of tools and translators for extracting data from SDTS to various formats. In some cases this involves extraction to earlier standards such as DLG[115] (see above) which suggests SDTS is a wrapper around other formats. GDAL (see GML above) support a SDTS Abstraction Library for geo-referencing[116].	Well supported as a data exchange standard but may be US centric.
SEG 2 (.sg2, .dat) Openly published[117] Binary Raw data Seismic survey including GPR: Ground Penetrating Radar	An update to various SEG formats including SEG Y by the Society of Exploration Geophysicists (SEG). Rather strangely there seems to be numbers of SEG 2 to SEG Y converters available. Does this mean SEG Y is still better supported? Seismic Unix is a popular freeware package for working with SEG and other seismic formats[118].	Possible exchange format. Export to ASCII with suitable metadata for preservation.

[110] http://www.asprs.org/society/committees/lidar/lidar_format.html

[111] www.ceg.ncl.ac.uk/heritage3d/downloads/TLS%20**formats**%20V1.pdf

[112] http://www.ceg.ncl.ac.uk/heritage3d/downloads%5Caddendum2006.pdf

[113] http://www.english-heritage.org.uk/server/show/nav.001002003003007001

[114] http://mcmcweb.er.usgs.gov/sdts/standard.html

[115] http://www.fws.gov/data/gisconv/sdts2av.html

[116] http://home.gdal.org/projects/sdts/

[117] http://www.seg.org/publications/tech-stand/seg_2.pdf

[118] http://www.cwp.mines.edu/cwpcodes/index.html

SEG Y (.segy) Published[119] Binary Raw data Seismic survey including Sub-bottom profiling Sidescan sonar GPR: Ground Penetrating Radar	An openly published format by the Society of Exploration Geophysicists (SEG). Originally (rev. 0) developed in 1973 for use with IBM 9 track tapes and mainframe computers and using EBCDIC (an alternative to ASCII encoding rarely used today) descriptive headers. The standard was updated (rev. 1) in 2001 to accommodate ASCII textual file headers and the use of a wider range of media. It should be noted that in the interim between revisions a number of flavours of SEG Y appeared trying to overcome the limitations of rev. 0. SEG Y to ASCII converters exist as, for example, made available by the USGS[120]. A limited functionality SEG Y viewer can be downloaded from Phoenix Data Solutions[121].	Can be converted to ASCII for preservation purposes. Possibly useful as a data exchange format as it appears widely supported.
eXtended Triton Format (.xtf) Proprietary but Publicly Available Specification[122] Binary Raw data or can be Sidescan sonar Sub-bottom profiling Bathymetric data	As described by the Triton Imaging Inc 'The XTF file format was created to answer the need for saving many different types of sonar, navigation, telemetry and bathymetry information. The format can easily be extended to include various types of data that may be encountered in the future'. Currently a Publicly Available Specification. Also described as an 'industry standard' for sonar. Some packages supporting XTF provide for ASCII text exports.	Possibly very suited for data exchange if industry support is widespread. Where possible ASCII text exports with suitable metadata would provide the best long term preservation environment.

6.2 Dissemination strategies

As with data transfer between creator and archive the dissemination of Big Data to a wider audience is often seen as problematic. The preference by users is for online access to file downloads. Whilst archival organisations are often hooked into high bandwidth systems many end users are not. For this reason the ADS, as an example, restricts file download sizes so users don't unwittingly affect their networks. On occasion larger files are made available for download by special arrangement for users known to have suitable connections. This may be one solution.

[119] http://www.seg.org/publications/tech-stand/

[120] http://pubs.usgs.gov/of/2005/1311/of2005-1311.pdf

[121] http://www.phoenixdatasolutions.co.uk/seisvu.htm

[122] http://www.tritonimaginginc.com/site/content/public/downloads/FileFormatInfo/Xtf%20File%20Format_X21.pdf

Other network technologies that were investigated included BitTorrent[123], a peer to peer (P2P) communications protocol for file sharing which appears to have possibilities as a means of distribution. To share a file an initial peer creates a 'torrent' which is a small file containing metadata about the file(s) to be shared, and about the computer that coordinates the file distribution which is known as the 'tracker'. When the first peers pick up the torrent and download the file(s) using BitTorrent clients they are expected as part of the process to become distributors of a small piece of the file(s). The tracker maintains a manifest of which peer has which part of a file and tells new peers where to download each piece. As the number of peers build up the load is increasingly shifted off the seed computer. Clearly the system needs peers or clients to have largely persistent network connections so that others can access the file fragments.

The above works very well with audio and video data that will have a high download usage and hence lots of potential peers. Research by CableLabs in 2006 suggests that 'some 18% of all broadband traffic carries the torrents of BitTorrent'[124]. This could provide a distributed archiving model; however, the reuse of Big Data is likely to be an occasional and limited activity with the consequence that BitTorrent is unlikely to provide an advantageous service where within a small community there will be limited downloads and thus peers. To quantify this file fragments are typically between 64 KB and 1 MB each; taking the upper value a 1 GB file would need 1,000 peers. There would be some advantage to the original seed but anyone attempting to reuse the data will experience even longer download times because of administration overheads.

High speed 'Point of Access' (PoA) optical networks and Grid Computing were also considered. UKLight 'is a national facility to support projects working on developments towards optical networks'[125] [126]. Data is transferred across dedicated 10 gigabit channels in a continuous stream rather than the conventional breaking down into small packets of data which are variously routed to their destination with a propensity for packet loss and the need to retransmit. As well as speed the dedicated channels mean other network users are unaffected in terms of bandwidth loss. The HP Vista Centre within the Institute of Archaeology and Antiquity at the University of Birmingham is a UKLight member[127] and is connected to the PoA in London. The ADS discussed the possibility of adding a spur to an existing UKLight connection with Computing Services at the University of York (where the ADS is based).

---

[123] http://en.wikipedia.org/wiki/Bittorrent

[124] http://www.multichannel.com/article/CA6332098.html

[125] http://www.uklight.ac.uk/

[126] http://www.managinginformation.com/news/content_show_full.php?id=3080

[127] http://www.rsconference.bham.ac.uk/documents/RSC06_IS_Computational___Visualisation_Support.pdf

The cost; however, of several thousand pounds prevented this proceeding further. Interestingly, although part of the academic network, UKLight is not exclusive in that collaborative projects within a wider community are considered. This may be worth investigating further as a way to link up academic and other archaeological organizations.

Grid Computing has a number of meanings[128]. Of specific interest are data grids which are concerned with 'the controlled sharing and management of large amounts of distributed data'. Data grids may be combined with computational grid systems. A number of open source middleware applications have been developed to support grids as a means of data sharing[129]. The in depth investigation of using grids would be a project in its own right. The ADS is actively investigating the possibility of a project proposal to the e-Science programme[130]. This would use Big Data archives or data coming from the Virtual ExploratioN of Underwater Sites or VENUS project[131] in which the ADS is a partner. This will feed back into the wider archaeological community.

Currently the most consistent way of disseminating large datasets is likely to be on portable media; DVDs for the lower end of Big Data and external hard drives for anything bigger. As noted already one terabyte portable hard drives are available for under £300 and can be supplied and returned.

Acquiring large files is likely to be expensive in one way or another whether it is terms of taking up bandwidth or of costs for preparing media. Clearly potential users need to be able ascertain the relevance to them of available data. Traditionally this has been done through descriptive metadata. The use of 'tasters' such as thumbnail images or movie clips is also a well established decision support mechanism. Big data throws up some perhaps more unusual mechanisms such as fly-throughs and point cloud models. These are generally project outcomes and tend to use decimated datasets but they will inform on the relevance of the associated raw data. For example, the point cloud models produced by the Big Data case study **Breaking through Rock Art Recording**. These models are available through the ADS website[132] as Visualisation Toolkit (.vtk) files which can be viewed with 3D visualisation software including the freely available ParaView[133].

[128] http://en.wikipedia.org/wiki/Grid_computing

[129] http://www.omii.ac.uk/

[130] http://www.rcuk.ac.uk/escience/default.htm

[131] http://piccard.esil.univmed.fr/venus/

[132] http://ads.ahds.ac.uk/catalogue/resources.html?btrar_ahrb_2005

[133] http://www.paraview.org/HTML/Download.html

Comments and recommendations

The provision of a well formed Submission Information Package or SIP is essential for the successful reuse of data [p 36 – 37]

That the data in the SIP is in or has migration paths to suitable Big Data formats for dissemination is essential for the creation of the Dissemination Information Package or DIP [p 36]

The documentation including metadata in the DIP provides the basis of the framework for the successful ongoing management of the data [p 36 -37]

Grid technologies and specialised optical networks need further investigation [p 40 -41].

Currently the only consistent way of disseminating large datasets within a small community such as Archaeology is on portable media [p 41]

7 Summary

Archival Strategies at large

The long term preservation and dissemination of Big Data (indeed any data) should ideally be within an OAIS compliant framework (ISO 14721:2003 standard) [p 9 – 11]

Because the certification metrics are very new many archives are currently working towards OAIS compliance. As such trust must exist between creator and archive [p 11]

The Submission Information Package or SIP assumes major importance in the relationship between data producer and an OAIS compliant archive where as well as the data; documentation and metadata inform on preservation and reuse [p 10]

Data creation

In order to effectively undertake the long term preservation and dissemination of Big Data (indeed any data) archival organisations need a well formed Submission Information Package (SIP) [p 17 -22]

Consideration must be given to software and the formats it supports during data creation. Where long term reuse is a goal there must be clear migration paths for both preservation and reuse [p 13 – 17]

In general ASCII text is seen as the most stable format for data preservation whilst open binary formats suit the dissemination of Big Data because of a dramatic reduction in file size [p 15]

Inadequate documentation during data creation is the single biggest barrier to the future reuse of data. Documentation including metadata facilitates reuse as well as supporting in house administration and management during a project [p 17]

It is recommended that the UK GEMINI metadata standard which is compliant with the ISO (19115) Standard for Geographic Information is used to describe survey data. Further, maintenance of provenance and fixity metadata is identified as a crucial part of data creation [p 18 – 19]

Any other documentation that may facilitate reuse should also be included in the SIP. [p 22]

Acquisition, retention or disposal

Consideration should be given to producing a Guide to Good Practice for the archaeological community based on this report [p 24]

Only raw (or the rawest available) data and project outcomes should be considered for archiving [p 24]

Big Data can be accommodated in existing lifecycle cost models but concepts of retention and discard suggest revision may be necessary. An ongoing revision of the ADS charging policy is introduced [p 25 - 27]

Third party copyright is problematic for reuse; however, consideration should be given be alternative models offered by vendors such as Infoterra [p 27 -28

Preservation and ongoing management

The provision of a well formed Submission Information Package or SIP is essential for the successful long term preservation of data [p 30]

That the data in the SIP is in, or has migration paths to, suitable Big Data formats for preservation is essential for the creation of the Archival Information Package or AIP [p 30]

The documentation including metadata in the SIP provides the basis of the framework for the successful ongoing management of the data [p 30]

Access and use

The provision of a well formed Submission Information Package or SIP is essential for the successful reuse of data [p 36 – 37]

That the data in the SIP is in, or has migration paths to, suitable Big Data formats for dissemination is essential for the creation of the Dissemination Information Package or DIP [p 36]

The documentation including metadata in the DIP provides the basis of the framework for the successful ongoing management of the data [p 36 – 37]

Grid technologies and specialised optical networks need further investigation [p 40 – 41]

Currently the only consistent way of disseminating large datasets within a small community such as Archaeology is on portable media [p 41]

Acknowledgements

Project staff of the various case studies used by Big Data
Big Data Workshop participants
Big Data Questionnaire respondents
Tim Evans (ALSF Curatorial Officer based with the ADS)
Emma-Jane O'Riordan (IFA Workplace Trainee ADS\Internet Archaeology)
Julie Satchell and Julian Jansen van Rensburg (Hampshire and Wight Trust for Maritime Archaeology)
Justin Dix (University of Southampton)
Stephen Chafer (Infoterra Global) who was very helpful
EH Commissions staff for their patience

Appendix A: Cost Model for a Big Data case study

Various costings are given for the archive. The first uses the existing ADS charging policy. The others use a revision of this policy currently under development (see 4.3). The former reflects a traditional approach of preservation for as long as is possible or practical. The others reflect the emphasis placed on retention and discard policies within a lifecycle management approach. A requirement for the Big Data case studies was to archive their resources for five years and hence a retention and possible discard scenario.

Breaking through Rock Art Recording

Ongoing refreshment

Size of archive: 37.001 GB = 37,001 MB

Current policy

Ongoing management and refreshment under the existing cost model (at 50p per megabyte)

37,001 * 0.50p = £18,500.50 (one off payment)

Proposed policy

Management and refreshment costs under the proposed costing model detailed in section 4.3

37,001 * 0.XXp = (one off payments detailed below)

Retention period	Cost per MB (pence)	Refreshment costs
5 years	0.13	£4,810.13
10 years	0.22	£8,140.22
15 years	0.27	£9,990.27
20 years	0.28	£10,360.28
0ngoing	0.30	£11,100.30

File ingest (based on data collected during the ingest of this resource)

Action	Purpose	Duration (days)	Staff member
Secure archive	ingest	3.0	ADS Technical
Review (problems with archive)	Negotiate with depositors	1.0	ADS Manager
Interface design (downloads)	Design Interface	1.0	ADS Technical
Review	Negotiate with partners	0.5	ADS Manager
Internal review	Post review implementation	1.0	ADS Technical
Dissemination Review	Disseminate archive	0.5	ADS Manager

Staff member	Total days	Day rate	Cost
ADS Manager	2	£300	£600
ADS Technical	5	£250	£1250
Sub total			£1850
Overheads @ 25% of total staff costs			£462.50
Total staff costs			**£2,312.50**

Total cost

Ingest costs are the same for each example. The following totals are one off charges (ingest + ongoing refreshment) for a given retention period

Current policy

Retention period	Refreshment costs	Ingest costs	Total cost
Ongoing	£18,500.50	£2,312.50	**£20,813.00**

Proposed policy

Retention period	Refreshment costs	Ingest costs	Total cost
5 years	£4,810.13	£2,312.50	**£7,122.63**
10 years	£8,140.22	£2,312.50	**£10,452.72**
15 years	£9,990.27	£2,312.50	**£12,302.77**
20 years	£10,360.28	£2,312.50	**£12,672.78**
0ngoing	£11,100.30	£2,312.50	**£13,412.80**

+ VAT