*Tony Austin – Jenny Mitcham – Julian D. Richards*

# From Questions to Answers: Outcomes from the "Big Data" Project

*Abstract:* It is axiomatic that archaeologists will expand their research to encompass the computing resources available to them and thus the use of larger and larger datasets. These often take the form of seismic, maritime and satellite surveys using proprietary equipment and software. That there might be problems preserving and reusing this data led to the commissioning of the "Big Data" project. This paper summarises the outcomes as presented in the project's final report.

## *Introduction*

Archaeologists push the boundaries of the computing resources available to them. For example, they often work with increasingly large satellite, seismic and marine survey data. The possibility of problems associated with such datasets surfaced at a Heritage 3D workshop in 2004 and subsequently English Heritage commissioned the Archaeology Data Service (ADS) to investigate *Preservation and Management Strategies for Exceptionally Large Data Formats* or the 'Big Data' project as it is commonly known. The project used a range of approaches consisting of a literature search, an online questionnaire[1], a workshop[2], a formats review[3] and case studies[4].

## *Archival Strategies: the Bigger Picture*

Digital data usually has a reuse value. The long term preservation of data and its dissemination for reuse is generally seen as a specialised task to be undertaken by archives or data centres. Archaeological examples include the ADS[5] which incorporates AHDS Archaeology, part of the Arts and Humanities Data Service. Other not specifically archaeological organisations may hold data of interest to archaeologists. A good example of this is the Natural Environment Research Council (NERC) data centres[6]. These UK-based examples represent a global picture. An examination of the policy and guidance documentation of many of these organisations witnesses a revolution in archival theory and practice during the last few years.

The latter part of the last decade saw the development of lifecycle frameworks for the management and preservation of digital resources. Two influential publications in 1998 described lifecycle approaches to data. Firstly *A Strategic Policy Framework for Creating and Preserving Digital Collections* (AHDS) by Neil Beagrie and Dan Greenstein[7] defined a framework in detail. Lifecycle stages are generally agreed to be

- Creation
- Acquisition
- Preservation
- Reuse

which form the basis for the rest of this paper. Subsequently Tony Hendley in a *British Library Research and Innovation Report* (106) developed a cost model for this framework[8]. It should be noted that there are other archival strategies such as technology preservation and software emulation. Both have their problems; the former a reliance on computer museums and the latter cost and copyright issues, with the consequence that lifecycle approaches have come to dominate.

[1] http://ads.ahds.ac.uk/project/bigdata/survey_results/bigdata_quest_final.doc
[2] http://ads.ahds.ac.uk/project/bigdata/workshop.html
[3] http://ads.ahds.ac.uk/project/bigdata/formats.html
[4] http://ads.ahds.ac.uk/project/bigdata/caseStudies.html
[5] http://ads.ahds.ac.uk/
[6] http://www.nerc.ac.uk/research/sites/data/
[7] http://www.ukoln.ac.uk/services/papers/bl/framework/framework.html
[8] http://www.ukoln.ac.uk/services/elib/papers/tavistock/hendley/hendley.html

Currently, the development of the Open Archival Information System (OAIS) reference model by the Consultative Committee for Space Data Systems (CCSDS) of the US National Aeronautics and Space Administration (NASA) is taking the archival community by storm. The OAIS reference model has recently become an ISO (14721:2003) standard[9]. The CCSDS also have a technical recommendation available for consultation[10]. The purpose of the model is to provide a conceptual framework for considering functional requirements for systems concerned with the long term management and preservation of digital resources. The core entities[11] within the model include various information packages. A data creator (a producer in OAIS terminology) produces a Submission Information Package or SIP. The requirement for a SIP in effect formalises the guidance offered on data creation in series such as the AHDS Guides to Good Practice[12]. It contains data and documentation including metadata that will facilitate preservation and reuse. The SIP is passed on to an archival organisation. The SIP feeds into the generation of an Archival Information Package (AIP) for preservation and Dissemination Information Package (DIP) for reuse by a consumer. All three information packages are described more fully below.

The latest development is the publication of a certification document *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist* [13] by the Online Computer Library Center (OCLC), the Centre for Research Libraries (CRL) and the National Archives and Records Administration (NARA). This closely reflects the OAIS reference model in what it expects of an archive. The archival community have been rushing to claim, or at least state they are seeking, compliance with the OAIS reference model through this certification process. It should, however, be remembered that these metrics are very new and for the time being, trust must exist between data creator and archive.

## *The Process of Data Creation*

A design phase normally precedes the creation or acquisition of data. If reuse is a goal migration paths for the preservation and dissemination of any data must exist. Consideration must be given as to what documentation including metadata is necessary to facilitate reuse. Thus consideration of the Submission Information Package or SIP assumes importance even before the lifecycle of a resource begins.

The Big Data workshop teased out a definition for "big data" as "that which creates storage and transportability problems within a system with a system defined as creators, users, data centres and the computing technology they have access to". From this definition, disciplines better resourced than archaeology will have a different and bigger notion of what is considered to be "big data". Furthermore, as the ratio of power and storage to the cost of computing technology increases with seemingly relentless vigour, archaeologists expand the datasets they are working with. This is reflected at the ADS where archives measured in kilobytes were once the norm with today those measured in gigabytes (1 GB = 1,000,000 kB) becoming increasingly common.

The Big Data project was specifically concerned with the file formats associated with what are considered to be potential "big data" technologies. The latter were initially abstracted from the technologies being used by the Big Data *case studies*. The Big Data *questionnaire* (Q1) largely confirmed that these were the main technologies in use, including 3D laser scanning, LiDAR survey, digital video, Computer Aided Design (CAD), geophysical survey, Geographic Information Systems (GIS), bathymetric (single beam and multibeam sonar) survey, sidescan sonar and sub bottom profiling.

The *questionnaire* (Q3) went on to ask what software respondents used. Of the 101 packages noted, an astonishing 52 were unique responses which suggests a wide range of proprietary software. This was confirmed by the questionnaire with over 80%

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

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

[11] ibid *Fig. 4.1*

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

[13] http://www.crl.edu/areastudies/ASC/index.htm

proprietary as opposed to open source. Fortunately, most software packages support multiple formats with, for example, nearly half of the above able to export as structured ASCII text a preferred preservation format amongst the archival community.

Archivists have a hierarchy of preferred format types for data. It reflects cost of archiving and risk (there is more information on specific formats in the Big Data *formats review*). In order of preference these are:

- ASCII (American Standard Code for Information Interchange): A character encoding based on the English alphabet. It is an international and open standard suitable both unstructured and structured data when used with a delimiter. More structurally complex ASCII data is acceptable as long as the structure is defined as an open standard.
- Widely used open binary standards: An example of this is the openly published TIFF standard which is currently seen as the best preservation option for raster images. This has similarities with the next option in that it involves version migration but is preferred because it is an open standard.
- Version migration within a proprietary format: Until recently this has been the only real option for the preservation of CAD drawings in using Autodesk's DXF (Drawing eXchange Format) despite the largely erratic updates to the publication of changes between versions. Clearly, this is only an option with very widely supported and used formats. The cost of version migration for minority software packages would be prohibitive for any archive.

In a European context, the preference for ASCII text may appear Anglocentric as most European languages contain characters outside the basic ASCII range. However, the Big Data technologies considered here invariably produce or are viewed as vector graphics. Such data represents numeric spatial coordinates which can be represented within the ASCII character range, although other documents associated with an archive are likely to contain characters specific to its language. Other attributes can be associated with a coordinate such as numeric colour (red, green, blue) values. ASCII text can be accommodated in unicode transformation formats such as UTF-8[14] if associated attributes contain non-ASCII characters. ASCII character assignments (values) remain consistent within UTF-8 and are still represented by a single byte; an important consideration in terms of big data. Any non-ASCII characters can require up to four bytes for representation so file sizes would increase.

The project undertook an in depth *formats review* in order to identify generic formats suitable for the long term preservation and dissemination of data generated by these technologies. These are discussed below (sections *Preservation* and *Reuse*) but clearly the software used for data creation needs to support, or have migration paths to, such formats.

Poor documentation is probably the biggest single obstacle to long term preservation and reuse of data. Any documentation including metadata that facilitates these objectives should be included in the SIP. The importance of metadata is well documented in, for example, the AHDS Guides to Good Practice[15]. Beyond the generic elements common to most metadata sets, the ISO (19115) *Standard for Geographic Information – Metadata* is increasingly seen as the best standard for spatially referenced data. A number of initiatives such as the *UK GEMINI geospatial metadata standard*[16] developed jointly by the Association for Geographic Information (AGI) and the Cabinet Office e-Government Unit and the Spatial Information in Europe (INSPIRE) project's *Draft Implementing Rules for Metadata* stress compliance with it[17]. See the Big Data final report for more information on documentation.

## *Acquisition*

The process of ingest is well documented by archival organisations. For instance, they will have collections and charging policies, guidelines and FAQs. A well formed SIP will aid the physical process of ingest but there are possible problems pertaining to Big Data.

[14] http://en.wikipedia.org/wiki/UTF-8 for an overview

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

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

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

Survey data is often obtained from third parties who retain copyright. For example, the Where Rivers Meet *case study* acquired LiDAR data from infoterra, a provider of satellite and aerial data. Clearly this cannot be archived with an external archive without the permission of the copyright holder. This is unlikely to be forthcoming if the data has commercial value. An optimistic view would be to see this as a distributed archive with an organisation such as the ADS holding derived data and the third party supplier archiving the raw data. The problem is, however, that reuse would involve the need for purchase.

Processing of the Big Data case studies suggested that once procedures had been defined and migration paths existed, the archiving of big data was no more consuming in terms of human resource than other files in more familiar formats. It does take much longer to move files around, for example, from delivery media but this can be accomplished as a background task. Similarly, the generation of checksum or fixity values that check whether a transfer has been successful takes much longer. The physical storage requirements of Big Data archives will by definition be more than a conventional archive. As an example, the ADS charges for the latter by megabyte with the consequence that Big Data archives fit comfortably within the current ADS charging policy[18].

Transfer of Big Data archives has been suggested by some as problematic. It is certainly more involved than burning a CD or DVD but the cost of external hard drives has been dropping dramatically with, for example, one terabyte drives available in the UK for as little as £300 (approximately €440). Delivery media can be supplied or returned.

## *Preservation*

Tab. 1 summarises a sample of Big Data formats that are thought to have applicability for long term preservation. The table was abstracted from the Big Data *Formats Review* which contains further information about these and a range of other formats including the identification of migration tools in a number of cases.

If data in the Submission Information Package or SIP is in, or has established migration paths to, suitable preservation formats and the accompanying document is sufficient, an archive should have little problem generating a robust Archival Information Package (AIP). It should be noted that the above table is not seen as exclusive in that there may be other formats suited to a preservation role.

## *Reuse*

Over 70% of respondents to the Big Data Questionnaire noted that they reuse data at least once a year (Q6). Nearly 80% noted that they would allow access to their data to others. Clearly there is strong support for reuse of data. This could be because of the need, for example, to monitor change over time. It could be to incorporate earlier work into a new project or any number of reasons. Many of the preservation formats already noted can be used for data exchange. The following *(Tab. 2)* also have potential (again this is not exclusive).

Data and documentation in a SIP supplied by a data creator is used to create a Dissemination Information Package (DIP). For example, data should be in or have migration paths to formats suitable for dissemination, and documentation should provide resource description and discovery metadata.

Dissemination may be problematic for Big Data archives in that online delivery is not really an option for most users because of bandwidth issues. It is probable that delivery would have to be on DVDs or external hard drives. The possibility of using BitTorrent[19], a peer to peer communications protocol for file sharing, was investigated and has possibilities as a means of distribution. Basically copies of a file are stored by a number of clients or peers. A 'tracker computer' coordinates the file distribution and tells a new client download request where to get various pieces of the file from amongst the peers. This spreads the load at the supply end but there would still be a bottleneck at the receiver end.

[18] http://ads.ahds.ac.uk/project/userinfo/charging.html
[19] http://en.wikipedia.org/wiki/BitTorrent

| Format/Technologies | Properties* | Comment |
|---|---|---|
| ASCII text (.txt, .dat, etc.) | Published standard<br>ASCII<br>Raw | Preserve as ASCII text with support documentation |
| DXF: Drawing eXchange Format (.dxf)<br>3D including Point cloud, CAD, Mesh | Proprietary published (currently)<br>ASCII and binary<br>Processed usually | ASCII DXF and version migration still seem to be the best preservation option but other options emerging. |
| GML: Geography Markup Language (.gml)<br>Geospatial data including GIS | Open Published standard<br>ASCII<br>Processed | GML is very suited for preservation and data exchange of geospatial data. |
| MGD77 (.mgd77)<br><br>Geophysical data including Bathymetric, Magnetic, Gravity | Published standard<br>ASCII<br>Raw or can be | In being ASCII based and published could act as a preservation format. Has support as a data exchange format. |
| MPEG 1 (.mpg, .mpeg)<br><br>Video, Audio | Published open standard<br>Binary<br>Processed usually<br>Video quality | Suitable for preservation and data exchange |
| MPEG 2 (.mpg, .mpeg)<br><br>Video, Audio | Published open standard<br>Binary<br>Processed usually<br>DVD quality | Suitable for preservation and data exchange |
| MPEG 4 (.mp4)<br><br>Video, Audio | Published open standard<br>Binary<br>Processed<br>Streaming | Can be used for preservation. In being an online streaming standard could be used for data sharing |
| NTF: National Transfer Format (.ntf)<br>Geospatial including Point cloud, CAD, DEM, Lidar | Published open standard<br>ASCII<br>Raw and processed | In being ASCII based and published it should be suited for both transfer and preservation. Unclear; however, as to how wide its usage is outside of the UK Ordnance Survey where it is being superseded by GML |
| NetCDF: Network Common Data Form (.nc)<br>Scientific including Bathymetric, Lidar and others? | Published open standard<br>Binary / ASCII dumps<br>Raw or can be | Conceivably this could provide a mechanism for preservation and data sharing through storing once and generating binary as requested using an associated toolkit. |
| OBJ (.obj)<br><br>3D including Laser scanning, Mesh, Point cloud | Published<br>ASCII<br>Raw data or can be | Wide support suggests a possible data exchange format. In being ASCII based it could act as a preservation format |
| XML: eXtensible Markup Language (.xml or can be – see GML)<br>Increasing range of technologies | Published open standard<br>ASCII<br>RAW or processed | Ideal for exchange and preservation if an established schema exists |
| XYZ (.xyz .xyzrgb)<br><br>Laser scanning, Lidar | XYZ coordinates sometimes with colour values<br>ASCII (can be binary)<br>Raw(ish) | ASCII text is seen as the best option for long term preservation along with suitable metadata |

Tab. 1. Contains to Chapter 'Preservation'.
* Links to standards are provided in the Big Data *formats review*

| Format/Technologies | Properties | Comment |
|---|---|---|
| GSF: Generic Sensor Format (.gsf)<br>Bathymetric | Openly published standard<br>Binary<br>Raw or can be | Possible data exchange format. It was noted as such during the Big Data Workshop |
| LAS (.las)<br>Lidar<br>Laser scanning (not formalised as yet) | Openly published standard<br>Binary<br>Raw or can be | Specifically designed for the exchange of data; a role for which it has strong support. |
| SDTS: Spatial Data Transfer Standard (various including .ddf)<br>Range of spatial data | Openly published standard<br>Binary<br>Raw data or can be | Well supported as a data exchange standard but may be US centric. |
| SEG Y (.segy)<br><br>Seismic survey including Sub-bottom profiling, GPR | Openly published<br>Binary<br>Raw or can be | Possibly useful as a data exchange format as it appears widely supported by proprietary systems. |
| SEG 2 (.sg2, .dat)<br><br>Seismic survey including GPR | Openly published<br>Binary<br>Raw data | A possible exchange format. |
| eXtended Triton Format (.xtf)<br><br>Sidescan sonar, Sub-bottom profiling, Bathymetric | Proprietary but currently a publicly available specification<br>Binary<br>Raw or can be | A possible exchange format but with no guarantee that the specification will remain in the public domain. |

Tab. 2. Contains to Chapter 'Reuse'

## *Conclusion*

The curation of Big Data appears to be largely unproblematic providing the data is in or has migration paths to formats suitable for long term preservation and dissemination for reuse, and sufficient documentation including metadata is provided to facilitate these processes. In short the provision of a well formed Submission Information Package or SIP by a data creator will ensure the smooth transfer to an archival environment.

The size of very large datasets is problematic in terms of storage costs and dissemination but neither is insurmountable if the data in question is considered to have sufficient value to a user community.

*Tony Austin*

*Archaeology Data Service, Department of Archaeology*
*The King's Manor, University of York*
*Exhibition Square*
*York YO1 7EP, United Kingdom*

*Jenny Mitcham*

*Archaeology Data Service, Department of Archaeology*
*The King's Manor, University of York*
*Exhibition Square*
*York YO1 7EP, United Kingdom*

*Julian D. Richards*

*Archaeology Data Service, Department of Archaeology*
*The King's Manor, University of York*
*Exhibition Square*
*York YO1 7EP, United Kingdom*
*jdr1@york.ac.uk*