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

Review of nature of technologies and formats

Tony Austin & Jen Mitcham
10 May 2007

This report has been produced as part of the Big Data Project. It is a technical review of each of the 'Big Data' technologies currently practised by archaeologists with a consideration of data formats for preservation and future dissemination. As well as data acquisition there will be an analysis phase to any project. Survey normally involves a series of traverses over a spatially defined area. Composite mosaics can be produced as either part of acquisition or as part of post processing. The composite can then be fed into a range of geospatial tools including 3-D visualization. Examples include Geographical Information Systems (GIS) and Computer Aided Design (CAD) software.

Discussion as indicated by the Big Data questionnaire[1] and the project case studies[2] focuses on the following technologies

- Sonar (single beam, bathymetry and sub bottom profiling)
- Acoustic Tracking
- 3D Laser Scanning
- Geophysics
- Geographic (eg GIS)
- LiDAR
- Digital Video

Raster (still) images and Computer Aided Design (CAD) also featured in the questionnaire but are covered more than adequately elsewhere. See, for example, the recent AHDS **Digital Image Archiving Study**[3] and the **CAD: A Guide to Good Practice**[4]

---

[1] http://ads.ahds.ac.uk/project/bigdata/survey_results/bigdata_quest_final.doc

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

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

[4] http://ads.ahds.ac.uk/project/goodguides/cad/

A tabular summary of Big Data formats can be found at the end of this document (table 1). This structure has been adopted rather than considering formats under each technology as the formats often span technologies. For example, SEG Y available in a number of maritime applications (see table 1) is a generic seismic survey format. This summary does not pretend to be inclusive but rather a representative flavour of the vast range of formats that seem to be associated with Big Data.

Sonar

Sonar (**SOund NA**vigation and **R**anging) is a simple technique used by maritime archaeologists to detect wrecks. It uses sound waves to detect and locate submerged objects or measure the distance to the floor of a body of water and can be combined with a Global Positioning System (GPS) and other sensors to accurately locate features of interest. A useful overview or Maritime Survey techniques can be found on the Woods Hole Science Center (part of the Unites States Geological Survey) website[5].

Bathymetry (single beam and multibeam sonar)

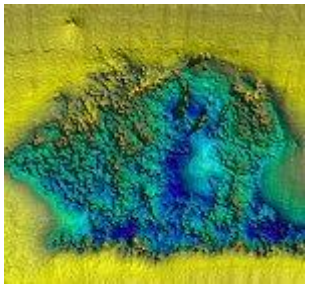

Illustration: top view of the multibeam data of Hazardous, lost in November 1706, when she was run aground in Bracklesham Bay © Wessex Archaeology

Single beam scanning sends a single pulse from a transducer directly downwards and measures the time taken for the reflected energy from the seabed to return. This time is multiplied by the speed of sound in the prevalent water conditions and divided by two to give the depth of a single point.

Multibeam sonar sends sound waves across the seabed beneath and to either side of the survey vessel, producing spot heights for many thousands of points on the seabed as the vessel moves forward. This allows for the production of accurate 3D terrain models of the sea floor from which objects on the seabed can be recorded and quantified. Wessex Archaeology used multibeam bathymetry during the Wrecks on the Seabed project (Big Data

[5] http://woodshole.er.usgs.gov/operations/sfmapping/index.htm

case study)[6]. As well as the raw data itself, 3D terrain models, 3D fly through movies and 2D georeferenced images were created. The 2D images were then used as a base for site plans and divers were able to use offset and triangulation to record other objects on to the plans.

The data

Why should we archive?

For future interpretation of data. Seeing anomalies in the results not seen before?
For monitoring condition and erosion of wreck sites
For targeting areas for future dives/fieldwork

Problems and issues

Many bathymetric systems use proprietary software. The extent to which this software supports open standards or openly published specifications is largely unknown. Data exchange between systems may also be problematic.

Specialised metadata

Metadata to be recorded alongside the data itself includes:
Equipment used (make and model)
Equipment settings
Assessment of accuracy?
Methodology
Software used
Processing carried out

Associated formats include

Generic Sensor Format (.gsf), HYPACK (.hsx, .hs2), MGD77 (.mgd77), eXtended Triton Format (.xtf), Fledermaus (.sd, .scene – visualisation)

[6] http://www.wessexarch.co.uk/projects/marine/alsf/wrecks_seabed/multibeam_sonar.html

Sidescan sonar

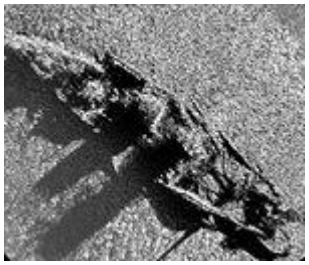

Illustration: This image created with sidescan data clearly shows a ship wreck protruding from the seabed © Wessex Archaeology

Sidescan sonar is a device used by maritime archaeologists to locate submerged structures and artefacts. The equipment consists of a 'fish' that is towed along behind the boat emitting a high frequency pulse of sound. Echoes bounce back from any feature protruding from the sea bed thus recording the location of remains. The sidescan sonar is so named because pulses are sent in a wide angle, not only straight down, but also to the sides. Each pulse records a strip of the seabed and as the boat slowly advances, a bigger picture can be obtained. As well as being a useful means of detecting undiscovered wreck sites, sidescan data can also be used to detect the extents and character of known wrecks.

The data

The data tends to be in a wide range of little known proprietary and binary formats. Although there are some open standards such as SEG Y around. The software packages associated with sidescan sonar may support ASCII or openly published binary exports.

Why should we archive?

For future-interpretation of data. Seeing anomalies in the results not seen before?
For monitoring condition and erosion of wreck sites
For targeting areas for future dives/fieldwork

Problems and issues

Many sidescan systems use proprietary software. The extent to which this software supports open standards or openly published specifications is largely unknown. Data exchange between systems may also be problematic.

Specialised metadata

Metadata to be recorded alongside the data itself includes:
Equipment used (make and model)
Equipment settings

Assessment of accuracy?
Methodology
Software used
Processing carried out

Associated formats include

eXtended Triton Format (.xtf), SEG-Y, CODA (.cod, .cda), Q-MIPS (.dat), HYPACK (.hsx, .hs2), MSTIFF (.mst)

Sub bottom profiling

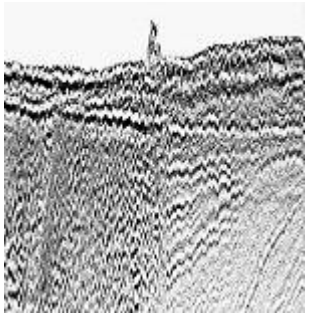

Illustration: example of sub-bottom profiler data © Wessex Archaeology

Powerful low frequency echo-sounders have been developed for providing profiles of the upper layers of the ocean bottom. Specifically sub-bottom profiling is used by marine archaeologists to detect wrecks and deposits below the surface of the sea floor. The buried extents of known wreck sites can be traced using an acoustic pulse to penetrate the sediment below the sea bed. Echoes from surfaces or the horizons between different geological layers are returned and recorded by the profiler and the sequence of deposition and subsequent erosion can be recorded. The case study, Wessex Archaeology, utilised sub bottom profiling[7] for the Wrecks on the Seabed project

The data

The data tends to be in a wide range of little known proprietary and binary formats. Although there are some open standards such as SEG Y around. The software packages associated with sub bottom profiling may support ASCII or openly published binary exports.

Why should we archive?

For future-interpretation of data. Seeing anomalies in the results not seen before?
For monitoring condition and erosion of wreck sites
For targeting areas for future dives/fieldwork

Problems and issues

Many systems use proprietary software. The extent to which this software supports open standards or openly published specifications is largely unknown. Data exchange between systems may also be problematic.

[7] http://www.wessexarch.co.uk/projects/marine/alsf/wrecks_seabed/sub_bottom_profiler.html

Specialised metadata

Metadata to be recorded alongside the data itself includes:
Equipment used (make and model)
Equipment settings
Assessment of accuracy?
Methodology
Software used
Processing carried out

Associated formats include

CODA (.cod, .cda), QMIPS (.dat), SEG Y (.segy), eXtended Triton Format (.xtf)

Acoustic Tracking

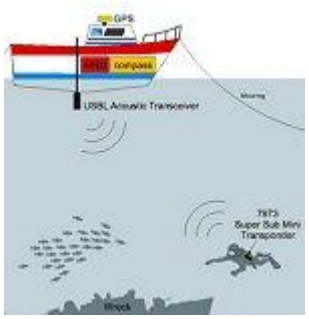

Illustration: diagram showing how acoustic tracking devices keep track of the divers' location at any one time © Wessex Archaeology

Acoustic tracking can be used to keep a log of a diver's location throughout the dive. Sound signals are emitted by a beacon attached to the diver and picked up by a transceiver attached to the side of the boat. The relative position of the diver underwater can be calculated and these relative co-ordinates can be used to calculate an absolute location for the diver. Additional equipment may be needed to compensate for the motion of the vessel in the water. Acoustic Tracking was utilised for the Wrecks on the Seabed project[8].

The data

Normal practice is to use a data logger for collection. Generally the data will be in the form of structured ASCII text. As such it will be easy to import into other packages such as a GIS or database. Wessex Archaeology supplied their Acoustic Tracking data as a Microsoft Access database

Why should we archive?

For Wessex Archaeology this data was seen as crucial to the project archive as it sets much of the other maritime archaeology project data in context. Will need to refer to this database to establish where the diver was when individual photographs were taken, segments of digital video recorded or general observations made.

Problems and issues

Possibly processed and not the raw data.

Specialised metadata

Metadata to be recorded alongside the data itself includes:
Equipment used (make and model)

[8] http://www.wessexarch.co.uk/projects/marine/alsf/wrecks_seabed/acoustic_diver_tracking.html

Equipment settings
Assessment of accuracy
Methodology
Software used
Processing carried out

Associated formats include

ASCII text formats

3D Laser Scanning

Illustration: Solid model created from point cloud laser scan data from stone 7 of Castlerigg Stone Circle in Cumbria - image from Breaking Through Rock Art Recording project

There are a wide variety of applications of laser scanning as a tool for capturing 3D survey data within archaeology. A common application of this technology is as a tool for recording and analysing rock art, but subjects can range from a small artefact to a whole site or landscape. A 3D image of Rievaulx Abbey was recently created by Archaeoptics in 10 minutes. The benefit of this technique is that a visually appealing and reasonably accurate copy of a real world site or object can quickly be created and manipulated on screen.

When a laser scanner is directed at the subject to be scanned, a laser light is emitted and reflected back from the surface of the subject. The scanner can then calculate the distance to this surface by measuring the time it takes, and x, y and z points relative to the scanner can be recorded. Absolute co-ordinates can then be created by georeferencing the position of the scanner. Some scanners may also record colour values for each point scanned and the reflection intensity of the surface (see Trinks *et al*, 2005[9])

Huge datasets are produced using this technology. The recent project by Wessex Archaeology and Archaeoptics to scan Stonehenge reported that each scan took "3 seconds to complete and acquiring 300,000 discrete 3D points per scan. A total of 9 million measurements were collected in just 30 minutes" (see Goskar *et al*, 2003[10]). It is not surprising that laser scanner data files can be many gigabytes in size.

The data

[9] Trinks I, Díaz-Andreu M, Hobbs R & K. Sharpe, K. 2005. 'Digital rock art recording: visualising petroglyphs using 3D laser scanner data', Rock Art Research 22, p. 131-9
Also online at http://www.dur.ac.uk/m.diaz-andreu/articles/2005_RAR_Trinks_et_al.pdf

[10] Goskar T, with Carty A, Cripps P, Brayne C, & Vickers D. 2003. 'The Stonehenge Lasershow', British Archaeology 73
Also online at http://www.britarch.ac.uk/ba/ba73/feat1.shtml

There are a number of different types of data that are created as a laser scanning project progresses:

Primary data produced through this technique is point cloud data. Point clouds essentially consist of raw XYZ data, to locate each point in space, plus if recorded, RGB data to record the colour of each point.

Firstly there are the raw observations as collected by the scanning equipment in a number of different proprietary formats.

Numerous scans may be carried out to record a complex subject, with the scanning equipment moved to a different position each time. This will create a large number of data files. All of these individual scans would then need to be stitched together in order to create a composite mosaic of the whole subject. From the point cloud data, it is possible to create a solid model of the subject, such as that illustrated above. A cut down or decimated version of the raw XYZ data may be used to create a dataset of a more manageable size for processing, viewing and analysing.

Why should we archive?

In archaeology it is thought that perhaps one of the main opportunities we will gain from storing and re-using this data in the future is that successive scans of the same sites may be used to monitor erosion or other physical changes to the site. The Durham University Fading Rock Art Landscapes project[11] was set up with just this in mind.

Data could also be re-processed in different ways to create new models and allow for new interpretations of the data. With technologies such as this it is very easy to create large datasets from a high resolution scan and then be hampered by a lack of storage space and processing power when attempting to view and interpret the resulting dataset.

Problems and issues

There are a fairly small range of software tools for viewing laser scan data. Huge file sizes may hamper reuse - may be more appropriate for researchers to interrogate a cut down or decimated version of the laser scanning data as this will be easier to process. No standard data format currently exists for laser scanning data. This should be addressed.

Which data do we actually need to archive? The raw data as created by the laser scanner? ie: a separate file for each scan - not yet combined to produce full composite scan of whole object. Or perhaps the composite scan is fine - will this have undergone additional processing? Processed results are also useful. If theories were reached as a result of looking at a particular version of

[11] http://www.dur.ac.uk/prehistoric.art/

the dataset, is it worth keeping this also so future researchers can see how a particular theory came about?

Specialised metadata

Re-use potential is maximised if relevant metadata exists for laser scanning data. Both technical information about the survey and more obvious information about the context of the scan are required. Lists are published by Heritage3D[12] include:

Date of capture
Scanning system used
Company name
Monument name
Weather during scanning
Point density on the object
Technical information relating to the scanning equipment itself - may include triangulation, timed pulse, phase comparison

Associated formats include

XYZ (.xyz), Visualisation ToolKit (.vtk - processed), LAS (.las), Riscan Pro (.3dd), National Transfer Format (.ntf), OBJ (.obj), Spatial Data Transfer Standard (various), Drawing eXchange Format (.dxf - processed)

[12] http://www.ceg.ncl.ac.uk/heritage3d/downloads/TLS%20formats%20V1.pdf

Geophysics

As stated in the ADS **Geophysical Data in Archaeology Guide to Good Practice**[13], the increasing size and sampling resolution of geophysical surveys in archaeology is resulting in the accumulation of increasing quantities of data. However the most common techniques, resistivity and magnetometer surveying generally do not produce datasets that are large enough to fall under the remit of the Big Data Project. The one land-based geophysical technique that can produce exceptionally large datasets is Ground Penetrating Radar.

In a Ground Penetrating Radar survey, the instrument is dragged along the ground at a constant speed and electromagnetic pulses are sent into the ground by an antenna. As the pulses come into contact with objects and layers within the ground, they are reflected back to the instrument picked up by a receiving antenna. A single GPR transect creates a vertical 2D image of the subsurface features. If numerous transacts are carried out within a grid, these images can be combined to create a 3D depiction of the results.
The amount of data being collected using Ground Penetrating Radar is likely to increase as the technology moves along. For example, Terravision[14] by Geophysical Survey Systems, Inc (GSSI) is a new and very advanced piece of equipment for carrying out GPR survey. It features a 14 antenna array and 6 foot wide survey path with a data collection speed of up to 10mph. The internal storage capacity of the equipment can be up to 32GB. The quantities of data that could be produced if archaeologists started to use equipment like this would be substantial.

The data

Basically these will be x, y and z co-ordinates. A good overview of GPR data formats is available on one of the USGS websites including a number of format specifications[15]. GPR data from the Where Rivers Meet case study was in the proprietary and binary DZT format. This stores data as radians

[13] http://ads.ahds.ac.uk/project/goodguides/geophys/

[14] http://www.geophysical.com/TerraVision.htm

[15] http://pubs.usgs.gov/of/2002/ofr-02-0166/ofr02_166.pdf

Why should we archive?

For future-interpretation of data. Seeing anomalies in the results not seen before?
For monitoring condition and erosion of the archaeology
For targeting areas for future fieldwork
Interpretation of GPR data can be fairly subjective, users wishing to re-use the GPR data may wish to go back to raw results rather than use someone elses subjective interpretations

Problems and issues

Metadata! If the raw GPR data is to be archived and re-used it has limited value without the field notebooks used to record location of each transect. This metadata will most probably be in a paper format and would require some time and effort to digitise and rationalise for general re-use.
Specialised metadata
Metadata to be recorded alongside the data itself includes:
Equipment used (make and model)
Equipment settings
Assessment of accuracy?
Methodology
Software used
Processing carried out
Check this against G2GP, pres handbook, proc doc
Data formats/file formats

Associated formats include

DEM (.dem), DLG (.dlg), RADAN™ DZT (.dzt), NTF (.ntf), RAMAC – RD3/RAD (.rd3, .rad), Fledermaus (.sd, .scene – processed), SDTS (various plus .ddf), SEG 2 (.dat, .sg2), SEG Y (.segy)

Geographic (eg GIS)

Illustration: High Resolution raster images such as this aerial photograph of Aberystwyth can be used within a Geographic Information System. Image taken from **Mapping Medieval Townscapes: a digital atlas of the new towns of Edward I** project[16]

Though many archaeologists use GIS without creating exceptionally large datasets, large file sizes can be an issue for some GIS projects. High resolution raster layers such as scanned aerial photographs, satellite imagery and sometimes digital elevation models (DEM) within a GIS can be very large. An high quality scanned colour aerial photo at a 'ground-resolution' of, 20 cm per pixel, could be as large as 250 MB. Hundreds of these photos could be needed to give coverage of a large study area such a county. This would have huge implications for the future archiving and re-use of the project data. See the AHDS GIS Guide to Good Practice[17]

The data

Data can consist of a georeferenced high resolution raster images such as a geotiff or vector x,y,z data. Data is generally collected using other technologies and processed within a GIS to generate project outcomes.

The established preservation strategy for GIS vector data is to migrate to ESRI Shape (SHP) and Export (E00) formats which are generally seen as *de facto* standards in light of alternatives[18]. However, the development of the Open Source GDAL Geospatial Data Abstraction Library[19] is allowing popular GIS formats to be abstracted to GML, an XML based language. This is being adopted as a preservation strategy by the ADS. The process can be reverse engineered if required.

Why should we archive?

For future interpretation of data. Seeing anomalies in the results not seen before?

---

[16] http://ads.ahds.ac.uk/catalogue/specColl/atlas_ahrb_2005/

[17] http://ads.ahds.ac.uk/project/goodguides/gis/

[18] http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf 5.4

[19] http://www.gdal.org/

For monitoring condition and erosion of landscapes
For targeting areas for future fieldwork

Problems and issues

Copyright?! Satellite imagery and aerial photographs will often have been obtained from other organisations and are not necessarily owned by the project.
Specialised metadata
How data acquired/created
Copyright
Processing carried out

Associated formats include

GEOTIFF (.tiff), TIFF World file (.tfw), JPEG World file (.jgw), DEM (.dem), ESRI Shape file (.shp), ESRI export (.e00), MOSS export (.exp), GML (.gml), GRASS (various), MapInfo (.tab), MapInfo interchange Format (.mif), IDRISI raster (.rst, rdc), IDRISI vector (.vct and others)

LiDAR

Illustration: LiDAR image of Hambledon Hill showing the hillfort with an internal long barrow © Environment Agency.

The name LiDAR comes from '**Li**ght **D**irection **A**nd **R**anging'. The technique[20] [21] involves scanning a pulsed laser along the ground from an aircraft. By monitoring the direction and speed of the incoming reflected pulse, the layout of the landscape can be recorded with some accuracy. Once xyz co-ordinates have been collected in this way, they can be used to generate a digital elevation model of the study area that can be viewed within a GIS package. The benefits to archaeologists are obvious[22]. Any undulations in the ground surface will be recorded, thus even very slight rises and depressions of archaeological earthworks will be visible within the elevation model. The researcher is able to alter the light direction and intensity within the model to look for features which would otherwise be hard to pick out.
English Heritage[23] reports that LiDAR can measure between 20,000 to 100,000 points per second. It is therefore not surprising that exceptionally large datasets are rapidly created.

Lidar data is normally supplied by specialist organisations; usually commercial, although it is worth noting that some infoterra[24], a leading supplier of 'geo-information solutions' data is available to academic users through MIMAS (Manchester Information & Associated Services) under a CHEST agreement[25]. Also some material is available from the Environment Agency[26] and the same for academic use through MIMAS (Manchester Information & Associated

[20] http://aarg.univie.ac.at/events/aarg2000/abstracts/abstracts2000.html#72

[21] http://www.english-heritage.org.uk/server/show/nav.8730

[22] http://aarg.univie.ac.at/events/aarg2000/abstracts/abstracts2000.html#73

[23] http://www.english-heritage.org.uk/server/show/nav.8730

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

[25] http://landmap.ac.uk/download/infoterra_download.htm

[26] http://www.environment-agency.gov.uk/science/monitoring/131047/?lang=_e

Services)[27] but this turns out to be mostly processed data either as raster images or in GIS formats.

The data

Raw data as recorded from the aeroplane is slightly different, containing time of flight of the laser and position of the plane in WGS84[28] (World Geodetic System) coordinates. This is normally converted into a local system such as OSGB36[29] xyz which is much more useful. It is not believed that there is any benefit in archiving the actual raw data.

XYZ ASCII files (point clouds) are referenced to a coordinate system such as the British National Grid. Data may also include an i value to record the intensity of the returned signal. Two data files will be produced; first pulse and last pulse. The first pulse data records the highest point or the first pulse returned to the aircraft. This could be the top of a tree for example. Last pulse data records the last pulse back to the sensor and this would represent the height of the ground underneath the tree. Last pulse data could therefore be used to create Digital Terrain Models (DTM = DEM or Digital Elevation Model) of the landscape.

Why should we archive?

For future-interpretation of data. Seeing anomalies in the results not seen before?
For monitoring condition and erosion of landscapes
For targeting areas for future fieldwork

Problems and issues

Copyright of data. Archaeologists are unlikely to be able to afford the equipment to carry out their own LiDAR survey so will either purchase existing data or get contractors to carry out the survey for them. This was the case with one of the Big Data case studies, Where Rivers Meet, where the data was supplied by infoterra. Clearly this may prevent reuse by or archiving with a third party

Specialised metadata

Metadata to be recorded alongside the data itself includes:
Most important metadata to accompany LiDAR data
Instrument make and model (thus wavelength of laser)
Altitude of flight
Effective point spacing (ie: 1 point per metre)

---

[27] http://landmap.mimas.ac.uk/lidar/lidar.html

[28] http://en.wikipedia.org/wiki/WGS84

[29] http://en.wikipedia.org/wiki/British_national_grid_reference_system

Copyright
Other metadata that it would be useful to have
Time of year of flight
Weather conditions/ground condition report (weather for previous week too as need to assess ground water levels)
Instrument settings
QA tolerance (for example ± 15cm)

Associated formats include

LAS (.las), XYZ (.xyz), DML (.dml), Blue Sky (.txt), NTF (.ntf), Doppler Markup Language (.dml)

Digital Video/Audio

Illustration: Screenshot taken from a digital video extract mounted by Wessex Archaeology on their 'Wrecks on the Seabed' project web pages © Wessex Archaeology

Digital video, notorious for producing large file sizes, consists of a series of digital images that when viewed in succession, create an impression of movement. It may or may not be associated with digital audio. Digital video is becoming more and more popular as a means of recording archaeology, particularly amongst maritime archaeologists where sites are less easily accessible than terrestrial sites. If a whole dive is recorded in this way by means of a hat-mounted camera, the generated file will be substantial in size. Digital video can also be used to record terrestrial archaeology, for example to record excavations in progress, condition surveys, experimental archaeology and interviews. English Heritage have been very involved in using video to benefit Archaeology including the production of video diaries of fieldwork[30]. TheBamburgh Research Project[31] has more recently made extensive use of digital video to record both the archaeological processes and the social context of their training excavations. Though Internet Archaeology noted in 1997 (EVA Conference paper) that few archaeologists have access to the technology to create digital video[32], it has moved along quickly. With digital cameras that shoot video becoming cheaper and more accessible to a wider range of users plus the easy availability of video editing software, it is likely that use of digital video within archaeology will increase in popularity. Similarly, digital audio can be used by archaeologists. Perhaps not as useful as video to record excavations, artefacts and sites, but oral history projects and interviews with archaeologists are increasingly using digital audio as a medium.

The recent AHDS **Digital Moving Images and Sound Archiving Study**[33] provides informative guidance.

The data

[30] http://www.english-heritage.org.uk/upload/pdf/Conservation_Bulletin_27.pdf

[31] http://www.bamburghresearchproject.co.uk/media.htm

[32] http://intarch.ac.uk/news/eva97.html

[33] http://ahds.ac.uk/about/projects/archiving-studies/moving-images-sound-archiving-final.pdf

Data has often originally recorded onto DV tape. This is the most economical way to store it but it should be noted that tape degrades relatively quickly. Also video tape is rapidly being superseded by disc based technologies both in terms of recording and viewing. Can be transferred to disk based storage where large scale storage devices are increasingly affordable.

Round 1 of Wessex Archaeology's Wrecks on the Seabed project (a case study – appendix D) has produced somewhere in the region of 75 gigabytes worth of dive footage. Even if all the data is worth retaining something of this size should be unproblematic.

Why should we archive?

In the case of the Wessex data associated with the track log (see Acoustic Tracking above) digital video in a maritime context provides an important record of what was seen by the diver at what particular points on the underwater site. Though not many future users would wish to view full un-edited footage of the dive, it is important to preserve this information. Digital video could be utilised as a tool to assess the condition of a wreck site and monitor damage over time. In short it pulls together components of a project just as the traditional paper site diary and more recent video diaries do. Such videos also become a source for historiography.

Problems and issues

Can be very long - Wrecks in the Seabed project has produced around 40 hours of digital video, much of which would be unclear and murky with sections where not very much is happening! Not sure if someone would want to sit down and view the whole lot. Most users would probably be happy with a cleaner edited version showing some of the highlights of the dive.

DV tape has an increasingly short lifespan and should be migrated to disc based storage. DVDs also have a finite lifespan with hard drives (internal or external) providing the securest medium for storage.

Specialised metadata

Digital video in an underwater context may be associated with some record of where the diver and thus the camera was at any one time.
Also, the metadata we would expect from any deposit of digital video:
Software, version and platform
Name and version of video codec (where appropriate)
video dimension (in pixels)
frame rate per second (fps)
bit rate
Name and version of audio codec including sample frequency, bit-rate and channel information
Length (hours, minutes, seconds) of file
File size

Associated formats include

MPEG 1 (.mpg, .mpeg), MPEG 3 (.mpg, .mpeg), MPEG 4 (.mpg4) , MXF (.mxf)

Table 1: Formats review

File extension, name and technologies	Description	Properties	Comments
.3dd RiSCAN Pro[34] Laser scanning Including Point cloud Mesh	Proprietary, binary format used by Riegl's Laser (3D) scanning software; RiSCAN Pro. This can export as a variety of other formats including ASCII, DXF, OBJ and VRML. RiSCAN apparently supports 'Smooth data transfer by using the well documented RiSCAN PRO XML-project format'[35] which could provide supporting metadata although we were unable to locate the schema definition.	Proprietary Binary Raw data or can be	Export to ASCII based format for preservation along with suitable metadata.
.cod .cda CODA[36] Seismic survey including Sidescan sonar Sub-bottom profiling	Difficult to locate information about CODA formats. A number of companies specializing in processing sonar data state that they don't accept data in these formats; however, products produced by CODA technologies generally support other formats such as XTF and SEG Y.	Proprietary Binary Raw data or can be	As a general guideline export to more open standards if possible and then to ASCII. Will need supporting metadata
.csv .dat .txt .xyz and others Delimited text	Delimited as the name suggests is structured (usually ASCII) text. Comma Separated Values (CSV) is perhaps the best known example and is associated particularly with spreadsheets. Other popular delimiters include tab and pipe. The .txt extension can reference structured as well as unstructured data. While .xyz is fairly common with coordinate data.	Open standard[37] ASCII Raw data or can be	The archival dream for the long term preservation of data with header information about data collection, etc held as metadata

[34] http://www.3dlasermapping.com/uk/3d/software/RiSCAN%20PRO.pdf

[35] http://www.riegl.com/terrestrial_scanners/3d_software_selection_guide_/software_package_/riegl_software_index.htm

[36] http://www.codaoctopus.com/survey/coda/index.asp

[37] http://en.wikipedia.org/wiki/ASCII for an introduction

.dat (see .csv)			
.dat QMIPS Sidescan sonar Sub-bottom profiling	Appears to be generated by data loggers such as the Triton Technology ISIS Data Logger system. Such data is often converted to other formats such as SEG Y as, for example, by the US Geological Survey (USGS) using an in-house script called qmipstosegy for onward processing[38]. Information about the QMIPS format can be viewed on the USGS website including the file header details[39].	Proprietary? Binary Raw data	Appears to be migrated as a matter of course to other formats (see entries for these).
.ddf DDF: Data Description File	See SDTS		
.dem DEM: Digital Elevation Model DEM: Digital Elevation Models Mesh Pointcloud	ASCII based format developed by the United States Geological Survey (USGS). Sources note that these are raster images. Described as largely superseded by the SDTS standard (see below) but older datasets still in DEM formats. Still supported by many geospatial processing packages. There is a freely available viewer for many USGS formats which is a limited-feature version of commercial software called Global Mapper[7]. Possibly US centric?	Published standard[40] ASCII Raw data or can be	Suited to data exchange and preservation but is essentially deprecated (but is it – see SDTS).
.dlg DLG: Digital Line Graph DEM: Digital Elevation Models	'The U.S. Geological Survey's (USGS) digital line graph (DLG) files are digital vector representations of cartographic information. Data files of topographic and planimetric map features are derived from either aerial photographs or from cartographic source materials using manual and automated digitizing methods'[41]. There is a freely available viewer for many USGS formats which is a limited-feature	Published standard[43] ASCII Raw data or can be	Suited to data exchange and preservation but is essentially deprecated (but is it – see SDTS).

[38] http://woodshole.er.usgs.gov/operations/sfmapping/SUchirp.htm

[39] http://woodshole.er.usgs.gov/operations/sfmapping/qmips.htm

[40] http://rockyweb.cr.usgs.gov/nmpstds/demstds.html

[41] http://edc.usgs.gov/guides/dlg.html

	version of commercial software called Global Mapper[42]. Like DEM more recent data is in SDTS format. Possibly US centric?		
.dxf DXF: Drawing eXchange Format 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[44] 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[45].	Proprietary but Published (currently) ASCII or Binary Raw and processed	Until recently Version migration was 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 the emergence of OpenDWG, IGES and STEP as described in the recent Digital Image Archiving Study[46]
.dzt RADAN™ DZT GPR: Ground Penetrating Radar	Proprietary, binary format in use with Geophysical Survey Systems, Inc. (GSSI) applications[47]. A DZT limited functionality viewer and a RADAN to ASCII converter are available from GSSI[48].	Published[49] (currently) Binary Raw data	The ASCII export may be suitable for preservation with supporting metadata
.e00 ESRI Export file GIS	'The ESRI E00 interchange data format combines spatial and descriptive information for vector and raster images in a single ASCII file. It is mainly used to exchange files between different versions of ARC/INFO, but can also be read by many other GIS	Proprietary Not published ASCII Processed (usually)	Usable as an exchange format. No longer seen as the best option for preservation.

[43] http://rockyweb.cr.usgs.gov/nmpstds/dlgstds.html

[42] http://mcmcweb.er.usgs.gov/drc/dlgv32pro/index.html

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

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

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

[47] http://www.geophysical.com/

[48] http://www.geophysical.com/softwareutilities.htm

[49] http://pubs.usgs.gov/of/2002/ofr-02-0166/ofr02_166.pdf

	programs. It is a common format for GIS data found on the Internet'[50]. This format is proprietary and not in the public domain. An informal analysis of the format is available[51]. Despite this it has been recognised for many years as the best option for exchange and preservation purposes in being ASCII based and having wide vendor support. A better option is now available in the form of migrating ESRI Shape files directly to GML using GDAL libraries (see below).		
.exp MOSS: Map Overlay Statistical System GIS vector	Use of this original open source GIS package has waned to the extent that it proved quite difficult to track down its source code[52]; however, its export format is still used for data exchange due to support by other packages. 'MOSS export files contain polygon data extracted from the U.S. Department of Interior's MOSS public domain GIS. These consist of points, lines, or closed polygon loops (possibly with islands), and a 30-character attribute field referred to as the subject value'[53].	Published ASCII Processed (usually)	Possible exchange format
various[54] including .fla .sfw .as .asc .flv Flash® 2D, 3D animation	Flash supports 'vector and raster graphics, a scripting language called ActionScript and bi-directional streaming of audio and video'[55]. SFW files are binaries compiled from source FLA files. AS files contain simple ActionScript source code (which can also be embedded in a SFW file) whilst ASC files contain server-side ActionScript. FLV files represent Flash video clips. Adobe® took	Proprietary but published under licence Binary (deliverables) Processed	Not suited for preservation

[50] http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf 5.4

[51] http://avce00.maptools.org/docs/v7_e00_cover.html#OVERALL

[52] ftp://ftp.blm.gov/pub/gis/

[53] http://ahds.ac.uk/preservation/gis-preservation-handbook.pdf

[54] http://en.wikipedia.org/wiki/Adobe_Flash#Related_file_formats_and_extensions

[55] http://en.wikipedia.org/wiki/Adobe_Flash

	over Macromedia and hence Flash and now maintain the specification which is currently freely available under licence[56]		
.gml GML: Geography Markup Language 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[57]. 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[58]	Published standard[59] ASCII Processed	GML is very suited for preservation and data exchange of geospatial data.
various GRASS GIS	Grass is an Open Source package[60]. Like other GIS a vector graphic in recent versions of GRASS is represented by a number of files grouped in a directory; coor, topo, cidx (all binary) and head, dbln, hist (ASCII)[61]. Attribute data is stored in an associated database. GRASS also has GDAL libraries built in so exports to other formats including GML should be an option.	Openly published (see footnote to left) Binary and ASCII Processed (usually)	Not a preservation option. Export to GML if possible
.gsf Generic Sensor Format Bathymetric data	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[62]. As well as the generic it allows attributes specific to a wide range of	Published Binary Raw data	Possible use as an exchange format if widely supported.

[56] http://www.adobe.com/licensing/developer/

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

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

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

[60]

[61] http://grass.itc.it/grass63/manuals/grass63vectorlib_2007_04_21_refman.pdf

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

	bathymetric surveying systems to be included.		
.hsx .hs2 HYPACK Inc. Sidescan sonar Bathymetric data (single beam?)	To quote from the Hypack 'HYSWEEP® survey has a Text logging option (HSX format), allowing raw data to be stored in a format that can be inspected and modified by most editing program (Windows Wordpad for example). Easy inspection of files is the advantage of text logging - the disadvantage is larger files and slower load time. If file size and load time are important to you, it is best to choose the HYSWEEP® binary format (HS2)[63]. The manual also contains format specifications.	Proprietary ASCII (.hsx) Binary (.hs2) Raw data or can be	Text logging provides the basis for a preservation option.
.jgw JGW: JPEG World file	Identical to TIFF World files (see below)		
.las LAS 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'[64]. The American Society for Photogrammetry & Remote Sensing (ASPRS) endorses and supports the use of LAS along with industry stakeholders[65]. Discussions of extending LAS to additionally handle terrestrial laser scanning data are actively taking place[66]. 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[67] [68].	Published[69] Binary Raw data or can be.	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 suited to a long term preservation role as ASCII text alternatives exist.

[63] http://www.hypack.com/documentation.asp

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

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

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

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

	Such usage has not been formalised as yet.		
.map .tab .dat .id .ind MapInfo TAB GIS	MapInfo's native format. It is regulated by MapInfo as a proprietary format and is not openly published. It comprises of a number of related files[70]. These are a mixture of binary and ASCII. The best option for preservation is to export to GML using GDAL libraries (see above).	Proprietary Not published Binary and ASCII files Processed usually	Not suited for data exchange or preservation. Export to GML or MIF with support metadata
.mgd77 MGD77 Geophysical data Including Bathymetric Magnetic Gravity	Described as a 'format for the exchange of digital underway (?) geophysics data'. It was developed by the US National Geophysical Data Center (NGDC) following an international workshop in 1977 'Workshop for Marine Geophysical Data Formats'[71]. Has been revised relatively recently. UNESCO note of MGD77 that the 'format has experienced much success over the last 20 years. It has been sanctioned by the Intergovernmental Oceanographic Commission (IOC) as an accepted standard for international data exchange, and it has been translated by IOC into French, Japanese and Russian. Most contributors of data to NGDC now send their data in the "MGD77" format'[72]	Published ASCII Raw	Possible exchange format Possible preservation format
.mif .mid MIF: MapInfo Interchange format GIS	Like the interchange format of its competitor, ESRI (Export .e00), MIF files are ASCII and support is widespread in GIS applications. The MIF file contains geometric data whilst the optional MID file has header and attribute data as delimited text. The format specification is currently	Proprietary Published (currently)[73] ASCII Processed (usually)	Possible interchange format because of widespread support

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

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

[70] http://en.wikipedia.org/wiki/MapInfo_TAB_format

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

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

[73] http://extranet.mapinfo.com/common/library/interchange_file.pdf

	available from MapInfo.		
.mpg .mpeg MPEG-1: 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.	Published open standard[74] Binary Processed usually	Suitable for preservation. The AHDS currently recommend de-multiplexing of video and audio channels.
.mpg .mpeg MPEG-2 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'[75] and hence much higher quality	Published open standard[76] Binary Processed usually	Suitable for preservation. The AHDS currently recommend de-multiplexing of video and audio channels.
.mpg4 MPEG-4 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'[77].	Published open standard[78] Binary Processed	In being an online streaming standard could be used for dissemination
.mst Marine Sonic Technology MSTIFF Sidescan sonar	Based on TIFF format v. 5. A format specification is currently accessible[79]. Notes that 'Although TIFFs allow for customization of the format, MSTL decided it was better to use the basic structure and create our own MSTL specific tags instead of trying to fit all of our proprietary information into the TIFF'.	Proprietary Published (currently) Binary Raw data or can be	As a general guideline export to more open standards if possible.
.mxf MXF: Material Exchange Format Video Audio	A generic wrapper or container for moving images. Developed as an open standard by the US Society of Motion Picture and Television Engineers (SMPTE). The recent Digital Moving Images and Sound Archiving	Published open standard[81] Binary and ASCII Processed usually	Generally seen as an emerging standard that may have the potential to become a preservation

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

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

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

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

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

[79] http://www.marinesonic.com/documents/mstiff.pdf

	Study undertaken by the AHDS notes that 'MXF is related to the AAF format (see above). It was specifically developed for optimised interchange and archiving of multimedia content. Although currently (2006) too new to be widely used it is emerging as a standard'[80].		standard.
.nc NetCDF: Network Common Data Form 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'[82].Openly published[83]. 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)[84].	Published Binary Raw or can be	This could provide an ideal mechanism for preservation and data sharing through storing once and generating binary or ASCII as requested
.ntf NTF: National Transfer Format Geospatial data including Point cloud CAD Digital Elevation Models (DEM) Lidar	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'. A wide range of NTF converters are available to, for example, popular GIS formats. Lidar data as supplied has often been processed in terms of coordinate transformation and decimation.	Published standard[85] ASCII 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 OS where it is being superseded by GML
.obj OBJ	A simple ASCII based format for representing 3D geometry. Initially developed by Wavefront	Published ASCII Raw data or	Wide support suggests a possible data

[81] http://www.smpte.org/standards/

[80] http://ahds.ac.uk/about/projects/archiving-studies/moving-images-sound-archiving-final.pdf 5.2.3

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

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

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

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

3D including Laser scanning Mesh Point cloud	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[86] we were unable to identify a format maintainer. There are numerous converters available for OBJ files.	can be	exchange format. In being ASCII based it could act as a preservation format
.rd3 .rad RAMAC – RD3/RAD GPR: Ground Penetrating Radar	Used natively by Malå GeoScience equipment. Data is in a binary RD3 file whilst header information is in an ASCII text file. Apart from the file extension they share the same name. A number of open source tools exist for manipulating RD3 files. For example, GPR IDL tools[87] can be down loaded from Source Forge and used to convert files into x, y, z.	Published[88] (currently) Binary (ASCII header) Raw	Move to ASCII text for preservation purposes
.rst .rdc IDRISI raster GIS	Native raster image format for Clark Labs[89] IDRISI GIS software. Associated file types include Raster Documentation (RDC) files. Can be viewed as ASCII.	Proprietary Published[90] (currently) Processed (usually)	
.sd .scene Fledermaus Visualisation of a range of spatially referenced data including Multibeam sonar Digital Elevation Models (DEM = DTM) GIS (various) CAD (.dxf)	Visualisation toolkit for 2D, 3D and movies. Supports the import and export of a large range of spatially referenced data types[91]. A reduced functionality viewer is available for download[92].	Proprietary Binary (part ASCII) Processed	Can represent a project outcome (i.e. a presentation format). Not suited for preservation but can export as a range of formats

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

[87] http://gpr-idl-tools.sourceforge.net/tutorial.html

[88] http://pubs.usgs.gov/of/2002/ofr-02-0166/ofr02_166.pdf

[89] http://www.clarklabs.org/about/

[90] http://www.clarklabs.org/about/

[91] http://www.ivs3d.com/products/fledermaus/

[92] http://www.ivs3d.com/download/iview3d_download.html

Various including .ddf SDTS: Spatial Data Transfer Standard 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[93] (see above) which suggests SDTS is a wrapper around other formats. GDAL (see GML above) support a SDTS Abstraction Library[94].	Published standard[95] Binary Raw data or can be	Well supported as a data exchange standard but probably US centric.
.sg2 .dat SEG 2 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[96]	Openly published[97] Binary Raw data	Possible exchange format. Export to ASCII with suitable metadata if possible
.segy SEG Y Seismic survey including Sub-bottom profiling Sidescan sonar GPR: Ground Penetrating Radar	An openly published format[98] 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	Published Binary Raw data	Can be converted to ASCII for preservation purposes. Possibly useful as a data exchange format as it appears widely supported.

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

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

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

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

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

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

	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[99]. A limited functionality SEG Y viewer can be downloaded from Phoenix Data Solutions[100]		
.shp and lots of associated formats[101] ESRI SHAPE file GIS	Well documented[102] and supported by other GIS vendors such as MapInfo[103]. It has been described as 'developed and regulated by ESRI as a (mostly) open specification for data interoperability among ESRI and other software products'[104]. ESRI Export format (E00) and more recently GML (see entries above) are seen as the best preservation options.	Proprietary Published (mostly?) Binary Processed usually	Because of industry support can usually be used as an exchange format
.svg SVG: Scalable Vector Graphics 2D vector images	XML (and hence ASCII) based format for 2D vector graphics. Specification developed and maintained by the World Wide Web Consortium (W3C)[105]. The specification notes that 'For accessibility reasons, if there is an original source document containing higher-level structure and semantics, it is recommended that the higher-level information be made available somehow, either by making the original source document available, or making an alternative version available in an alternative format which conveys the higher-level information, or by using SVG's facilities to include the higher-	Published open standard XML (ASCII) Processed data	Suited for both dissemination and the long term preservation of simpler 2D vector graphics. It should be noted that Adobe has recently dropped its support for SVG[106] but support is still widespread.

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

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

[101] http://ahds.ac.uk/preservation/gis-preservation-handbook.pdf

[102] http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

[103] http://extranet.mapinfo.com/products/Features.cfm?ProductID=1044&productcategoryid=1#Importing%20and%20Exporting%20Data

[104] http://en.wikipedia.org/wiki/Shapefile

[105] http://www.w3.org/TR/SVG/

[106] http://www.adobe.com/svg/eol.html

	level information within the SVG content. This suggests that for archival purposes use might need restricting to simpler models.		
.tfw TFW: TIFF World file ESRI GIS products (others?)	A mechanism for georeferencing 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[107]. TIFF World files will be small in themselves but may be associated with large images.	Proprietary ASCII (but associated image will be binary) Processed	That the metadata (spatial information is in ASCII could be seen as good for preservation.
.tiff GEOTIFF TIFF 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'[108]. TIFF is also a public domain format currently maintained by Adobe® [109]. It should be noted that the size of a TIFF file is limited to 4GB[110].	Public domain[111] Binary Processed	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.
.txt (see .csv)			
.txt Blue Sky Lidar Point cloud Mesh	Blue Sky are a leading European are described as 'at the forefront of imaging technology and geospatial data'[112]. Straight forward XYZ data (see below)[113]. Lidar data as supplied has often been processed in terms of	? ASCII Raw (ish)	Suited for preservation with suitable metadata

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

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

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

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

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

[112] http://www.bluesky-world.com/index.html

[113] http://www.bluesky-world.com/dem-lidar.html

	coordinate transformation and decimation.		
.vct .vdc .mdb .adc IDRISI vector GIS	Native vector graphics format for Clark Labs[114] IDRISI GIS software. Associated file types include Vector Document (VDC) files, additional attributes within Microsoft Access database (MDB) files and Attribute Documentation (ADC) files. See RST for IDRISI raster format. Can be viewed as ASCII	Proprietary Published[115] (currently) Binary (ASCII export) Processed (usually)	Export to ASCII along with suitable metadata for preservation.
.vpf VPF: Vector Product Format Vector graphics	Developed by the U.S. Defense Mapping Agency. 'The Vector Product Format (VPF) is a standard format, structure, and organization for large geographic databases that are based on a georelational data model and are intended for direct use'[116]. It may be difficult to work with 'Don't try to read VPF unless absolutely necessary. It's a dog of a format'[117]	Open standard[118] ASCII Raw or processed	Possible transfer format but may be complex.
.vtk Visualization Toolkit (VTK)[119] 3D computer graphics, image processing, and visualization and thus Laser scanning	X, Y and Z co-ordinates plus the file header and tail which make the data readable in programs such as Paraview[120] which uses the Visualization Toolkit and is a freely downloadable viewer for the visualization of large data sets such as point clouds. It should be noted that Paraview and VTK support a very wide range of technologies and formats and that recent implementations support XML based formats with older ASCII/binary formats now considered as legacy but still supported. VTK has not been considered here beyond point	Published (i.e. the toolkit is open source) ASCII or Binary Processed (probably)	With an open source viewer the .vtk files supplied by one of the case studies seem a very reasonable way of disseminating point cloud data. That this is now seen as a legacy format might contradict this; however. In being ASCII text suited for preservation of

[114] http://www.clarklabs.org/about/

[115] http://geography.rutgers.edu/courses/06fall/321/Public/Idrisi_Guide/Andes%20Manual.pdf ch. 5

[116] http://www.nga.mil/portal/site/nga01/index.jsp?epi-content=GENERIC&itemID=a2986591e1b3af00VgnVCMServer23727a95RCRD&beanID=1629630080&viewID=Article

[117] http://www.vterrain.org/Culture/vector.html

[118] http://www.nga.mil/NGASiteContent/StaticFiles/OCR/vpf_main.pdf

[119] http://public.kitware.com/VTK/

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

	cloud data supplied by the Breaking through Rock Art case study.		research outcomes.
.wav WaveForm			
.wmv .asf WMV: Windows Media Video Video Audio	Proprietary video and audio codec		
.wrl VRML: Virtual Reality Modelling Language 3D graphics	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[121]. Apparently still popular especially for the exchange of CAD drawings but is slowly being superseded by other standards such as X3D (below)	Published open standard[122] ASCII Processed	Possible exchange format. In being ASCII based has the potential to act as a preservation format but aging.
various X3D 3D graphics	Developed as a replacement for VRML (above) by the web3D consortium[123] 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.	Published open standard[124] ASCII and binary flavours Processed usually	With XML being ASCII based this has archival possibilities.
.xml XML: eXtensible Markup Language Increasing range of technologies	XML[125] is a general-purpose markup language geared towards facilitating the sharing of data. An XML document is said to be 'well formed' and when it conforms to XMLs 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	Published open standard[127] ASCII RAW or processed	Ideal for exchange and preservation if an established schema exists

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

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

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

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

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

	extension, for example .gml (see above). Others such as MIDAS XML developed by the Forum on Information Standards in Heritage (FISH)[126] are explicit in having the .xml extension.		
.xtf eXtended Triton Format[128] 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	Proprietary Binary Raw data or can be	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
.xyz (see .csv)			
.xyz .xyzrgb XYZ Laser scanning	Point cloud data - simply the X, Y and Z coordinates of each scanned point, sometimes with Red, Green and Blue colour values also. XYZ data is often 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 (can be binary) Raw(ish)	ASCII text is seen as the best option for long term preservation along with suitable metadata

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

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

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