Preservation Handbook

Prepared by the

Archaeology Data Service

1

Table of contents

Introduction

The VENUS project aims to develop scientific methodologies and deliver technological tools for the virtual exploration of deep underwater archaeology sites.

Underwater archaeological sites, such as shipwrecks, offer extraordinary opportunities for archaeologists due to their low light, low temperature and a low oxygen environment which is favourable for archaeological preservation. However, these sites are difficult to experience at first hand and are in constant jeopardy from activities such as deep trawling. The VENUS project will improve virtual access to underwater sites by generating thorough and exhaustive 3D records for virtual exploration.

The project team plans to survey shipwrecks at various depths and to explore advanced methods and techniques of data acquisition through autonomous or remotely operated unmanned vehicles with innovative sonar and photogrammetry equipment. Research will also cover aspects such as data processing and storage, plotting of archaeological artefacts and information system management. This work will result in a series of best practice procedures for data acquisition, dissemination and archiving. VENUS will also develop virtual reality and augmented reality tools for the visualisation of and immersive interaction with a 3D digital model of one of the projects chosen underwater sites. This model will be made accessible online as an example of digital preservation, but also for demonstrating new approaches to site investigation in a safe, cost-effective and pedagogical environment. The virtual underwater site will provide archaeologists with an improved insight into the data and the general public with simulated dives to the site.

The VENUS consortium, composed of eleven partners, is pooling expertise in various disciplines: archaeology and underwater exploration, knowledge representation and photogrammetry, virtual reality and digital data preservation. The ADS are the partner with direct responsibility for long term preservation of a subset of the project's digital outputs as well as the development of a Guide to Good Practice (G2GP) on managing marine archaeological data. The G2GP, along with an exemplar archive of project outputs are the final project deliverables from the ADS. This handbook is the first stage along the road of producing both the G2GP and the archive.

Handbook Objectives

The purpose of this handbook is threefold firstly it is to act as a primer on current approaches to digital archiving and preservation for VENUS partners covering generic data types, secondly it is intended to highlight specific issues in the realm of archiving and preservation that are pertinent to marine archaeology drawing on the experience of VENUS partners and the marine archaeology community more widely. Lastly, the handbook provides a starting point for the planning and creation of an exemplar marine archaeological archive, holding examples of everything from raw data to highly processed outputs.

This handbook has been informed by both existing standard practice in the Archaeology Data Service and emerging best practice in the marine archaeology

sector, particularly in the UK. The handbook draws on conclusions and approaches highlighted in English Heritage's recent attempts to address the specific issues surrounding long term preservation and dissemination of marine archaeological data, which included and questionnaire and analysis of current practice[1], as well as a broad range of other sources. It is not intended to be a comprehensive guide to all existing guides to good practice or technical advice that is available on the subject of digital archiving nor to draw them together in one document. Rather, the handbook should be seen as a source both of general information of digital archiving and of references to additional online and published resources.

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

Archival Approaches

Organisations with responsibility for the long term preservation and management of digital data should have well documented archival strategies and procedures in place. Other organisations act as advisory bodies. Documentation can range from generic policy statements through to the quite specific, for example, a series of Preservation Handbooks produced by the UK's Arts and Humanities Data Service (AHDS) and its subject specific data centres[2], including the ADS. Other national and international organisations providing useful documentation in terms of strategies and procedures include the UK Data Archive (UKDA)[3], the British Library[4], the Library of Congress[5], the National Library of Australia[6], the United Kingdom Hydrographic Office (UKHO)[7], NASA's National Space Science Data Centre (NSSDC)[8], the Electronic Resource Preservation and Access Network (ERPANET)[9], The Digital Preservation Coalition (DPC)[10] and the Digital Curation Centre (DCC)[11]. 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.

Open Archival Information System

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[12]. A technical recommendation is also available for consultation on the CCSDS website[13]. 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 both to 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

---

[2] http://ahds.ac.uk/preservation/ahds-preservation-documents.htm
[3] http://www.data-archive.ac.uk/
[4] http://www.bl.uk/about/collectioncare/digpresintro.html
[5] http://www.digitalpreservation.gov/
[6] http://www.nla.gov.au/padi/
[7] http://www.ukho.gov.uk/amd/ProvidingHydrographicSurveys.asp
[8] http://nssdc.gsfc.nasa.gov/
[9] http://www.erpanet.org/
[10] http://www.dpconline.org/
[11] http://www.dcc.ac.uk/
[12] http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=24683&ICS1=49&ICS2=140&ICS3
[13] http://public.ccsds.org/publications/archive/650x0b1.pdf

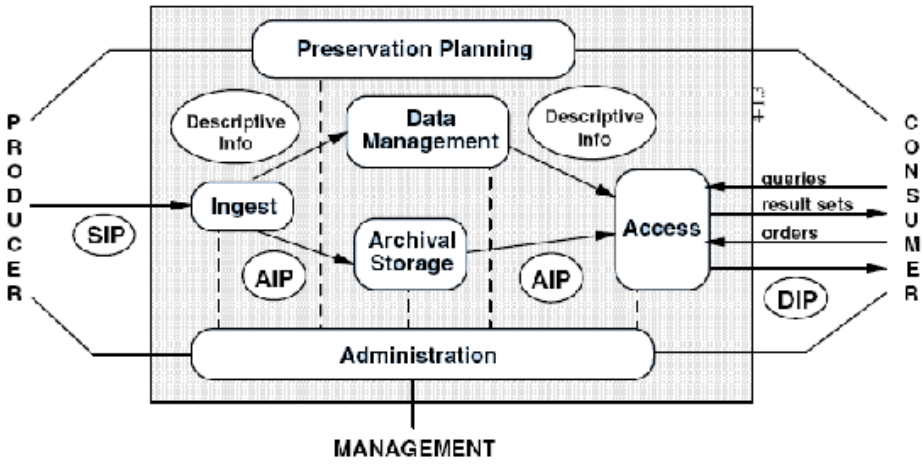

Fig. 1 OAIS Functional Entities (after CCSDS Fig.4.1[14])

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 converted 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'. In practice most repositories offer guidelines to depositors about acceptable formats, delivery media, copyright issues and necessary documentation and metadata. Many existing guidelines will be relevant to VENUS project data but particular issues that have might arise are discussed in following sections.

The most recent development is the publication of a certification document *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist*[15] 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, plus 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.

[14] http://public.ccsds.org/publications/archive/650x0b1.pdf
[15] http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=9

The CRL is currently undertaking a project to test the RLG-NARA metrics through actual audits of subject digital archives and one archiving system[16]. 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[17].

In general the archival community, including the ADS, are actively seeking to become compliant with the reference model through this process of certification. It should, however, be noted that the audit checklist is very recent development. For the time being a state of trust needs to exist between creator and archive, thus for the VENUS project and exemplar archive the ADS approach, modelled on OAIS, is appropriate even though the processes of certification are yet to be fully resolved.

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 digital preservation to life*[18] 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)[19] 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. Beagrie 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[20].

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*[21] and the subsequent development of this framework into a cost model by Tony Hendley in a British Library Research and Innovation Report (106)[22]. 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'[23]. This 1998 report will undoubtedly feed into many archival policies.

---

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

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

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

[19] Terotechnology Handbook (1978) HMSO

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

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

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

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

Other strategies

It is worth looking briefly at alternatives to OAIS, which may in the future be relevant to digital data generated by maritime archaeology projects such as VENUS, although no example of where these alternatives are currently implemented can yet be identified.

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.

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'[24]. 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[25], although some of the 'exhibits' have been used in earnest. In the context of the VENUS project it is likely that substantial amounts of very specific hardware would need to be preserved. Some of this may be data acquisition hardware, but it is especially true with regard to Virtual Reality dissemination outputs that rely on specialist equipment such as head mounted displays, hemispherical displays etc.

Emulation

Rothenberg favours emulation as an alternative preservation strategy. 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[26].

Recommendations

---

[24] http://www.clir.org/PUBS/reports/rothenberg/pub77.pdf (section 6.3)
[25] http://ads.ahds.ac.uk/project/museum/
[26] http://www.dpconline.org/graphics/orgact/storage.html

The long term preservation and dissemination of VENUS project data should ideally be within an OAIS compliant framework (ISO 14721:2003 standard).

Because the certification metrics are very new many archives (including the ADS) are currently working towards OAIS compliance. As such trust must exist between creator and archive.

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.

Acquisition, retention and metadata

Once the long term preservation of digital outputs of any process – marine archaeological or otherwise – has been identified as desirable, then it is best approached as a task right from the initial planning stages of a project. 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.

Many of the software packages associated with the VENUS project data acquisition procedures (see next section) are both proprietary and produce binary format 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 English Heritage's 'Big Data'[27] questionnaire create binary files, these group of respondents are engaged in very similar procedures to the VENUS partners. Fortunately nearly 50% use, or can export, data as ASCII text.

It should be noted that a tension exits between users and archivists 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[28] and Microsoft® Office 2007 (Office Open XML)[29], the Geospatial Data Abstraction Library (GDAL/OGR)[30] is a cross platform C++ translator library for raster and vector geospatial data formats, released under an X/MIT style Open Source license by the Open Source Geospatial Foundation[31]. 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

[27] http://ads.ahds.ac.uk/project/bigdata/
[28] http://xml.openoffice.org/
[29] http://office.microsoft.com/en-us/help/HA100069351033.aspx
[30] http://www.gdal.org/index.html
[31] http://www.osgeo.org/

Markup Language (GML) [32]. 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. This fact may provide an archival solution for some VENUS project 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*[33] undertaken by the AHDS for the, UK Higher Education body, JISC (Joint Information Systems Committee).

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[34]. 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.

[32] http://www.opengis.net/gml/
[33] http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf
[34] http://www.nationalarchives.gov.uk/documents/image_compression.pdf (currently a draft)

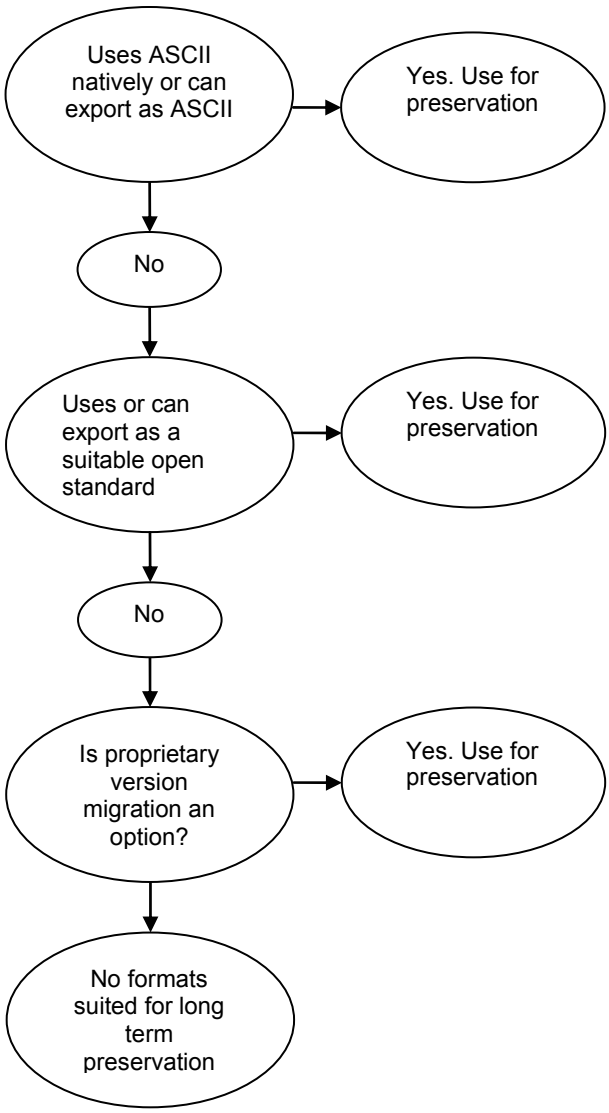

Fig 2. An example software package preservation formats decision tree

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[35]. 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[36].

[35] http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=8446698
[36] http://www.adobe.com/aboutadobe/pressroom/pressreleases/200701/012907OpenPDFAIIM.html

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.

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 can be questioned as information is often implicit within the files themselves; however, this does not facilitate resource discovery and data management and these are key to successful reuse of the data.

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[37]. The recent AHDS Digital Image Archiving Study notes that ISO 19115:2003 Standard for Geographic Information – Metadata[38] and ISO/TS 19139:2007,[39] the XML schema implementation, are the 'ultimate metadata' for GIS data[40]. The relevance here is wider in that the Standard can encompass any geospatially referenced dataset.

Geographic metadata is assumed to have special relevance to the VENUS project given the prominent geographic component to observed acquisition techniques. Essentially nearly all VENUS outputs, raw or processed, have a spatial element whether a region or a specific point.

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.

With regard to harmonisation various VENUS partners' country-specific geospatial metadata standards with the Infrastructure for Spatial Information in Europe (INSPIRE) project's, Draft Implementing Rules for Metadata is already underway. The INSPIRE draft standard[41] maps to ISO 19115 so this should be unproblematic.

The relevant UK standard (i.e. it complies with ISO 19115) would be UK Gemini.

---

[37] http://ads.ahds.ac.uk/project/goodguides/g2gp.html
[38] http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020&ICS1=35
[39] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32557
[40] http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf (section 7.6)
[41] http://www.agi.org.uk/SITE/UPLOAD/DOCUMENT/policy/draftINSPIREMetadataIRv2_20070202.pdf

Much of the information that appears important to the successful management and reuse of VENUS project data does not obviously fit into a UK GEMINI standard, including 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 (or other INSPIRE compliant European equivalent) 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 ISO 19115 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. Some ISO 19115 standards 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.

Other forms of metadata are associated with good archival practice as, for example, indicated in the OAIS Reference Model[42]. The model describes the Preservation Description Information (PDI) 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 but file level metadata such as fixity values and provenance which includes 'processing history' need addressing.

The London Charter

The metadata approaches indicated by the London Charter[43] are also likely to be of significance regarding the VENUS three dimensional dissemination outputs. The London Charter specifies eight principles that should be adhered to in the creation and documentation of three dimensional visualisations for use in the cultural heritage sector. These are as follows:

- **Principle 1- Subject Communities.** The aims and objectives of this Charter are valid across all domains in which 3D visualisation can be applied to cultural heritage. Related specialist subject areas should therefore adopt and build upon the principles established by this Charter.

[42] http://public.ccsds.org/publications/archive/650x0b1.pdf
[43] http://www.londoncharter.org/

- **Principle 2 - Aims and Methods.** Numerous types of 3D visualisation methods and outcomes exist, and can be used to address a wide range of research and communication aims. A 3D visualisation method should normally only be used to address an aim when it is the most appropriate available method for that purpose.

- **Principle 3 – Sources.** In order to ensure the intellectual integrity of 3D visualisation methods and outcomes, relevant sources should be identified and evaluated in a structured way.

- **Principle 4 - Transparency Requirements.** Sufficient information should be provided to allow 3D visualisation methods and outcomes to be understood and evaluated appropriately in relation to the contexts in which they are used and disseminated.

- **Principle 5 – Documentation.** The process and outcomes of 3D visualisation creation should be sufficiently documented to enable the creation of accurate transparency records, potential reuse of the research conducted and its outcomes in new contexts, enhanced resource discovery and accessibility, and to promote understanding beyond the original subject community.

- **Principle 6 – Standard.** Appropriate standards and ontologies should be identified, at subject community level, systematically to document 3D visualisation methods and outcomes to be documented, to enable optimum inter- and intra-subject and domain interoperability and comparability.

- **Principle 7 – Sustainability.** 3D visualisation outcomes pertaining to cultural heritage and created in accordance with the principles established by this Charter, constitute, in themselves, a growing part of our intellectual, social, economic and cultural heritage. If this heritage is not to be squandered, strategies to ensure its long-term sustainability should be planned and implemented.

- **Principle 8 – Accessibility.** Consideration should be given to the ways in which the outcomes of 3D visualisation work could contribute to the wider study, understanding, interpretation and management of cultural heritage assets.

Provenance

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)'[44]. 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[45] which generates MD5 hashes and File Checksum Integrity Verifier (FCIV)[46] 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
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

[44] http://en.wikipedia.org/wiki/Checksums
[45] http://www.fastsum.com/
[46] http://support.microsoft.com/kb/841290

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

The three example tables above and their descriptions are derived from the ADS metadata examples and the Big Data report[47].

Additional 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 VENUS project data where a number of survey techniques involve a series of traverses over a spatially defined area

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

(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 an 'Additional Information Source' or similar element in an ISO 19115 compliant metadata standard to point to such information. A robust and adhered to file naming convention can also reinforce this.

Archival strategy

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

1) Use of software supporting formats with clear migration paths for both preservation and reuse

2) The creation of adequate documentation to facilitate this as well as supporting in house administration and management during the project

Other key elements of an archival strategy are of course access to an adequate hardware system and a robust backup strategy.

Data storage is largely unproblematic for most projects with terabyte external hard drives available for under £300 (c378 Euro). The FISH (Forum on Information Standards in Heritage) Fact Sheet no. 1 'A Six Step Guide to Digital Preservation' provides a brief overview of back up strategies[48]. 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[49] and also maintains copies of data in the UK Data Archive (UKDA) 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. For a project like VENUS it is undeniable that unless large volumes of data are being transmitted onshore during the data acquisition phase of the project there will be a vulnerable window where data is only stored in a single location (possibly at sea). Practical solutions to this issue would have to be covered on a case by case basis.

Comments and recommendations

In order to effectively undertake the long term preservation and dissemination of VENUS data (indeed any data) archival organisations need a well formed Submission Information Package (SIP)

[48] http://ads.ahds.ac.uk/newsletter/issue19/fishsheet1.pdf
[49] http://www.york.ac.uk/services/cserv/offdocs/keynotes/oct01.pdf

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

Although in general ASCII text is seen as the most stable format for data preservation, whilst open binary formats suit the dissemination of some of the formats used in the VENUS project because of a dramatic reduction in file size

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.

It is recommended that a ISO 19115 compliant metadata standard (this may vary from country to country or be covered by INSPIRE) 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

All other documentation that may facilitate reuse should also be included in the SIP

Preservation Intervention Points

In addition to the more general points regarding acquisition and retention, the notion of selection needs special attention in a complex project like VENUS. Where a project has a series of data lifecycle stages where data is transformed by processes such as decimation, aggregation, recasting, and annotation and so on, as well as being migrated from format to format, then there may be more than one point in the process at which intervention for the purposes of preservation might be desirable (preservation interventions points or PIPs).

This is shown diagrammatically below (Figure 3). Going from left to right it can be seen that data streams are initiated by various (hardware based) techniques in the field undergo a series of transformations until the project dissemination products are created. The example stages indicated in the diagram are not comprehensive or definitive (even for the VENUS project) but include:

- Data stream generation – e.g. image capture from ROV cameras, bathymetric survey by sonar, device specific locational information from DGPS and or radio triangulation.

- In-device processing – e.g. sample rates can be altered, say the rate at which images are captured is adjusted, or the lighting conditions are altered. This can be considered processing as it is variable (adjustable) and can, depending on the device, require the discarding of captured information.

- In-field processing – e.g. data is discarded as being outside the area of interest or sample rates are altered (either at this stage or by changing a device variable to alter in-device processing, hence the feedback loop).

- Post-processing – e.g. 1) XYZ coordinates are converted to a triangulated irregular network or used to create a digital terrain model with derived data points 2) captured real world dimensions of an amphora type are used to create an idealised three dimensional model, say using photogrammetry.

- Dissemination versions of three dimensional models are created for specific dissemination modes – e.g. HIVE, HMD, Hemispherical display.

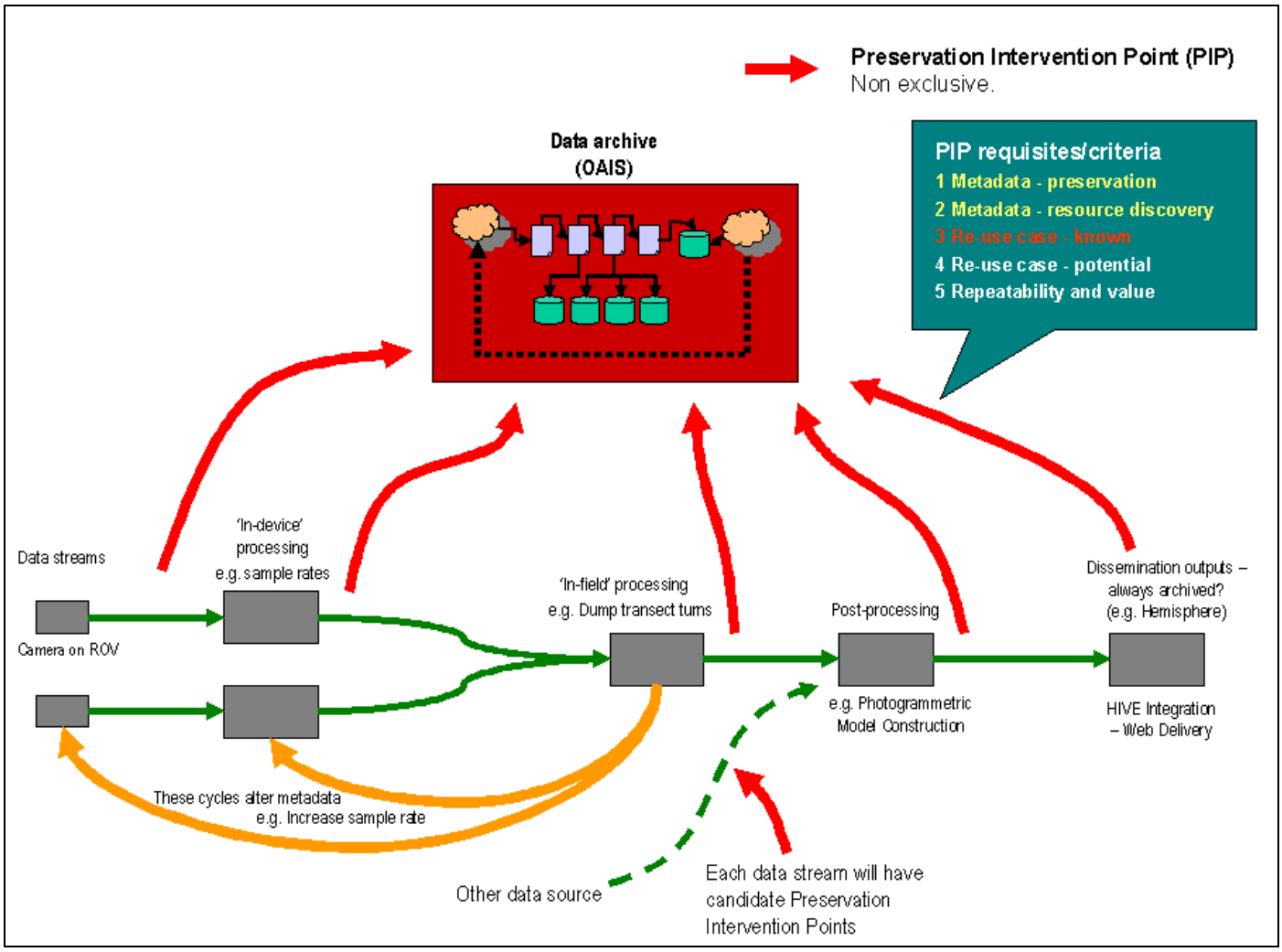

Fig 3. Data streams and preservation intervention points.

While it is clear that this in no way represents the totality of possible stages in the entire VENUS life-cycle it highlights the fact that there are a number of stages where it might be appropriate to intervene to take a preservation copy of the data to be accessioned into an OAIS based archive. Although this it is generally considered good practice that data in as raw a state as possible is ideal for preservation, because the transformations applied can be recreated, however this is not always the case. Just one example of where this idealised approach falls down is in the area of photogrammetry where a series of images are used to construct a three dimensional output. In this case a three dimensional output (say a model of an amphora or a DTM) may be

constructed from a series of high resolution images, the process by which the output was created may or not be proprietary or repeatable, in which case both the original images and the three dimensional outputs would represent preservation intervention points. In more complex processes there may be even more preservation intervention points. Once potential PIPs have been identified they then have to be judged against a series of criteria so the most appropriate PIP(s) for each data stream is identified). The broad criteria by which PIPs are judged are:

1. **Preservation Metadata –** are there appropriate levels of preservation metadata available, i.e. can the data be made actually be reusable rather than simply preservable?
2. **Resource discovery metadata –** are there appropriate levels of resource discovery metadata, i.e. are there meaningful ways of differentiating and discovering this data that distinguish it from other resources? (this mostly applies to legacy data).
3. **Identifiable migration paths –** are there clear AIP and DIP options for this data, i.e. will it fit into an OAIS model?
4. **Reuse cases –** This is probably both the most important criteria and occasionally the hardest to judge. Where the data is in a form that can obviously be used by other researchers or in other contexts then the question is simply whether this is likely to happen, i.e. it CAN BE reused, but is it LIKELY TO BE reused? The other complication is that for certain types of data a reuse case that can be imagined as feasible can be identified, although it is currently not being enacted. An example of this might be a form of data that would lend itself to a post-processing technique which is under development or merely envisaged as being possible in the future (or an enhancement to an existing technique).
5. **Repeatability –** Is the process that created this data repeatable? If yes, an earlier stage may be an appropriate PIP, if not then this intervention point should be selected.
6. **Retention policy –** does the data match the retention policy of the target archive? (See below).
7. **Value –** is the cost of intervening to preserve data at this particular point value for money, given that no project has an unlimited budget. 'Value' here also means the value of the material to be archived – e.g. it might be worth preserving data produced by a repeatable process if that process were particularly expensive and difficult to reproduce. Value is therefore to do with balancing the "value" of the data against the cost of archiving (See section below on archiving costs).

The above criteria are NOT ranked in order of importance and each has to be balanced out against the other. In some case it may be that there is a very strong reuse case for the data, but there is no identifiable AIP, it may be that this point in the process is still selected as a PIP. The process of examining a projects data life cycle (or in the case of VENUS a projects data life cycles) should not be done by the archive alone, this is a consultative process between the archive and the depositor.

Retention

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 the 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 (such as the ADS) 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 marine archaeological data.

Retention and disposal

The question of what to preserve is relevant to all data, but particularly so for the types of data generated by the VENUS project 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 and repeatable it seems unnecessary to keep intermediate data (see the section below on 'Preservation Intervention Points'). 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.

Cost of archiving

For the purposes of the VENUS project the cost of creating an exemplar archive for the Marseille phase of the project has been budgeted for, however in order for the VENUS partners to get an impression of the factors involved, for future reference, the following might be helpful.

So long there are well formed SIPs it is likely that maritime archaeology data is no more consuming in terms of human resource than the archiving of any other data. However, where datasets are large, 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. That being said VENUS project data archives would, for example, fit comfortably within the current ADS charging policy[50] 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. In order to take advantage of technological advances and decreasing costs in certain areas archives have to periodically upgrade systems or parts thereof.

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

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.

From experience at the ADS, 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 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 (c9cents). 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[51]. 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 (c63cents) 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 (c63cents) charge can be reduced significantly. The 'per megabyte' charge is shorthand for what has been described above which might be better described 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

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

As an example, if R = 9p, DR = 3p, E = 4p and DE = 1p (all pence per megabyte charges are for example purposes only – they could equally be though of as cent per megabyte charges) then

Retention period	Cost for refreshment (pence)	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 (c38cents) 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 the equivalent body in whatever archive is being used).

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 the current 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.

Copyright

Any potential copyright issues associated with the VENUS project are not yet clear, (they will be clarified at the November workshop). However, in general survey data is

often owned by third parties. Clearly an external organisation cannot archive this without the permission of the copyright holder which may not be forthcoming if the data still has a commercial value. For the exemplar archive package of VENUS project it is likely to be the case that only material that can be deposited using the standard ADS deposit license will be used[52].

Data transfer

The transfer and ultimate dissemination of VENUS datasets has been suggested as problematic, i.e. how are the large volumes of data generated by partners in the field actually transferred to the archive. Given the capacities of optical media in comparison to the volumes generated this could represent a significant management and administration issue. 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. The future potential for network transfers is discussed below. This may not ultimately be relevant for the VENUS project, but is likely to be a valuable approach for large datasets, such as those generated by marine archaeology, in the future.

[52] http://ads.ahds.ac.uk/project/userinfo/deposit_guidelines/deposit_how.cfm

Data acquisition processes and types

The two completed field missions of the project have represented a progression towards a standardised set of digital survey outputs. Both in terms of the data formats produced and the packaging of the associated metadata each mission's product can be seen as work in progress.

Much of the focus of development in this area has been the integration of the raw data streams as captured from in-water devices. The aim being to combine these into a coherent and usable set of sampled photographic and navigational data which, alongside the relevant bathymetric datasets, can be passed to the photogrammetric modelling stage of the data cycle.

The processes of refining and combining the data streams represent a series of down-sampling stages for the raw data acquired from the devices. Outlined here is the data acquisition schema detailed by ISME, based upon the actual data acquisition process during the first field mission at Pianosa, subsequent non-VENUS surveys undertaken by UNIVPM at Tremiti, and the intended methodology for data acquisition during the Sesimbra mission. This focuses on the automated acquisition by ROV of images and positioning data suitable for photogrammetry.

ROV raw data acquisition

Navigational data

Acquired by an IMU (Intertial Measurement Unit). It is an aircraft unit and has some problems with filtering data in these low speed situations.

- Pitch, roll, yaw etc at 25Hz sample level
    - Pitch
    - Roll
    - Yaw
    - Wx (Vx?)
    - Wy (Vy?)
    - Wz (Vz?)
    - Ax
    - Ay
    - Az
- RPM (motor velocities) - not used
- Depth - 100Hz sample (analogue recording)
- Heading - 20Hz

Imaging data

Digital video stream and concurrent high resolution digital still camera images are collected.

- Analogue video - 3CCD Sony camera - PAL output
    - now using Sony DV camera (TRV-50e - Firewire output, 64mb memory)

- Flash photo still capture via analogue video - 1.3MPx
- NIKON DHS2 still camera (COMEX)
    - 10 MPix
    - Slow flash cycle - 4 sec => 0.25Hz sample
    - Flash memory card

Positional data

Positional data comes from two sources, sonar heads on the ROV itself to record its distance from the seabed and from a Ultra Sound Base Line (USBL) acoustic positioning system.

- Sonar - Kongsberg
    - 2 heads - conical (2.7 deg) and fan (1.7 deg)
    - Only aim is to record distance of ROV from seabed

- USBL (Ultra Sound Base Line) acoustic positioning
    - 0.5 Hz sample
    - Uses GPS or DGPS where available (DGPS 1m error, GPS 7m error)
    - Boat recordings
        - heading
        - pitch
        - roll
        - DGPS position
    - ROV recordings
        - Depth
        - X position
        - Y position
    - also records signal error/variability
    - Using 2 transceivers to track either
        - 2 ROVs
        - ROV and diver
        - Point on site and ROV - allows later correction of GPS positioning

Synchronisation

- Synchronisation signal (BUS) - serial connection (9600 baud)

Low level data collection

- Data capture
- Filter
- Out to shared memory
    - Data
    - Timestamp (milliseconds)

Higher level data collection

- 1Hz cycle
- Collect all data, plus device information, plus timestamp

- Recorded to hard disk drive

JPEG EXIF 'extension'

JPEG frames sampled from the DV-tape (1Hz cycle) have the higher level data recorded in the EXIF structure of the JPEG, using the "User comments" field. EXIF itself does not have an extensible structure, but the inserting structure data into the open "User comments" field allows it to be embedded with the image. Two formats have been used, a full format including a wide range of acquitted data, and a slimmer VENUS format which includes that acquired data which is specifically required for the photogrammetry process – this format is not yet finalised.

- Data is semicolon separated
- Full format
  - x, y, and z from GPS data, corrected to record centre point of image on seabed (reference point)
  - Distance of sonar head to reference
  - Depth (z) of ROV from internal sensor
  - USBL derived ROV depth (z)
  - Full USBL data
  - ROV heading, pitch and roll
  - RPMs (ROV motor velocity)
  - Sonar data
- A VENUS format is under developement. Currently it consists of...
  - timestamp
  - camera calibration data
    - x
    - y
    - z
    - pitch
    - roll
    - heading
    - dist
  - REC ON
  - FLASH ON - to indicate availability of equivalent image from high resolution digital still camera

Example of embedded EXIF data

```
Exif.Image.Model    0x0110 IFD0   Ascii  11 ROV camera

Exif.Image.ExifTag  0x8769 IFD0   Long   1 49

Exif.Photo.UserComment     0x9286 Exif   Undefined 4113

type=FULL;file_name=Captured_1068788.jpeg;index=196;time_stamp_image_relativ
e=1068788.000000;time_stamp_data_relative=1068788.000000;x=0.499696;y=-
0.270246;z=-289.689971;dist=0.100000;z_depth=-
289.390800;z_scout=0.000000;yaw=344.025700;pitch=4.125400;roll=-
1.060200;RPM=0.000000      0.000000      0.000000      0.000000      0.087500
       0.087500;scout_data=105507,4206.39299,N,01529.27257,E,2,12,0.1,-
19.2,M,0.0,M,2.2,0362*49;sonar_data=0.000000   847.000000   166.000000
```

```
182.000000   255.000000   255.000000   255.000000   255.000000
255.000000   255.000000   169.000000   94.000000    47.000000
47.000000    2.000000     18.000000    7.000000     44.000000
42.000000    36.000000    38.000000    3.000000     3.000000
6.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     0.000000     0.000000     0.000000
0.000000     0.000000     1.000000     0.000000     1.000000
0.000000     2.000000     0.000000     3.000000     0.000000
6.000000     0.000000     34.000000    45.000000    145.000000
255.000000   255.000000   255.000000   252.000000   100.000000
45.000000    11.000000    10.000000    3.000000     2.000000
0.000000     1.000000     0.000000     0.000000     0.000000
       0.000000     0.000000     0.000000     0.000000     0.000000;
```

Data structure and extent estimates

The ROV acquired data as the potential to be significantly voluminous. It is important to understand the potential extent of the data and the implications this might have on its preservation.

The ROV data captured is stored in project directory structure with a schema similar to this:

- YYYMMDDD_MISSION_SURVEY
  - INTERNAL
    - YYMMDD_MISSION_SURVEY_DEVICE.TXT
  - IMAGES
    - CAPTURE_TIMESTAMP.JPG
    - ...
  - YYMMDD_MISSION_SURVEY_FULL
    - CAPTURE_TIMESTAMP.JPG
    - ...
  - YYMMDD_MISSION_SURVEY_VENUS
    - CAPTURE_TIMESTAMP.JPG
    - ...
  - COMEX (images from Nikon DHS2)
  - XML

The XML structure is under development, but will contain a minimum of:

- Survey

- Camera calibration data
- timestamp
- JPEG/EXIF from ROV index
- JPEG/EXIF from COMEX index

ROV acquired data brought home from each mission (ideal)

- DV-tape of each survey
- DV-tape of calibration
- device.txt
- Captured JPEG images (800x600) with 'full format' EXIF (300-400kb)
- Captured JPEG images (800x600) with 'VENUS format' EXIF (300-400kb)
- COMEX camera images
- EXIF data etc in XML format

Data extent estimates and calculations

Total collected

10 minutes DV-tape = 1.5Gb
For Pianosa 5 VHS tapes were used => up to 15 hours footage
Estimated 3 DV-tapes - captured video at 10 fps = *c* 400Gb
80Gb navigational data

1 'full format' JPEG EXIF frame = *c* 300kb
With necessary overlap of frames 1m of 2m wide corridor = *c* 600kb =>
10m corridor = 6Mb Estimate of total EXIF jpegs for single run
Pianosa survey would be 60Mb Surveys always completed twice => 120Mb
COMEX still photos (0.25 fps) = *c* 50 frames @ 1.7Mb = approx 100Mb
DV camera stills (*c* 500kb per frame - 1360x1020) = approx 100Mb

Raw DV-tape of 200 second survey (25 fps) = 450Mb Converted to DVD = 130Mb
The raw DV tape records whole survey from location of survey site:

1. Search for site
2. Start record
3. Survey corridor
4. Turn ROV
5. go to 3
6. Stop
7. Repeat.

Only the corridor sections are required for generating the photogrammetry for the site, but there may be valuable material contained in the DV footage relating to areas outside of the survey area. Retention or disposal of this material, based on its potential value, is a question that must be addressed

Estimated total survey time = 60mins Therefore 60 x 1200Mb = total of 12Gb DV-tape

Estimated totals for survey - only sampling corridor survey material:

120Mb EXIF JPEGs
100Mb COMEX camera JPEGs
100Mb DV camera stills (flash)
450Mb DV-tape (130Mb DVD) + calibration data (XML)
= approx 1 GB of data per survey

Pianosa
For the Pianosa mission the data produced represented a mission database, multi beam sonar survey, ROV imaging and positional data and diver generated photography.

Sesimbra
For the Sesimbra mission the data represents a side scan sonar survey in XTF format (processed using SonarWizMAP from Chesapeake Technology, Inc, including navigation mitigation through DGPS interpolation, to produce GeoTIFF), multi beam sonar survey. ROV imaging and positional data and diver generated photography are also represented.

Processed Outputs

Although the above section details the likely raw data forms from VENUS project data acquisition phases, there are a whole range of post-processing techniques being applied to VENUS projects data (as discussed in the Preservation Intervention Points section). The outputs from this phase are likely to include 3D models, text (publications), databases, and animations. Unlike the raw acquired data the management of these data types is much better understood and are unlikely to be problematic, should they be selected for inclusion in the exemplar archive (see the following section). However there are two important points to note:

- The exact processes raw data will undergo (the life-cycle) are still being settled as part of the VENUS research and development process, therefore PIPs cannot be selected with utter confidence, therefore neither can formats of ingest procedures.

- The relationship between INTERACTIVE three dimensional models and their underlying datasets, either 3D datasets or other forms of data is now being recognised as very difficult to capture fully in metadata. This is may require pragmatic approaches to the archiving of this material.

The following section, ‘general points on file formats and file naming’, is derived from ADS’s standard guidelines for depositors[53] which have extensive documentation on data structuring and deposition with the ADS (much of which will be generally applicable to any number of target archives).

General points on file formats and file naming

[53] http://ads.ahds.ac.uk/project/userinfo/deposit.cfm

There are several important things to consider when depositing data; that the files are in the correct format; that proper file naming conventions are used; and that they are accompanied by appropriate documentation.

Check that the files you are depositing are accepted by the ADS (see table below). If Don't include duplicate, draft or spurious files within your deposit. We aim to preserve and disseminate quality data but do not edit or proof read the contents of deposited files. We therefore ask the depositor to make sure that the data deposited is in its final form.

The ADS UNIX operating system requires that certain filenaming conventions should be adhered to when transferring files. Files must have a file extension that helps the ADS and future users of the resource determine the file type, these are normally 3 characters long and we recommend they are in lower case. Do use only alphanumeric characters (a-z, 0-9), the hyphen (-) and the underscore (_). Both upper and lower case characters and numbers can be used in a filename but keep file names within your project consistent and ensure that supplied documentation accurately reflects the case of your filenames. Don't use spaces or full stops (.) within filenames. Full stops should only be present where the filename is separated from the file extension e.g. .doc or .pdf. Spaces can usually be replaced with the underscore (_) character.

Use a consistent scheme and case when naming files. A descriptive filename helps explain the contents of the file, for example `12102007_trench_1.tif` could be a digital photograph of ROV transect 1 taken on 12/10/2007. A non-descriptive file name might be a unique id number allocated to an image within an accompanying image catalogue database. Non-descriptive filenames are acceptable but their content must be adequately described in accompanying metadata. Consistent use of case ensures that files can be reliably identified on case-sensitive operating systems such as UNIX where `report.doc` would be recognised as a different file to `Report.doc`. In either case, it is advised that, regardless of file structure, individual filenames should be unique within a project.

In order for us to undertake proper archiving of your data, we need to have as much information about it as possible, so we would ask that you provide metadata for your files. Include a list of the files you have sent. This should list filename, file size, software package and version used to create the file and a short description of the file's contents. This documentation should be created in digital form to be preserved alongside the data files themselves for future users of the resource.

Below is the standard ADS guide to acceptable file types (which covers most expected data types for processed VENUS data, but not raw)

	File Format		Documentation (**software, version and platform; mandatory,** if relevant)
CAD (Vector graphics)	Preferred	AutoCAD – DWG SVG	• AI, CDR, SVG – Relationship to other documents, caption • DWG, DXF – AutoCAD/DXF version, significance of conventions (layers, colours, linetypes, hatch styles, symbols, etc.), relationship to other files (databases, object libraries, etc.)
	Accepted	Adobe Illustrator - AI DXF CorelDraw - CDR	

Databases	Preferred	Access – MDB OpenDocument Database - ODB	• A data dictionary i.e. a list of all tables and their fields, including the data types and field sizes for text fields, relationships between tables. • Row counts for each table. • For delimited text the delimiters and qualifiers.
	Accepted	Dbase - DBF	
GIS	Preferred	ESRI Shapefile - SHP + SHX + DBF Geo-referenced TIF Image - TIF + TFW	• What is the purpose of the GIS? • What does each layer represent? • Use of Coordinate system or arbitrary site grid? And how the data is • relates to the chosen grid. • Method of capture (Total station survey etc) • Data source (Purchased from OS etc) • Scale/resolution of data capture • Scale/resolution at which data is stored • Assessment of data quality (Root mean square error etc) • Date of capture/purchase
	Accepted	ArcInfo Ungen GML ESRI Grid MapInfo Interchange Format – MIF + MID Spatial Data transfer standard - DDF MOSS - EXP Vector product Format - VPF	
Images	Preferred	Uncompressed Baseline TIFF v.6 -TIF	• Caption mandatory. • Also consider depositing a database or table of metadata, possible fields include: title, photographer, date taken, period, monument type, object type, method, country, district, parish.
	Accepted	Portable Network Graphics – PNG Joint Photographic Expert Group – JPG Graphics Interchange Format - GIF	
Movies	Preferred	MPEG1 & 2 – MPG, MPEG MPEG4 – MPG4	• Name and version of video codec, video dimension (in pixels), frame rate (fps) and bit rate. • Name and version of audio codec - sample frequency, bit-rate & channel information. • Length (hours, minutes, seconds) of file and size. • Copyright clearances where required (e.g. oral history). • Caption and short description for each movie file.
	Accepted	DivX – DIVX, AVI	
Spreadsheets	Preferred	CSV Microsoft Excel – XSL OpenDocument Spreadsheet - ODS	• Purpose & content of spreadsheet & worksheets. • Content of each column and row if not obvious. • Data type and scale for each column. • Key for any codes within data. • Column and row counts. • Documentation of any extra features the spreadsheet may contain, i.e. formulae, macros, charts, comments and any significant characteristics to be preserved.
	Accepted	Lotus 1-2-3 - 123,WK4, WK3, WK1, WKS	
Statistics	Preferred	Delimited text	• Source(s) of the data, collection methodology • Purpose of data • Details of tables and samples (columns) • Number of rows • Type and scale of variables • Full description of any coding used • Analyses performed on data
	Accepted	SPSS - SAV, POR, SPO SAS - SAS7DBAT, SAS Microsoft Excel – XLS OpenDocument Spreadsheet - ODS SYLK - SLK Microsoft Access - MDB	
Texts	Preferred	Word – DOC OpenDocument Text - ODT	• DOC, WPD – software, version and platform • SXW, RTF, ODT – software, version • TXT – text encoding • HTML, XHTML – Software used in creation, doctype with HTML schema • XML – Text encoding, DTD or Schema • SGML – Text encoding
	Accepted	RTF Microsoft Word – DOCX, DOCM OpenOffice.org 1.0 - SXW WordPerfect -WPD TXT HTML, XHTML, XML, SGML	

Virtual Reality	X3D VRML Java3D QTVR	• Original data files that make up the model (image files, CAD models and so on) where available. A video 'fly-by' render of the original VR world to preserve the look and feel.

Exemplar archive

Mapping acquired data to OAIS packages and formats

As stated above, the various VENUS project formats are not yet fixed as the methodologies in use are still being developed, this is after all one of the functions of the project. However, this means that it is impossible to make hard absolute definitive statements regarding the either the most appropriate metadata or the most appropriate file formats for use in archiving marine archaeological data generated by the VENUS project. This section looks at the likely range of data formats and gives an indication of their utility in this respect. This should allow for the mapping process between identified PIPs and the most appropriate data formats for the Submission Information Package (SIP).

Data in the 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'[54]. That some of this information needs to be supplied by the data creator has been discussed above.

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 data formats that are considered to be applicable to long-term preservation and are likely to be appropriate to VENUS project data acquisition, post-processing or dissemination. This table is not intended to be exhaustive, there are undoubtedly other formats suited to preservation and other formats associated with the technologies under consideration. The tables are generated from a combination of field observation with VENUS partners, ADS practice and English Heritage's maritime division (Fort Cumberland) and the 'Big Data' survey of marine archaeological practice in the UK.

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

Format/Properties/Technologies	Description	Comment
ASCII text (.txt, .dat, .xyz, etc) Published standard ASCII Raw (logger)	E.G. Data logger outputs 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.
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[55] 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[56]. 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[57]. 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[58] 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). 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[59].	GML is ideally suited for preservation and data exchange of geospatial data.
MGD77 (.mgd77) Published ASCII Raw or can be Geophysical data including Bathymetric Magnetic	Developed by the US National Geophysical Data Center (NGDC) following an international workshop in 1977[60]. 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'[61]. The MGD77CONVERT toolset allows conversion to the binary NetCDF format[62] which offers an alternative and smaller means of dissemination.	In being ASCII based and published could act as a preservation format. Has support as a data exchange format.

[55] Walker, R. (ed.) 1993. *AGI Standards Committee GIS Dictionary*. Association for Geographic Information
[56] http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=8446698
[57] http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf
[58] http://www.opengis.net/gml/
[59] http://www.gdal.org/index.html
[60] http://www.ngdc.noaa.gov/seg/gravity/document/html/mgd77.shtml#general
[61] http://ioc.unesco.org/iocweb/iocpub/iocpdf/tc045.pdf
[62] http://www.soest.hawaii.edu/GMT/gmt/doc/html/mgd77convert.html

Gravity		
MPEG 1 (.mpg, .mpeg) Published open standard[63] 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'[64].	Suitable for preservation and data exchange.
MPEG 2 (.mpg, .mpeg) Published open standard[65] 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'[66] and hence much higher quality.	Suitable for preservation and data exchange.
MPEG 4 (.mp4) Published open standard[67] 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'[68].	In being an online streaming standard could be used for data sharing.
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'[69].Openly published[70]. 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)[71].	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	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	Wide support suggests a possible data exchange format. In being

[63] http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=25371
[64] http://mediacoder.sourceforge.net/
[65] http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=37679&ICS1=35&ICS2=40&ICS3=
[66] http://en.wikipedia.org/wiki/MPEG-2
[67] http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=38559
[68] http://en.wikipedia.org/wiki/MPEG-4
[69] http://www.unidata.ucar.edu/software/netcdf/
[70] http://www.unidata.ucar.edu/software/netcdf/docs/netcdf/File-Format-Specification.html#File-Format-Specification
[71] http://www.bodc.ac.uk/data/online_delivery/gebco/

Raw data or can be 3D including Laser scanning Mesh Point cloud Photogrammetry	source community. Whilst the format specification is available on numerous websites[72] we were unable to identify a format maintainer. There are numerous converters available for OBJ files.	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[73]. 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.
GEOTIFF (.tiff) Public domain[74] 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'[75]. TIFF is also a public domain format currently maintained by Adobe®[76]. It should be noted that the size of a TIFF file is limited to 4GB[77].	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[78] 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[79]. 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.

[72] http://people.scs.fsu.edu/~burkardt/txt/obj_format.txt
[73] http://support.esri.com/index.cfm?fa=knowledgebase.techArticles.articleShow&d=17489
[74] http://remotesensing.org/geotiff/spec/geotiffhome.html
[75] http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf 1.4.i
[76] http://partners.adobe.com/public/developer/tiff/index.html
[77] http://www.awaresystems.be/imaging/tiff/faq.html#q8
[78] http://www.web3d.org/x3d/specifications/vrml/
[79] http://vads.ahds.ac.uk/guides/vr_guide/sect37.html

X3D (various) Published open standard[80] ASCII and binary flavours Processed usually 3D graphics	Developed as a replacement for VRML (above) by the web3D consortium[81] 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.
XML: eXtensible Markup Language (.xml or can be) Published open standard[82] ASCII RAW or processed Increasing range of technologies	XML[83] 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).	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.

That the data in the SIP is in, or has migration paths to, suitable VENUS project data formats for preservation is essential for the creation of the Archival Information Package or AIP.

The documentation including metadata in the SIP provides the basis of the framework for the successful ongoing management of the data.

[80] http://www.web3d.org/x3d/specifications
[81] http://www.web3d.org/x3d/
[82] http://www.w3.org/XML/
[83] http://en.whttp://www.web3d.org/x3d/specificationsikipedia.org/wiki/Xml

6 Access and use

Dissemination Information Packages

Preservation 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. A well formed SIP will facilitate the generation of the Dissemination Information Package (DIP)[84].

Many of the formats noted as suitable for preservation are also suitable for dissemination. This is the ideal situation; especially for large data sets, 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[85] 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.
SDTS: Spatial Data Transfer Standard (various including .ddf) Published standard[86] 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[87] (see above) which suggests SDTS is a wrapper around other formats. GDAL (see GML above) support a SDTS Abstraction Library for geo-referencing[88].	Well supported as a data exchange standard but may be US centric.

[84] see footnote 70
[85] http://www.ldeo.columbia.edu/res/pi/MB-System/formatdoc/gsf_spec.pdf
[86] http://mcmcweb.er.usgs.gov/sdts/standard.html

SEG Y (.segy) Published[89] 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[90]. A limited functionality SEG Y viewer can be downloaded from Phoenix Data Solutions[91].	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[92] 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.

Dissemination strategies

As with data transfer between creator and archive, the dissemination of VENUS 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.

Other network technologies that were investigated included BitTorrent[93], 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

[87] http://www.fws.gov/data/gisconv/sdts2av.html
[88] http://home.gdal.org/projects/sdts/
[89] http://www.seg.org/publications/tech-stand/
[90] http://pubs.usgs.gov/of/2005/1311/of2005-1311.pdf
[91] http://www.phoenixdatasolutions.co.uk/seisvu.htm

[92] http://www.tritonimaginginc.com/site/content/public/downloads/FileFormatInfo/Xtf%20File%20Format_X21.pdf
[93] http://en.wikipedia.org/wiki/Bittorrent

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'[94]. This could provide a distributed archiving model; however, the reuse of much of the raw or partially processed VENUS 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 limited 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.

Currently the most consistent way of disseminating large datasets is likely to be on portable media; DVDs for the lower end of VENUS project data and external hard drives for anything bigger. As noted already one terabyte portable hard drives are available for under £300 (c378 Euros) 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. VENUS project data throws up some perhaps more unusual mechanisms such as fly-throughs and point cloud models. These are will be project outcomes and tend to use decimated datasets but they will inform on the relevance of the associated raw data. A current example of this approach used by the ADS is the point cloud models produced by the **Breaking Through Rock Art Recording** project. These models are available through the ADS website[95] as Visualisation Toolkit (.vtk) files which can be viewed with 3D visualisation software including the freely available ParaView[96].

Comments and recommendations

The provision of a well formed Submission Information Package or SIP is essential for the successful reuse of data.

That the data in the SIP is in or has migration paths to suitable formats for dissemination is essential for the creation of the Dissemination Information Package or DIP.

[94] http://www.multichannel.com/article/CA6332098.html
[95] http://ads.ahds.ac.uk/catalogue/resources.html?btrar_ahrb_2005
[96] http://www.paraview.org/HTML/Download.html

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

VENUS Partner Workshop

The final preparation phase before material is selected for the VENUS exemplar archive will be the partner workshop on digital preservation scheduled to be held at the King's Manor, University of York on November 3rd 2008. This workshop will be an opportunity for each partner to work through the 'preservation intervention point' decision tree for each of the different sequences of data acquisition, post-processing and dissemination that they are involved in. It is a very important point to bear in mind that although the archive can advise on best practice for the construction of a well formed SIP it cannot, and should not, be the final arbiter on issues such as archaeological or academic value and existing and emergent reuse cases for the data itself. Experience and common sense can be brought to bear on these issues from the side of the archive (especially in relation to financial implications), however the depth and breadth of knowledge required to make sound decisions should be expected supplied from the depositor. Clearly this means that the process of SIP and DIP creation and data deposition is actually a process of discussion and negotiation between the depositor and the archive. The workshop in November will be an opportunity for all the VENUS partners to engage in these discussions in the light of the information on process and format available in this handbook.