Marine Remote Sensing and Photogrammetry
A Guide to Good Practice

By Tony Austin, Jonathan Bateman, Stuart Jeffrey, Jen Mitcham, Kieron Niven

Edited by Kieron Niven
Section 1. Introduction to the Guide

1.1 Guide Objectives
1.2 The VENUS Project
1.3 Archival Strategies

Section 2. Creating Marine Survey Data

2.1 Identifying File Formats from the Outset
2.2 Documenting Data Creation and Processing
2.3 'Big Data' Issues
2.4 Selection and Retention
2.5 Preservation Intervention Points (PIPs)
2.6 Specific Copyright Considerations for Marine Data.

Section 3. Archiving Marine Survey Data

3.1 Data From Marine Projects
3.2 Sonar
3.3 Positional Data
3.4 Navigational Data
3.5 Digital Photography and Photogrammetric Data
3.6 Digital Video
3.7 Structuring and Depositing Your Archive
Section 1. Introduction to the Guide

1.1 Guide Objectives

The objective of this guide is to highlight specific issues in the realm of archiving and preservation that are pertinent to marine archaeology and to provide guidance on the planning and creation of a marine archaeological archive, holding examples of everything from raw data to highly processed outputs. In addition the guide also aims to act as a primer on current approaches to digital archiving and preservation and, in doing so, covers generic concepts such as OAIS, selection, retention and metadata.

This guide has been informed by both existing standard practice in the Archaeology Data Service (ADS) and emerging best practice in the marine archaeology sector, particularly in the UK. The guide also 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 a questionnaire and analysis of current practice, 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, this guide should be seen as a source both of general information on digital archiving and of references to additional online and published resources.

1.2 The VENUS Project

In addition to the sources mentioned above, this guide also draws heavily on ADS involvement in the VENUS project. The project aimed to survey shipwrecks at various depths and to explore advanced methods and techniques of data acquisition through autonomous or remotely operated unmanned vehicles (ROVs) with innovative sonar and photogrammetry equipment. Research also covered aspects such as data processing and storage, plotting of archaeological artefacts and information system management. This work has resulted in a series of best practice procedures for data acquisition, dissemination and archiving as well as demonstrating new approaches to site investigation in a safe, cost-effective and pedagogical environment.

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 aimed to improve virtual access to underwater sites by generating thorough and exhaustive 3D records for virtual exploration.

1.3 Archival Strategies

For organisations with responsibility for the long term preservation and management of digital data it is imperative that well documented archival strategies and procedures are in place. 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, including the ADS. Other national and international organisations providing useful documentation in terms of strategies and procedures include the UK Data Archive (UKDA), the British Library, the Library of Congress, the National Library of Australia, the United Kingdom Hydrographic Office (UKHO), NASA’s National Space Science Data Centre (NSSDC), the Electronic Resource
Preservation and Access Network (ERPANET), The Digital Preservation Coalition (DPC) and the Digital Curation Centre (DCC). 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. It is important for the producers of data, marine or otherwise, to be aware of these systems and concepts as they form the basis for archival strategies and inform selection, retention and preservation policies as well as metadata requirements.

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. A technical recommendation is also available for consultation on the CCSDS website. 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 Figure 1.

![Figure 1: OAIS Functional Entities](after CCSDS Fig.4.1)

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 DOC file might be converted to an XML based format such as an Open Office text document (ODT) 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 are relevant to marine project data and particular issues that 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 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.

The CRL recently undertook a project to test the RLG-NARA metrics through actual audits of subject digital archives and one archiving system. 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 also been undertaken.

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 a recent development and, for the time being, a state of trust needs to exist between creator and archive.

**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 the 2006 conference The LIFE Project: Bringing digital preservation to life, Neil Beagrie’s paper ‘The LIFEcycle model, from paper to digital’ discussed the evolution of lifecycle management from its beginnings in publications such as the Terotechnology Handbook (1978) 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.

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 and the subsequent development of this framework into a cost model by Tony Hendley in a British Library Research and Innovation Report (106). 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’. This report will undoubtedly feed into many archival policies.

The generally recognised categories of the lifecycle of digital assets are
Data creation

Acquisition, retention or disposal

Preservation and management

Access and use

These categories and elements within them provide the framework for this guide.

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 in 'Avoiding Technological Quicksand' (1999, section 6.3) notes the problems associated with this reliance on 'computer museums'. 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, although some of the 'exhibits' have been used in earnest. In the context of marine projects 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 (Digital Preservation Coalition Handbook). An interesting and confusing development came about during the CAMiLEON project which developed a strategy called ‘Migration on Request’ which in fact is emulation with a tool being built to process the original byte stream of a digital object on request.

Interestingly it was recently decided to move the interactive video created in 1986 by the BBC to celebrate the 900th anniversary of the Domesday Book from its dependence on outmoded media and computer hardware. Numbers of experts were approached including the CAMiLEON project who ‘argued that the slight faults in images as displayed from the <original> analogue discs were a part of that experience, and should not be cleaned up’ but the National Archive ‘wanted to preserve the data with the highest quality available consistent with longevity’ and hence opted for migration (see Ariadne issue 36).

Recommendations

The long term preservation and dissemination of marine project data as described in this guide is 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.
Section 2. Creating Marine Survey Data

2.1 Identifying File Formats from the Outset

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.

Binary and ASCII Formats

Many of the software packages associated with marine survey project data acquisition procedures 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' questionnaire create binary files though 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 have previously expressed a preference for binary data in openly published formats because file sizes are significantly smaller, which makes handling and exchanging data easier. Whereas it is clear that representatives from data centres prefer 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 and Microsoft® Office 2007 (Office Open XML), the Geospatial Data Abstraction Library (GDAL/OGR) 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. In short GIS files such as ESRI Shape files and MapInfo files can be migrated to an alternative supported format such as the XML-based Geography Markup Language (GML). 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 marine 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 undertaken by the AHDS for the, UK Higher Education body, JISC (Joint Information Systems Committee).

**Open Standards**

There are, however, 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 unsuitable (see The National Archives Digital Preservation Guidance Note:5 Image Compression). 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.
In addition, 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. It should also be noted that some proprietary formats do develop into open standards. For example, Adobe® announced in 2007 that they had begun the process of turning their very popular Portable Document Format (PDF) into an ISO standard, a decision which, in 2008, resulted in the creation of the ISO 32000-1 standard based upon PDF 1.7.

Figure 2: An example software package preservation formats decision tree
In summary, the file formats used during a marine project will be partly determined by the software or technique being used in their creation. In many cases most data will have migration paths to formats suitable for long term preservation and, where applicable, dissemination (Figure 2). Project-specific examples of these techniques and their relevant outputs are detailed in the Section 3 of this guide. In all cases it is assumed that 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.

2.2 Documenting Data Creation and Processing

In addition to suitable file types, 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. Again, as highlighted with the identification of file types, 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.

The Importance of Metadata

It is important that metadata is created while project data is being actively generated and processed. It is at these points that creators have the clearest idea of what information each file contains, where it was collected, how it was collected and how it was subsequently processed. As well as aiding data management and collection within a project, metadata and documentation is also designed to help others discover, re-use, interpret and manage data.

Metadata can be used to document many different aspects of a project at many different levels. It can be recorded broadly for the project as a whole or for datasets from specific techniques down through to metadata for specific files. When creating metadata it is important, where possible, to identify and adhere to relevant and established metadata and documentation standards. It is important to also realise that metadata standards not only vary according to the relevant techniques but also to geographic region. If long term preservation in a digital archive is one of the intended aims of a project it is therefore necessary to determine the metadata standards required by the archive in question.

This section aims to identify a number of established metadata standards that are specifically relevant to marine project data.

Metadata for Marine Projects

Geographic metadata is identified as having special relevance to marine projects given the prominent geographic component to observed acquisition techniques. Essentially nearly all marine outputs, raw or processed, have a spatial element whether a region or a specific point. The recent AHDS Digital Image Archiving Study, notes that ISO 19115:2003, Standard for Geographic Information - Metadata and ISO/TS 19139:2007, the XML schema implementation, are the 'ultimate metadata' for GIS data. The relevance here is wider in that the Standard can encompass any geospatially referenced dataset.
ISO 19115:2003 defines mandatory and conditional metadata sections, metadata entities, and metadata elements. The standard also defines the minimum set of metadata required to serve a full range of applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data). In addition, optional metadata elements, to allow for a more extensive standard description of geographic data, are described which allows the extending of metadata sets to fit specialized needs.

A relevant UK metadata standard (i.e. it complies with ISO 19115) for geo-spatial datasets would be UK Gemini.

UK GEMINI

Although much of the information that appears important for the successful management and reuse of marine project data does not obviously fit into a UK GEMINI standard, this information can be accommodated by certain elements of the standard. Metadata about the equipment used, settings, methodology, accuracy and software, as described in detail in Section 3 of this guide, 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 such as a brief survey overview. The lack of a relation element in the 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. 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 one of the above metadata standards although file level metadata such as fixity values and provenance which includes ‘processing history’ would still need addressing (see below).
The London Charter

The metadata approaches indicated by the London Charter are also of significance regarding three dimensional dissemination outputs from marine projects. 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.

**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'. Within provenance, reference information
is concerned with unambiguously identifying content information through, for example, the provision of an ISBN number for a publication. In addition, context information (in terms of OAIS) is concerned specifically 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)' (http://en.wikipedia.org/wiki/Checksums). 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. Checksums thus provide 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 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:

<table>
<thead>
<tr>
<th>File Metadata</th>
<th>Comments</th>
<th>Example data</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNIQUE_ID</td>
<td>Auto-generate - unique</td>
<td>1234567</td>
</tr>
<tr>
<td>FILE_LOCATION</td>
<td>Directory + filename</td>
<td>/adsdata/cottam_ba/jpg/fwking_plan.jpg</td>
</tr>
<tr>
<td>CHECKSUM_TYPE</td>
<td>MD5, SHA-1, etc</td>
<td>MD5</td>
</tr>
<tr>
<td>CHECKSUM_VALUE</td>
<td>Generated by algorithm</td>
<td>578cbb18f73a885988426797bcab8770</td>
</tr>
<tr>
<td>PROJECT_ID</td>
<td>Unique project ID</td>
<td>ADS-123</td>
</tr>
<tr>
<td>SURVEY_ID</td>
<td></td>
<td>Laser_05-Jun-2003</td>
</tr>
</tbody>
</table>
The example above 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:

<table>
<thead>
<tr>
<th>Process metadata</th>
<th>Comments</th>
<th>Example data</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROCESS_ID</td>
<td>Auto-generate - unique</td>
<td>1234567</td>
</tr>
<tr>
<td>PROJECT_ID</td>
<td>For example a survey ID</td>
<td>PRO-453</td>
</tr>
<tr>
<td>SOURCE_FORMAT</td>
<td></td>
<td>xyz</td>
</tr>
<tr>
<td>DESTINATION_FORMAT</td>
<td></td>
<td>shp</td>
</tr>
<tr>
<td>PROCESS_AGENT</td>
<td>Who did the processing</td>
<td>Mitcham, J</td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>PROCESS_COMMENTS</td>
<td>Referenced to WGS84</td>
<td></td>
</tr>
<tr>
<td>PROCESS_START_DATE</td>
<td>17-May-2007</td>
<td></td>
</tr>
<tr>
<td>PROCESS_COMPLETION_DATE</td>
<td>17-May-2007</td>
<td></td>
</tr>
<tr>
<td>PROCESS_DESCRIPTION</td>
<td>Import of XYZ data into ArcView for analytical purposes and dissemination as research outcome</td>
<td></td>
</tr>
<tr>
<td>PROCESS_GUIDELINES</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>PROCESS_HARDWARE_USED</td>
<td>Viglen Genie Intel Pentium 4</td>
<td></td>
</tr>
<tr>
<td>PROCESS_SOFTWARE_USED</td>
<td>ESRI Arcview 9.1</td>
<td></td>
</tr>
<tr>
<td>PROCESS_INPUT</td>
<td>/adsdata/pro-453/xyz/file.xyz</td>
<td></td>
</tr>
<tr>
<td>PROCESS_OUTPUT</td>
<td>/adsdata/pro-453/shp/file.shp</td>
<td></td>
</tr>
<tr>
<td>PROCESS_RESULT</td>
<td>Success</td>
<td></td>
</tr>
<tr>
<td>PROCESS_TYPE</td>
<td>See below Conversion - dissemination</td>
<td></td>
</tr>
</tbody>
</table>
The AHDS model restricts process types to a defined list (a lookup table) which should work within a wider setting:

**Figure 4**: Example processing metadata

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

**Figure 5**: Example process types
The three examples above and their descriptions are derived from the ADS metadata examples and the Big Data report.

**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 marine project data where a number of survey techniques involve a series of traverses over a spatially defined area. Composite mosaics can be produced as either part of acquisition or as part of post processing. In the latter case it is clearly critical to document 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.

**2.3 'Big Data' Issues**

The preceding sections attempt to outline two of the cornerstones of an archival strategy for marine 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.

It is a self-evident truth that archaeologists push the boundaries of available computing resources in the course of their work. Access to an exponential increase in computing power has allowed archaeologists to investigate and use a range of technologies that were developed in other disciplines such as the Earth Sciences. Specifically, the ever increasing storage capacity of digital media allows archaeologists to work with larger and larger datasets. Not so long ago to talk of megabytes of data seemed awesome. Today the gigabyte (1000 megabytes) is becoming a relatively common term with some research teams even working with terabytes (1 tb = 1000 gigabytes) of data as is the case with the North Sea Palaeolandscales project.

Examples of Big Data from marine projects include:

- bathymetric survey data such as Multibeam and Side-scan sonar
diver or ROV collected digital video

diver / ROV collected digital images and photogrammetric data

GPS data?

Figure 6: Data collection in the VENUS project using submarine

Is Big Data a Problem?

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

Development at the ADS reflects this. Back in the late 1990s the size of resources were thought of in kilobytes (= 1000 bytes). Today megabytes (= 1000 kilobytes) are the norm with the occasional resources in gigabytes (= 1000 megabytes) but archaeologists are already working with terabytes (=1000 gigabytes). To put this into a wider context, companies such as Google are already working with hundreds of petabytes. A petabyte =1000 terabytes of data which is 1015 or 1,000,000,000,000,000 bytes.

Storing Big Data

Data storage is largely unproblematic for most projects with terabyte external hard drives available for under £200 (c. €235). 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. 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 and also maintains copies of data in the UK Data Archive (UKDA). 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 marine projects 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.

The transfer and ultimate dissemination of marine datasets has thus also been suggested as problematic, i.e. how are the large volumes of data generated in the field actually transferred to an 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 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. Delivery media can of course be supplied or returned. Network transfers, discussed in Section 3.1, are likely to be a valuable approach for large datasets, such as those generated by marine archaeology, in the future.

Creating and Curating Big Data

While storing large datasets has a fairly straightforward solution, actively archiving and curating Big Data has a number of cost implications. While Big Data can be accommodated within existing archival cost models, it would appear astonishingly expensive to archive indefinitely. Interestingly the requirement within the Big Data project was to retain the archives derived from the case studies for five years only (project design 3.2.6). On a five year refreshment cycle data from these archives might only be refreshed once before being discarded. Clearly this needs to be reflected in any charges for archival services. Following on from this, the rise of interest in lifecycle modelling with its emphasis on retention and discard policies seems particularly relevant to the Big Data problem. In many ways this is a clear break with the tradition of trying to preserve everything for as long as possible and could provide the basis of a cost model reflecting the retention period of a resource e.g. archiving a resource for 10 years would require refreshing twice, a period of 15 years would see three refreshments and so on.

The concept of lifecycle models can also be usefully, and perhaps more significantly, applied to Big Data during the data creation stage. As with suitable file formats, the early identification of key points in the creation and processing stage at which data is transformed is key to informing which elements of the final dataset are to be archived. 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 Intervention Points are discussed in more detail in Section 2.5.

2.4 Selection and 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 at the ADS) is generally well documented with archival organisations or data centres providing, for example, collection 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 marine projects 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 Section 2.5 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 and 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

So long as (1) data are deposited in recommended formats (or alternatively that there are migration paths to convert them to such formats), (2) sufficient documentation is provided, and (3) ingest procedures are established, the processing of the large marine datasets should be no more consuming in terms of human resource than the archiving of any other data. Clearly, because of the physical size of the data, it takes much longer to move files around, for example, when moving from delivery media into an archival environment. Similarly, confirming the success of the transfer through generating fixity or checksum values is a much longer process because each byte in a file is referenced. Both of these processes can; however, be run as background tasks. By definition the physical storage requirements will be greater than a more conventional dataset.

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. 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 a five year cycle (noted in the Big Data project). 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 (c. 9 cents). Analysis of past and current trends suggests this will be 1p in five years time and negligible not long after that to be considered as zero cost (See The Decline of Magnetic Disk Storage Cost Over the Next 25 Years). 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:

<table>
<thead>
<tr>
<th>Retention period</th>
<th>Cost for refreshment</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 years</td>
<td>R + E</td>
</tr>
<tr>
<td>10 years</td>
<td>R - DR + E - DE</td>
</tr>
<tr>
<td>15 years</td>
<td>R - 2DR + E - 2DE</td>
</tr>
<tr>
<td>20 years</td>
<td>R - 3DR + E - 3DE</td>
</tr>
<tr>
<td>25 years</td>
<td>R - 4DR + E - 4DE</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Retention period</th>
<th>Cost for refreshment (pence)</th>
<th>Cumulative total (pence)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 years</td>
<td>( 9 + 4 = 13 )</td>
<td>13</td>
</tr>
<tr>
<td>10 years</td>
<td>( 9 - 3 + 4 - 1 = 9 )</td>
<td>22</td>
</tr>
<tr>
<td>15 years</td>
<td>( 9 - 6 + 4 - 2 = 5 )</td>
<td>27</td>
</tr>
<tr>
<td>20 years</td>
<td>( 9 - 9 + 4 - 3 = 4 )</td>
<td>28</td>
</tr>
<tr>
<td>ongoing</td>
<td></td>
<td>30</td>
</tr>
</tbody>
</table>

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 and aims to highlight the possible implications of storing complex or large data types (also discussed in relation to Big Data and Preservation Intervention Points). 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.

### 2.5 Preservation Intervention Points

In addition to the more general points regarding acquisition and retention, the notion of selection needs special attention when dealing with the often complex data collection methods involved in marine projects. Many marine projects feature a series of data lifecycle stages involving data transformation by processes such as decimation, aggregation, recasting, and annotation in addition to data being migrated from format to format. Within these lifecycles 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).

The identification of PIPs is shown diagrammatically below (Figure 6). Going from left to right it can be seen that data streams are initiated by various (hardware based) techniques in the field and then undergo a series of transformations until the project dissemination products are created. The example stages indicated in the diagram are not comprehensive or definitive 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, images capture rates 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.
While it is clear that this in no way represents the totality of possible stages in a marine project 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 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, 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 may 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 created be made 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?

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 2.4 Selection and Retention in relation to archiving costs).

The above criteria are NOT ranked in order of importance and each has to be balanced out against the other. In some cases 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 marine data, a projects data life cycles) should not be done by the archive alone, this is a consultative process between the archive and the depositor.

2.6 Specific Copyright Considerations for Marine Data

That copyright was a major issue first became apparent during the Big Data project. Unlike the VENUS project, which was highly focussed on data collection and processing, many marine projects use survey data that is often owned by third parties. For example, the 'Where Rivers Meet' Big Data case study acquired survey data from the satellite aerial data vendor, Infoterra, who retain copyright. Clearly an external organisation cannot archive this original dataset without the permission of the copyright holder which is unlikely to be forthcoming if the data still has a commercial value. Optimistically this could be seen as a distributed archive with the third party archiving the raw data and digital archives holding the derived data. The problem with this, however, is that reuse by anyone other than the original licensee (in this case Birmingham University who undertook the Where Rivers Meet project) would require them to again purchase the data. Interestingly, Infoterra did offer alternatives to the above scenario such as purchase at a reduced rate.

In addition to the original datasets, data licences may also place limitations on derived data. Restrictions can range from simply placing an acknowledgement and licence number on any derived data sets through to strict restrictions on how much of an original dataset can be reproduced and the methods of reproduction. Copyright issues, as previously noted, should be identified and clearly documented as early as possible in a project.
Section 3. Archiving Marine Survey data

3.1 Data From Marine Projects

The following sections aim to outline a number of significant marine data acquisition techniques, provide examples of the types of data generated in each case and suggest possible preservation issues. These sections also look at the likely range of data formats and give an indication of their utility in archiving marine archaeological data. The techniques covered here are:

- Sonar (Single Beam, Multibeam and Sub-bottom profiling)
- Positional Data
- Navigational Data
- Digital Photography and Photogrammetry
- Digital Video

Mapping data to OAIS packages and formats

Data in the Submission Information Package (SIP) should be in, or have migration paths to, suitable preservation formats and the associated documentation be sufficient to support the creation of an Archival Information Package (AIP). The AIP should consist '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'.

The PDI is defined 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' (OAIS Terminology section 1.7.2). That some of this information needs to be supplied by the data creator has been discussed in earlier sections.

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 tables in the following sections summarise a sample of data formats that are considered to be applicable to long-term preservation and are likely to be appropriate to marine project data acquisition, post-processing and dissemination. These tables are not intended to be exhaustive, there are undoubtedly other formats suited to preservation and other formats associated with the technologies under consideration. These tables have been 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.
Processed and 'intermediate' Data

Much of the focus of development, especially within the VENUS project, has been on the integration of the raw data streams as captured from in-water devices. The primary example here is the combination of data streams into a coherent and usable set of sampled photographic and navigational data which, alongside the relevant bathymetric datasets, can then be passed to the photogrammetric modelling stage of the data cycle. As well as raw data acquisition there will be an analysis phase to any project. For example, survey techniques normally involve a series of traverses over a spatially defined area and composite mosaics can be produced as either part of acquisition or as part of a post processing stage. The composite can then be fed into a range of geospatial tools including 3-D visualization, Geographical Information Systems (GIS) and Computer Aided Design (CAD) software.

Such processes of refining, combining and post processing data streams often represent a series of downsampling stages for the raw data acquired from the devices. This processing of data streams into 'new' data sets poses a number of archival issues in terms of selection and retention and Preservation Intervention Points which have been discussed in previous sections and will be identified in the following data and technique specific sections.

Dissemination Information Packages

A well formed SIP will also facilitate the generation of the Dissemination Information Package (DIP). 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 an ideal environment for this increasingly popular format.

Many of the tables in the following sections note formats considered to be suitable for disseminating data. These are additional to formats already noted as having suitability for preservation and dissemination.

Dissemination strategies

As with data transfer between creator and archive, the dissemination of marine 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 during the Big Data project included BitTorrent, 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, known as the 'tracker'. When the first peers pick up the torrent and download the file(s) using BitTorrent clients they are expected as part of the process to become distributors of a small piece of the file(s). The tracker maintains a manifest of which peer has which part of a file and tells new peers where to download each piece. As the number of peers build up the load is increasingly shifted off the seed computer. Clearly the system needs peers or clients to have largely persistent network connections so that others can access the file fragments.

The above works very well with audio and video data that will have a high download usage and hence lots of potential peers. Research by CableLabs in 2006 suggests that ‘some 18% of all broadband traffic carries the torrents of BitTorrent’. This could provide a distributed archiving model, however, the reuse of much of the raw or partially processed data from marine projects is likely to be an occasional and limited activity. As a consequence, 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 and, 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 marine project data and external hard drives for anything bigger. As noted already, one terabyte portable hard drives are available for under £200 (c. 235 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. Marine project data throws up some perhaps more unusual mechanisms such as fly-throughs and point cloud models. These 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 as Visualisation Toolkit (.vtk) files which can be viewed with 3D visualisation software including the freely available ParaView.

3.2 Sonar

Sonar (SOund NAvigation and Ranging) is a simple technique used by maritime archaeologists to detect wrecks. It uses sound waves to detect and locate submerged objects or measure the distance to the floor of a body of water and can be combined with a Global Positioning System (GPS) and other sensors to accurately locate features of interest.

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

Multibeam sonar sends sound waves across the seabed beneath and to either side of the survey vessel, producing spot heights for many thousands of points on the seabed as the vessel moves forward. This allows for the production of accurate 3D terrain models of the sea floor from which objects on the seabed can be recorded and quantified. Wessex Archaeology used multibeam bathymetry during the Wrecks on the Seabed project (a Big Data project case study). As well as the raw data itself, 3D terrain models, 3D fly through movies and 2D georeferenced images were also created. The 2D images were then used as a base for site plans and divers were able to use offset and triangulation to record other objects on to the plans.

**Side-scan Sonar**

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

![Side-scan Sonar data from the VENUS project](image)

**Sub-bottom profiling**

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

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

Problems and issues

As highlighted above, in many cases it is important that original, unprocessed survey data is effectively archived. As seen in many of the VENUS project missions, where bathymetric data is processed and feeds into a number of subsequent datasets, it is important to identify these key outputs and suitable points in the processing where data should be selected for archiving (Preservation Intervention Points). In most cases archiving the original collected data and significant derived outputs along with metadata recording the processes used to reach each stage is sufficient. It is important, however, that each PIP is identified not only on the basis of it being a final output but also as a point at which the dataset cannot easily be recreated using previous or subsequent incarnations of the data.

Many bathymetric systems use proprietary software. The extent to which this software supports open standards or openly published specifications is largely unknown. Data exchange between systems may also be problematic although there are some open standards such as SEG Y around and many software packages associated with sonar data may support ASCII or openly published binary exports.

Specialised metadata

Metadata to be recorded alongside the data itself includes:

- Equipment used (make and model)
- Equipment settings
- Assessment of accuracy
- Methodology
- Software used
- Processing carried out

Deposition and Archival File Formats

<table>
<thead>
<tr>
<th>Format</th>
<th>Properties/Technologies</th>
<th>Description</th>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCII text (.txt, .dat, .xyz, etc)</td>
<td>Published standard for ASCII Raw data, usually directly from a logger</td>
<td>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.</td>
<td>Suitable deposition and preservation format when stored with supporting documentation.</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>DXF: Drawing eXchange Format (.dxf)</td>
<td>Currently a proprietary published format. Uses ASCII and binary.</td>
<td>The format is published and maintained by AutoDesk, the vendors of AutoCAD. It has been seen for a long time as a de facto standard for the exchange of CAD files but then Autodesk stopped publishing details (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. 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 also GML). Also see OpenDWG, IGES and STEP as described in the recent Digital Image Archiving Study. These emerging standards are not well supported in terms of tools as yet and are thus not recommended here.</td>
<td>DXF usually represents processed data e.g. 3D including Point cloud, CAD and Mesh. ASCII DXF and version migration still seem to be the best preservation option but other options are emerging.</td>
</tr>
<tr>
<td>GML: Geography Markup Language (.gml)</td>
<td>Published standard ASCII format, usually processed data</td>
<td>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.</td>
<td>GML is ideally suited for preservation and data exchange of geospatial data such as GIS CAD</td>
</tr>
<tr>
<td>GEOTIFF</td>
<td>Format is in the Public domain, is binary and usually used to display processed</td>
<td>The GEOTIFF standard is in the public domain. It allows metadata, specifically georeferencing to be</td>
<td>Despite being a binary format TIFF has long been recognised as a de facto</td>
</tr>
<tr>
<td>Extension</td>
<td>Description</td>
<td>Details</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>(.tiff)</td>
<td>data</td>
<td>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 &lt;as preservation format&gt; is the best strategy at the current time, but a watching brief should be maintained on JPEG2000 as an emerging preservation format’. TIFF is also a public domain format currently maintained by Adobe®. It should be noted that the size of a TIFF file is limited to 4GB.</td>
<td></td>
</tr>
<tr>
<td>MGD77 (.mgd77)</td>
<td>Published format using ASCII, usually for raw data</td>
<td>Developed by the US National Geophysical Data Center (NGDC) following an international workshop in 1977 and 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’. The MGD77CONVERT toolset allows conversion to the binary NetCDF format which offers an alternative and smaller means of dissemination. In being ASCII based and published could act as a preservation format but primarily has support as a data exchange format for bathymetric data.</td>
<td></td>
</tr>
<tr>
<td>NetCDF: Network Common Data Form (.nc)</td>
<td>A published binary format, often used to record raw data (or can be).</td>
<td>NetCDF® is a set of software libraries and machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data’. Openly published, and the libraries are 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). NetCDF could provide an ideal mechanism for the preservation and data sharing of bathymetric data through storing once and generating binary or ASCII as requested.</td>
<td></td>
</tr>
<tr>
<td>OBJ (.obj)</td>
<td>Published format based on ASCII, used for storing raw data (or can be)</td>
<td>A simple ASCII based format for representing 3D geometry. Initially developed by Wavefront Technologies, the format is apparently open and has wide support amongst both software vendors and open source community. Whilst the format specification is available on Wide support suggests a possible data exchange format. In being ASCII based it could act as a preservation format for 3D datasets including laser scanning, mesh, point cloud and...</td>
<td></td>
</tr>
</tbody>
</table>
numerous websites we were unable to identify a specific format maintainer. There are numerous converters available for OBJ files.

<table>
<thead>
<tr>
<th>Format</th>
<th>Properties/Technologies</th>
<th>Description</th>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML: eXtensible Markup Language (.xml)</td>
<td>Published open standard based on ASCII. Can be used for raw or processed data and within an increasing range of technologies</td>
<td>XML 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).</td>
<td>Ideal for exchange and preservation if an established schema exists.</td>
</tr>
<tr>
<td>XYZ (.xyz .xyzrgb)</td>
<td>Primarily an ASCII format though can be binary.</td>
<td>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 and such processes should be documented.</td>
<td>ASCII text is seen as the best option for long term preservation along with suitable metadata.</td>
</tr>
</tbody>
</table>

### Specific Dissemination File Formats

<table>
<thead>
<tr>
<th>Format</th>
<th>Properties/Technologies</th>
<th>Description</th>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic Sensor Format (.gsf)</td>
<td>Published Binary format used for Raw data</td>
<td>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.</td>
<td>Possible use as an exchange and dissemination format for bathymetric data if widely supported.</td>
</tr>
<tr>
<td>Format</td>
<td>Description</td>
<td>Comments</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>SDTS: Spatial Data Transfer Standard (various including .ddf)</td>
<td>An Earth Science standard developed by the USGS for data exchange. Downloaded files are a tarred (zipped) directory which, in addition to 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 which suggests SDTS is a wrapper around other formats. GDAL (see GML) support a SDTS Abstraction Library for geo-referencing.</td>
<td>Well supported as a data exchange standard for geospatial data (e.g. DEM, terrain, image) but may be US centric.</td>
<td></td>
</tr>
<tr>
<td>SEG Y (.segy)</td>
<td>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. A limited functionality SEG Y viewer can be downloaded from Phoenix Data Solutions.</td>
<td>Can be converted to ASCII for preservation purposes. Possibly useful as a data exchange format for data from Sub-bottom profiling, Side-scan Sonar and Ground Penetrating Radar as it appears widely supported.</td>
<td></td>
</tr>
<tr>
<td>eXtended Triton Format (.xtf)</td>
<td>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 Possibly very suited for data exchange (e.g. Side-scan sonar, Sub-bottom profiling, etc.) if industry support is widespread. Where possible ASCII text exports with suitable metadata would provide the best long term preservation environment.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example data

Multibeam and sidescan sonar data was generated by a number of missions as part of the VENUS project. For the Pianosa mission the data produced came from multibeam sonar survey and was accompanied by ROV imaging and positional data and diver generated photography. The Sesimbra mission data represented a side scan sonar survey in XTF format (processed using SonarWizMAP from Chesapeake Technology, Inc, including navigation mitigation through DGPS interpolation, to produce GeoTIFF) and multibeam sonar survey. ROV imaging and positional data and diver generated photography are also represented. Multibeam sonar data was also collected by the VENUS Marseille mission and has been deposited with the ADS for archiving. The dataset consists of seven .xyz ascii format files (Figure 7) and is accompanied by a log file recording navigational and positional data (Figure 8).
In both cases, the sonar data has been deposited in formats suitable for ingest and long term preservation.

### 3.3 Positional Data

Datasets containing positional data can come from a number of sources on marine projects and are largely dependent on the data collection methodology. The VENUS project missions, for example, largely focussed on the collection of multiple data streams, including positional data, via unmanned Remotely Operated Vehicles (ROVs). Other projects may use a similar methodology to record the position of divers through a combination of acoustic tracking and GPS.

**ROV Positional data**
Within the VENUS projects, positional data was collected from two sources, sonar heads on the ROV itself to record its
distance from the seabed and from an 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

**Acoustic Tracking**

Acoustic tracking can be used to keep a log of a diver's location throughout the dive. Sound signals are emitted by a
beacon attached to the diver and picked up by a transceiver attached to the side of the boat. The relative position of the
diver underwater can be calculated and these relative co-ordinates can be used to calculate an absolute location for the
diver. Additional equipment may be needed to compensate for the motion of the vessel in the water. Acoustic Tracking
was utilised for the Wrecks on the Seabed project .
Normal practice is to use a data logger for collection. Generally the data will be in the form of structured ASCII text. As such it will be easy to import into other packages such as a GIS or database. Wessex Archaeology supplied their Acoustic Tracking data as a Microsoft Access database.

**Reasons for archiving**

For both Wessex Archaeology's Wrecks on the Seabed project and the VENUS project this data was seen as crucial to the project archive as it sets much of the other project data in context. It forms a vital reference point for the position of the ROV or diver individual photographs were taken, segments of digital video recorded or general observations made.

**Problems and issues**

Positional data may be possibly processed and not the raw data. Any processing carried out should be suitably documented.

**Specialised metadata**

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

**Deposition and Archival File Formats**

<table>
<thead>
<tr>
<th>Format</th>
<th>Properties/Technologies</th>
<th>Description</th>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCII text (.txt, .dat, .xyz, etc)</td>
<td>Published standard for ASCII Raw data, usually directly from a logger</td>
<td>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.</td>
<td>Suitable deposition and preservation format when stored with supporting documentation.</td>
</tr>
<tr>
<td>TFW: TIFF World file (.tfw)</td>
<td>Proprietary ASCII format but associated image will be binary</td>
<td>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. TIFF World files will be small in themselves but</td>
<td>That the metadata (spatial information is in ASCII could be seen as good for preservation.</td>
</tr>
<tr>
<td>XML: eXtensible Markup Language (.xml)</td>
<td>may be associated with large images.</td>
<td>XML 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.</td>
<td>Ideal for exchange and preservation if an established schema exists.</td>
</tr>
</tbody>
</table>

### Example data

Within the VENUS project positional data combined with photography to create Tiff world files for use in photogrammetry and rapid mosaicing. In each case the image was stored with an ascii .tfw file containing the locational information (figure 9).

![Figure 11](image.png)

**Figure 11**: Sample locational data within a .twf file.

### 3.4 Navigational Data

In addition to positional data, navigational data relating to an ROV or vessel is often collected and again provides context for data recordings or observations.

Within the VENUS project, navigational data was acquired via an IMU (Inertial Measurement Unit). Such units record motion in a number or terms, as detailed below:

- Pitch, roll, yaw etc.
- Pitch
- Roll
Yaw
Wx (Vx?)
Wy (Vy?)
Wz (Vz?)
Ax
Ay
Az
RPM (motor velocities)
Depth
Heading

**Reasons for archiving**

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

**Problems and issues**

As with positional data, it is important to document any processing carried out on the original datasets.

**Specialised metadata**

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

**Deposition and Archival File Formats**

<table>
<thead>
<tr>
<th>Format</th>
<th>Properties/Technologies</th>
<th>Description</th>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCII text (.txt,.dat,</td>
<td>Published standard for ASCII Raw data, usually directly from a</td>
<td>E.G. Data logger outputs as structured ASCII text and incorporated into a database. There are well established archival</td>
<td>Suitable deposition and preservation format when stored with supporting</td>
</tr>
</tbody>
</table>
Example data

As highlighted in the Sonar section, navigational and positional data was recorded alongside Multibeam Sonar data. Figure 10, below, is a short sample from this log.

```
08;10;11;12;25;22;980#$PLSM,05,703215.42,4782495.06,96.60,11/10/2008
12:25:19'421
08;10;11;12;25;22;937#$PLSM,05,703215.37,4782495.13,96.64,11/10/2008
12:25:21'921
08;10;11;12;25;22;953#$HEHDT,31.471,T*1F
08;10;11;12;25;22;894#$PIXSE,ATITUD,-0.574,6.424*4D
08;10;11;12;25;23;000#$PIXSE,POSITI,43.16784270,5.49971093,-111.947*74
08;10;11;12;25;23;031#$PIXSE,SPEED_,0.021,0.036,-0.045*4B
08;10;11;12;25;23;046#$PIXSE,UTMWGS,T,31,703196.129,4782487.123,-111.947*34
08;10;11;12;25;23;062#$PIXSE,HEAVE_,-0.141,0.071,0.013*54
08;10;11;12;25;23;078#$PIXSE,STDHRP,0.039,0.002,0.002*7A
08;10;11;12;25;23;093#$PIXSE,STDPOS,39.18,39.02,291.22*47
08;10;11;12;25;23;109#$PIXSE,STDSPD,0.062,0.062,0.058*70
08;10;11;12;25;23;140#$PIXSE,ALGSTS,00000341,00000000*63
08;10;11;12;25;23;156#$PIXSE,STATUS,03000001,00000000*6D
08;10;11;12;25;23;187#$PIXSE,HT_STS,01FE5551*46
```
3.6 Digital Video

Digital video is becoming more and more popular as a means of recording archaeology, particularly amongst maritime archaeologists where sites are less easily accessible than those on dry land. If a whole dive is recorded in this way by means of a hat- or ROV-mounted camera, the generated file is likely to be substantial in size. Particularly with marine projects, digital video may or may not be associated with digital audio.

Digital video can also be used to record events related to marine archaeology, for example to record condition surveys, procedures and interviews. Although Internet Archaeology noted in 1997 (EVA Conference paper) that few archaeologists have access to the technology to create digital video, this situation has moved along quickly. With digital cameras that also record video becoming cheaper and more accessible to a wider range of users plus the easy availability of video editing software, it is likely that use of digital video within archaeology will only increase in popularity.

Digital video is often originally recorded onto DV tape. This is the most economical way to store it but it should be noted that tape degrades relatively quickly. Also video tape is rapidly being superseded by disc based technologies both in terms of recording and viewing. Video can also be transferred to disk based storage where large scale storage devices are increasingly affordable. As noted above, storage and file size can be an issue, for example round 1 of Wessex Archaeology's Wrecks on the Seabed project produced somewhere in the region of 75 gigabytes worth of dive footage.

The recent AHDS Digital Moving Images and Sound Archiving Study provides additional guidance.

Reasons for archiving

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

Problems and issues

Digital video can be very long e.g. the Wrecks on the Seabed project produced around 40 hours of digital video. If 'cleaner' edited versions of video footage are produced, e.g. showing some of the highlights of the dive, it is important to make informed decisions on what version(s) should be preserved and in what quality format. In addition, DV tape has an increasingly short lifespan and should be migrated to disc based storage. DVDs also have a finite lifespan with
hard drives (internal or external) providing the securest medium for storage. Again, in deciding what versions are migrated it is important to prioritise original rather than derived footage.

Specialised metadata

Digital video in an underwater context may be associated with some record of where the diver, and thus the camera, was at any one time. This positional data is discussed elsewhere in this guide. The metadata required specifically for digital video should cover:

- Software, version and platform
- Name and version of video codec (where appropriate)
- Video dimension (in pixels)
- Frame rate per second (fps)
- Bit rate
- Name and version of audio codec including sample frequency, bit-rate and channel information
- Length (hours, minutes, seconds) of file
- File size

Deposition and Archival File Formats

<table>
<thead>
<tr>
<th>Format</th>
<th>Properties/Technologies</th>
<th>Description</th>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPEG 1 (.mpg,.mpeg)</td>
<td>A published open standard, binary format for video and audio</td>
<td>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'.</td>
<td>Suitable for preservation and data exchange.</td>
</tr>
<tr>
<td>MPEG 2</td>
<td>Published open standard</td>
<td>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</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Suitable for preservation and data</td>
</tr>
</tbody>
</table>
Example data

Many of the VENUS project missions acquired digital video via ROVs. In many cases this data had the potential to be significantly voluminous and it is important to understand the potential extent of the data and the implications this might have on its preservation. For example, the Pianosa mission collected up to fifteen hours of footage.

From each VENUS mission the following ROV data would be acquired:

   DV-tape of each survey  
   DV-tape of calibration  
   device.txt (documentation)

When collecting digital video for later use in photogrammetry a set procedure was followed with the raw DV tape recording the whole survey from the location of the survey site:
1. Search for site  
2. Start record  
3. Survey corridor  
4. Turn ROV  
5. go to 3  
6. Stop  
7. Repeat.

While only the corridor sections of the video were required for generating the photogrammetry for each site, as discussed above 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.

3.5 Digital Photography and Photogrammetric Data

High resolution digital still camera images are a common type of data from marine projects and often form the basis for later photogrammetric models. Within the VENUS missions such images were captured concurrently with digital video.
by the ROV, though images can equally be taken by divers (e.g. the VENUS Pianosa mission). In other cases digital still images may also be derived from digital video as discussed below and in the section on Digital Video.

While digital images themselves are fairly straightforward to manage and archive, the use of these images as the basis for photogrammetry creates significantly more complicated lifecycles for these files in which Preservation Intervention Points (PIPs) play an important role. Two such stages that are highlighted by the VENUS project are the embedding of related data streams (e.g. positional and sonar data) and the photogrammetric rectification for rapid mosaicing.

**Photogrammetry**

In order to aid and semi-automate subsequent photogrammetry, the VENUS project explored the potential of embedding sonar and positional data within the EXIF metadata structure of JPEG files. Although EXIF itself does not have an extensible structure, by inserting structured data into the open ‘User comments’ field additional metadata, e.g. positional data, could be embedded within the image file. The combining of data streams at this stage represents a possible PIP and preservation of this distinct set of data should be considered alongside strategies and methods to preserve the other data streams.

Subsequent to embedding the data streams, the processing and rectification of images for the use in photogrammetric modelling represents an additional stage in the images' lifecycle. The VENUS project datasets automated this process using the images with embedded EXIF data to generate TIFF world files. Again, however, at this stage any processing (e.g. colour correction) or rectification of an image may result in a loss of quality.

![Figure 13: Still image from the VENUS project showing amphora on the seabed.](image)

**Reasons for archiving**

As with digital video, digital images provide an important record of what was seen by the diver or ROV at particular points on the underwater site, especially when associated with positional data. In contrast to digital video, however, digital photography can also provide higher quality images and in many cases are easier to manipulate and access.
than large digital video files. When used as the source for photogrammetric models, digital image collections can provide site plans to compliment bathymetric datasets.

**Problems and issues**

As photogrammetric processing requires that images are rectified and often results in diminished image quality, it is important to retain the original image files in order to enable future re-processing and analysis. With long lifecycles featuring processing at multiple points, as seen in the VENUS project, it is important to identify PIPs at which the images and associated data should be saved. With the VENUS data lifecycle it was decided that, aside from the final photogrammetric models, the original image files should be archived separately alongside the other data streams as subsequent outputs could be easily recreated from these component parts. At whatever point it is decided that data should be archived it is, however, also important to document the processing which has been carried out on the data.

**Specialised metadata**

Metadata to be recorded alongside the data itself includes:

- Equipment used (make and model)
- Equipment settings
- Assessment of accuracy
- Methodology
- Software used
- Processing carried out
- Image specific metadata i.e. subject, time, date, etc.

**Deposition and Archival File Formats**

<table>
<thead>
<tr>
<th>Format</th>
<th>Properties/Technologies</th>
<th>Description</th>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCII text (.txt,.dat, .xyz, etc)</td>
<td>Published standard for ASCII Raw data, usually directly from a logger</td>
<td>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</td>
<td>Suitable deposition and preservation format when stored with supporting documentation.</td>
</tr>
<tr>
<td>Format</td>
<td>Description</td>
<td>Details</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td><strong>GEOTIFF (.tiff)</strong></td>
<td>Format is in the Public domain, is binary and usually used to display processed data</td>
<td>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 &lt;as preservation format&gt; is the best strategy at the current time, but a watching brief should be maintained on JPEG2000 as an emerging preservation format'. TIFF is also a public domain format currently maintained by Adobe. It should be noted that the size of a TIFF file is limited to 4 GB. 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. It is generally used to store processed data for display.</td>
<td></td>
</tr>
<tr>
<td><strong>OBJ (.obj)</strong></td>
<td>Published format based on ASCII, used for storing raw data (or can be)</td>
<td>A simple ASCII based format for representing 3D geometry. Initially developed by Wavefront Technologies, the format is apparently open and has wide support amongst both software vendors and open source community. Whilst the format specification is available on numerous websites we were unable to identify a specific format maintainer. There are numerous converters available for OBJ files. Wide support suggests a possible data exchange format. In being ASCII based it could act as a preservation format for 3D datasets including laser scanning, mesh, point cloud and photogrammetry</td>
<td></td>
</tr>
<tr>
<td><strong>TFW: TIFF World file (.tfw)</strong></td>
<td>Proprietary ASCII format but associated image will be binary</td>
<td>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. 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.</td>
<td></td>
</tr>
<tr>
<td><strong>XML: eXtensible Markup Language</strong></td>
<td>A published open standard based on ASCII. Can be used for raw or processed data and within an increasing range of</td>
<td>XML 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 ideal for exchange and preservation if an established schema exists.</td>
<td></td>
</tr>
</tbody>
</table>
Example data

The image and photogrammetric data from the VENUS Marseille mission consist of a number of different datasets from different stages of data collection and processing. Firstly there is the original unprocessed images collected by the ROV, these were then combined with the relevant positional data (also collected by the ROV as XML), stored in each image using EXIF metadata (Figure 11), to created a second dataset.

Lastly these files were again processed to create TIFF world files for use in rapid mosaicing. The TIFF files and associated text files (which hold the relevant positional data for each images) are considered the third dataset and additional still images, created from the rapid mosaicing, represents a forth and final stage (Figure 12).
3.7 Structuring and Depositing Your Archive

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 relevant archive. Don't include duplicate, draft or spurious files within your deposit. Many archives will not edit or proof read the contents of deposited files and it is therefore important that the data deposited is in its final form.

Many archives, the ADS included, require that certain filenaming conventions be adhered to when transferring files. In the case of depositing with the ADS, files must have a file extension that helps the ADS and future users of the resource determine the file type. Such extensions are normally 3 characters long and are usually in lower case. Do use only alphanumeric characters (a-z, 0-9), the hyphen (-) and the underscore (_) when naming files. Both upper and lower case characters and numbers can be used in a filename but keep filenames 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 an archive to undertake proper archiving of your data, it is important that as much information as possible is recorded about your project. The project and file level metadata required by an archive can vary and has been discussed in general terms in Section 2.2. File level metadata, as previously highlighted, is largely dependant on the filetype but a list of the files that have been deposited should be included. This should list filename, file size, software package and version used to create the file and a short description of the file’s contents. Fixity values, as previously discussed, should also be included. This documentation should be created in digital form to be preserved alongside the data files themselves for future users of the resource. As with file level metadata, project level metadata requirements vary and a number of standards currently exist as detailed in Section 2.2.

Issues with the physical transfer of marine datasets have been discussed previously (Section 2.3 and Section 3.1) and individual archives will be able to specify exactly how they require datasets to be deposited.

Information and requirements on depositing data with the ADS can be found in the ADS Guidelines for Depositors.