

Archiving Archaeological Spatial Data: Standards and Metadata

Robert Shaw¹, Anthony Corns¹, and John McAuley²

¹ The Discovery Programme, Ireland

² Digital Media Centre, Dublin Institute of Technology, Ireland

Abstract

The Spatial Heritage & Archaeological Research Environment I.T. (SHARE IT) project was a collaborative venture supported by the Heritage Council (Ireland) under its Irish National Strategic Archaeological Research Programme 2008. Partners were drawn from research (the Discovery Programme), academia (School of Archaeology, UCD; Digital Media Centre, DIT), and archaeological consultancy (Margaret Gowen Ltd) with the aim to develop a strategy for the archiving and dissemination of spatial archaeology landscape data sets, initially LIDAR, aerial orthoimagery and geophysical survey. The project goal was to develop a pilot web mapping application tool for data exploration and use in further research. One of the key research challenges was to identify a suitable digital archiving strategy for spatial landscape data and this was approached by a review the current best practices that have been adopted within the cultural heritage sector and within the wider professional community. Standards organizations specific to cultural data such as the Archaeological Data Service (ADS) and ARENA (Archaeological Records of Europe Networked Access) were consulted on their prescribed policy. Issues addressed included:

- What are the adopted data formats and standards for the sharing and long term archival preservation of digital spatial data?
- Are there any prescribed metadata formats associated with the storage of digital archaeological and spatial data that should be adopted?
- Are there any standards organizations that can assist and integrate Irish digital spatial data into an international framework?

This paper discusses the findings of this process and how it shaped our recommendations for the management of spatial archaeological landscape data and the development of an archiving policy. Amongst the topics it will consider are:

- The importance of the OAIS model as an archival system.
- The need for metadata schema compliant with international standards such as ISO 19115 and INSPIRE.
- The advantages of expanding keyword fields to include controlled vocabularies and thesauri to standardize the description of geographical and cultural components.
- The definition of “preferred” data formats for archiving and the dissemination of information.
- The need for a comprehensive copyright and access policy to accompany the archiving process.
- The financial implications and cost models available to calculate the lifecycle costs of implementing an appropriate archival strategy.

In conclusion the paper will consider how the understanding gained from this approach to archiving spatial data may be applied to a wider range of cultural digital resources.

Key words: *OAIS, cultural heritage, spatial data, thesauri, archiving*

1 BACKGROUND

In Ireland, over the past 15 years, much financial and professional effort has been invested in the collection and analysis of spatial archaeological

data by government, research and commercial sectors. Within this digital asset, landscape data forms a substantial component and includes: aerial photography; topographic surveys created by both LiDAR (Light Detection and Ranging) and digital photogrammetry; and geophysical surveys. Once this data is recorded and interpreted, the printed

report is often seen as the final deliverable, while the digital assets created frequently remain hidden and unused within the source organizations, eliminating any possible knowledge transfer to the wider archaeological community. In the current economic climate the possibility for the loss of archaeological information is great as the digital data collected and held by commercial companies could potentially disappear.

Recently, several reports¹²³⁴ reviewing the current archaeological research framework within Ireland highlighted concerns that exist within the archaeological community. Some major problems to the successful development of the knowledge society in Irish archaeology were identified including:

- Underdeveloped and poorly resourced research infrastructure.
- The disconnected nature of archaeological information and key resources within the archaeological research community.
- A lack of accessible and sustainable digital archives for archaeological data, conforming to established standards and metadata.
- An inadequate return on the investment in primary data collection, from both development led and grant funded archaeological practice, resulting in the production of hidden archaeological material.

A potential solution to these problems lies in the creation of an effective complimentary ICT strategy to provide easy access to primary research information whilst offering a sustainable and robust digital archive that adheres to recognised international standards. Developments in Geographical Information Systems (GIS) have provided researchers with new mechanisms to access improved archaeological data sets. The

tools within GIS enable the visualisation, cataloguing and analysis of a varying scale of spatial data improving the investigative capacity of the researcher. Creating a coherent Spatial Data Infrastructure (SDI) where high quality landscape data is easily accessible will maximise the knowledge return from this resource and enhance future archaeological research.

2 THE SHARE IT PROJECT

The Spatial Heritage & Archaeological Research Environment I.T. (SHARE IT) project was a collaborative venture supported by the Heritage Council (Ireland) under its Irish National Strategic Archaeological Research Programme, 2008. Its main aims were to develop a strategy for the archiving and dissemination of landscape archaeology data sets (LiDAR, aerial imagery and geophysics) using ICT, (see fig. 1). The project hoped to bridge the gap between the potential and actual use of digital data, by providing access not only to the data, but to the technology to exploit it. The tasks were divided into six interlinked work packages, their content and connectivity building towards understanding, assessing, designing and implementing an appropriate ICT solution for the sharing and reuse of spatial archaeological landscape data.

¹ Gabriel Cooney, *Archaeology in Ireland: A Vision for the Future*, ed. (Dublin: Royal Irish Academy, 2006).

² *A Review of Research Needs in Irish Archaeology*, (Kilkenny: The Heritage Council, 2007).

³ Alison Harvey, *The Heritage Council Strategic Plan 2007 – 2013 Consultation Document*, (Kilkenny: The Heritage Council, 2006).

⁴ Roberta Reeners, ed., *Archaeology 2020. Repositioning Irish Archaeology in the Knowledge Society*, (Dublin: University College Dublin, 2006).



Figure 1. Examples of archaeological spatial data held by the Discovery Programme. From the left; orthoimagery of a neolithic settlement at Mullaghfarna, Co Sligo; magnetometry survey of an enclosure complex, Carns Co Roscommon; high resolution LiDAR hillshade model of part of the Hill of Tara earthworks, Co Meath.

The six work packages were:-

WP1: domain analysis – This module aimed to clarify the current situation and state of the digital archaeological landscape record that exists within the many commercial, government and research institutes within the island of Ireland. Components of this work package included the construction of a questionnaire to be completed by the archaeological community, and interviews with leading figures in Irish archaeology.

WP2: international best practice review – The second module explored and reviewed the current best practices that have been adopted by the cultural heritage sector and the wider professional community, particularly standards organizations specific to cultural data. Policy and standards outside of cultural heritage, such as engineering, were also examined, and the review included an examination of current legislation governing the sharing and reuse of spatial data, specifically the EU INSPIRE (Infrastructure for Spatial Information in Europe) directive⁵.

⁵ Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), in Official Journal of the European Union, http://www.ec-gis.org/inspire/directive/1_10820070425en00010014.pdf (accessed May 6, 2009).

It is the research for this work package that forms the basis of the core discussions in this paper.

WP3: system analysis & design – This module of the project dealt with the technical preparation for the successful development of the web mapping application pilot.

WP4: webgis pilot development – This was concerned with the construction of a web mapping application pilot using the acquired knowledge gained from WP1, 2 and 3. The methodology followed the setup, development and multiple iterative testing phases of the web mapping application pilot. During this process the WebGIS was seeded with core amount of landscape data from the Discovery Programme, Margaret Gowen & Co and UCD School of Archaeology.

WP5: dissemination – A combination of a project website, hosting public seminars and presentations at selected conferences / workshops were used to help disseminate the project to the archaeological community.

WP6: exploitation – For the web mapping application to have a sustainable existence following the pilot, a review of possible funding mechanisms and supporting actions was commissioned.

3 ARCHIVING DIGITAL DATA

Any research into archiving digital data quickly leads to the reference model of the Open Archival Information System (OAIS). The Consultative Committee for Space Data Systems (CCSDS) was formed in 1982 by the major space agencies of the world, including NASA, to provide a discussion forum for common problems in the development and operation of space data systems. One outcome has been the recommendation of standards for the preservation of space related data through the OAIS reference model. It defines the basic functional components of an archive and provides a comprehensive framework for describing and analysing preservation issues.

“An OAIS is an archive, consisting of an organization of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community” (CCSDS 2002, 1-1).

In 2002 OAIS was approved as ISO standard 14721⁶, to establish a system for archiving information, both digitalized and physical data.

Our discussions with Archaeological Data Services (ADS) and subsequent research of their publications, in particular the ‘Big Data’ project⁷ emphasized the benefit in using the OAIS as an archiving model. This is particularly the case when collaborating with external organizations as it provides a language and a set of terms that can aid communication. OAIS emphasizes the requirement for ongoing management and administration in digital preservation, i.e. the need for life cycle management, a theme which is returned to later in

this paper. The full OAIS ‘blue book’⁸ presents in detail the recommendations. Figure 2 and the following Table 1 listing the key components of the OAIS, gives a brief introduction to how the system is designed.

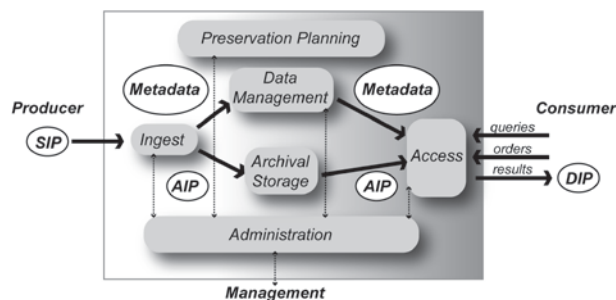


Figure 2. OAIS Functional Entities (after CCSDS 2002, 4-1).

⁶ International Organization for Standardization (ISO), ISO 14721:2003, Space data and information transfer systems - Open archival information system - Reference model, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683 (accessed May 6, 2009).

⁷ Preservation and Management Strategies for Exceptionally Large Data Formats: ‘Big Data’, Archaeology Data Service, <http://ads.ahds.ac.uk/project/bigdata/> (accessed 11 May 2009).

⁸ Reference Model for an Open Archival Information System, OAIS, (Washington DC: Consultative Committee for Space Data Systems, CCSDS 650.0-B-1 Blue Book, 2002).

| OAIS TERMINOLOGY | Description |
|--|---|
| Producer | The role played by those persons or client systems, which provide the information to be preserved. |
| Submission Information Package (SIP) | An Information Package that is delivered by the Producer (<i>in this case the archaeologist</i>) to the OAIS for use in the construction of one or more AIPs. |
| Ingest | The OAIS entity that contains the services and functions that accept Submission Information Packages from Producers, prepares Archival Information Packages for storage, and ensures that Archival Information Packages and their supporting Metadata (Descriptive Information) become established within the OAIS. |
| Archival Information Package (AIP) | An Information Package, consisting of the Content Information and the associated Metadata (Descriptive Information) which is preserved within an OAIS. |
| Archival Storage | The OAIS entity that contains the services and functions used for the storage and retrieval of Archival Information Packages. |
| Data Management | The OAIS entity that contains the services and functions for populating, maintaining, and accessing a wide variety of information. Some examples of this information are catalogues and inventories on what may be retrieved from Archival Storage, processing algorithms that may be run on retrieved data, Consumer access statistics, Consumer billing, Event Based Orders, security controls, and OAIS schedules, policies, and procedures. |
| Access | The OAIS entity that contains the services and functions which make the archival information holdings and related services visible to Consumers. |
| Dissemination Information Package (DIP) | The Information Package, derived from one or more AIPs, received by the Consumer in response to a request to the OAIS. |

Table 1. Selected OAIS terminology (taken from CCSDS-2002-1.7.2).

The OAIS standard identifies the following six mandatory responsibilities that an organization must discharge in order to be considered OAIS compliant:-

1. Negotiate for and accept appropriate information from information Producers.
2. Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation.
3. Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided.
4. Ensure that the information to be preserved is independently understandable to the designated community. In other words, the community should be able to understand the

information without needing the assistance of the experts who produced the information.

5. Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original.
6. Make the preserved information available to the Designated Community.⁹

⁹ *Reference Model for an Open Archival Information System, OAIS*, (Washington DC: Consultative Committee for Space Data Systems, CCSDS 650.0-B-1 Blue Book, 2002) Page 3-1.

4 ARCHIVAL DATA FORMATS

The data being considered by the share-IT project is initially limited to three data types, LiDAR, orthoimagery, and geophysical survey. These datasets not only have a geo-spatial graphical component (i.e. a map) but also have associated underlying data files, and potentially a cultural interpretation component.

A key data preservation issue is which file format is selected as the archival version, critical to the longevity and future access to the data. The archival information package is the version which will be held in perpetuity, and as such needs to be in a standard non-proprietary format. The choice of this format is critical as the submission format must be able to migrate into it, and the dissemination format be generated from it. Potential data formats include:-

LiDAR data:

LAS format – The LAS format is a public file format for the interchange of LIDAR data between vendors and customers. This binary file format is an alternative to proprietary systems or a generic ASCII file interchange system used by many companies.¹⁰

ASCII xyz – This is considered the standard format for text files. LiDAR data supplied by BKS to the Discovery Programme is in this format. Unlike the LAS format, ASCII can be easily understood by other software and opened easily to view and read by users.

Note: The preserved data must include the full data set before the creation of the DTM to ensure that improvements in processing algorithms over time can be applied to the data.

Orthoimagery:

GeoTIFF – These are files which have geographic (or cartographic) data embedded as tags within the TIFF file. The geographic data can then be used to position the image in the correct location and geometry on the screen of a geographic

information display. GeoTIFF is a metadata format, which provides geographic information to associate with the image data. But the TIFF file structure allows both the basic metadata and the image data to be encoded into the same file.¹¹ This is the currently used format at the Discovery Programme for orthoimagery created from PCI Geomatica 10 photogrammetric software. Although widely adopted by the community at large, this format is owned by Adobe Inc. and as such is deemed proprietary, adversely affecting its suitability as an archive format.

JPEG2000 format – the JPEG (Joint Photographic Experts Group) committee has addressed many of the limitations of the original JPEG format and its latest format, JPEG2000 has emerged as a new standard for the effective preservation of digital image data. The format is published as International Standard ISO/IEC 15444 Part 126. The particular advantages from an archiving perspective are:-¹²

- Metadata - the format embeds metadata within the file in a standard XML compliant environment. This allows for the possibility to incorporate descriptive information within the file.
- Lossy and lossless compression (with high quality lossless decompression available naturally through all types of progression)
- Progressive transmission by quality, resolution, component, or spatial locality
- Multiple resolution representation (images are decomposed into multiple resolutions in the compression process). This will dramatically increase the speed of display for large images, particularly important for high resolution data.
- No limit on file size, significant as image resolution increases.

¹⁰ Common Lidar Data Exchange Format - .LAS Industry Initiative, in American Society for Photogrammetry & Remote Sensing Online http://www.asprs.org/society/committees/lidar/lidar_for_mat.html (accessed May 6, 2009).

¹¹ GeoTIFF FAQ Version 2.3, <http://www.remotesensing.org/geotiff/faq.html#What%20is%20GeoTIFF%20and%20how%20is%20this%20different%20from%20TIFF?> (accessed May 6, 2009).

¹² Michael W. Marcellin, Michael J. Gormish, Ali Bilgin, Martin P. Boliek, "An Overview of JPEG-2000," *Proc. of IEEE Data Compression Conference* (2000), pp. 523-541.

(JPEG was the original JPEG committee standard for images (IS 10918-1) developed more than 15 years ago. It is generally not considered as an archive quality format primarily due to loss of quality on compression, and generation loss issues¹³).

The Open Geospatial Consortium (OGS) has adopted this format and defined the means by which the OpenGIS® Geography Markup Language (GML) can be used within the JPEG2000 format, GMLJP229. GML is an xml schema used to describe geographic information, including elements such as coordinate system, coverage, unit of measure, and also vector based objects (e.g. points, lines, and polygons). GMLJP2 is intended to handle a variety of imaging use cases including the following:

- Single geo-referenced images. GML describes the geometry and the radiometry.
- Multiple geo-referenced images of the same type. GML describes the geometry and the radiometry of the constituent images. Examples include a stereo photographic pair, a triangulation block of images, or image mosaics.
- Multiple geo-referenced images of various types. GML describes the geometry and the radiometry of the constituent images. Examples include combinations of images such as an optical image, FLIR and SAR images for target identification.
- Ortho-rectified images with or without associated digital elevation models.
- Digital Elevation Models that incorporate terrain-based constraints.¹⁴

With support at this level in the GIS community this format is rapidly being established as an industry standard for image archiving. This should be monitored with a view to its adoption as an archiving standard.

¹³ JPEG definition, from Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/JPEG#Recommended_usage (accessed May 6, 2009).

¹⁴ Ron Lake, David Burggraf, Martin Kyle, Sean Forde, *GML in JPEG 2000 for Geographic Imagery (GMLJP2), Implementation Specification*, (Open Geospatial Consortium Inc. 2005).

Geophysical Survey:

ASCII x,y,z – As with LiDAR (above) this is the standard for raw data files and is the common approach. The AIP should include the processed and unprocessed raw data, again to ensure that in the future improved algorithms can be applied.

5 METADATA STANDARDS

Metadata is often described as ‘data about data’.¹⁵ Its purpose is to provide context for data and to facilitate the understanding and management of a specific dataset. This is a similar function to that of a legend, north arrow and scale bar on a map. It provides the ‘who, what, where, why, when and how’ information which allow users to judge the quality or reliability of the data.

Metadata is an integral part of the OAIS model, a core component of the Ingest, Archival Storage and Access functions, see Figure 2. Metadata contains different levels of information which are all contained in the final schema. Three broad levels of metadata can be identified:-

Discovery - the minimum information to convey the nature and content of the resource.

Exploration - the information to ensure data is appropriate for purpose.

Exploitation - the information required to access, transfer, and apply the data in an application.

Dublin Core

The Dublin Core metadata element set is a standard for cross-domain information resource description. It provides a simple and standardized set of conventions for describing things online in ways that make them easier to find. Dublin Core is widely used to describe digital materials such as video, sound, image, text, and composite media like web pages. Implementations of Dublin Core typically make use of XML and are Resource Description Framework (RDF) based.¹⁶ The

¹⁵ Metadata definition, from Wikipedia, the free encyclopedia, <http://en.wikipedia.org/wiki/Metadata> (accessed May 6, 2009).

¹⁶ Dublin Core, definition, from Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Dublin_Core (accessed May 6, 2009).

Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description.¹⁷

ISO 15836:2003 defines the Dublin Core metadata element set which deals with cross-domain information resource description.

Qualified Dublin Core enables the extension of the core metadata element set to include additional schema such as controlled vocabularies. The ADS recommend the use of Dublin Core and have refined and defined how the elements should be created.¹⁸ In particular they define the schema for the subject element as the Thesaurus of Monument Types. However they note the flexibility of Dublin Core allows elements to be repeated, so to increase potential interoperability additional further subject element(s) could be added, possibly the Getty AAT controlled vocabulary to give an international dimension.

XML Schema

As noted, metadata is usually presented as an extensible markup language (XML) document. An XML schema is a description of a type of XML document with constraints on the structure and content beyond the basics imposed by XML itself. As the word extensible implies an XML schema has the flexibility to be extended or altered to suit the specific needs of particular user communities. Not surprisingly standard schema have been adopted for geospatial datasets and adopted by the International Standards Organisation (ISO).

ISO Standards

Many different metadata schemas exist specifically designed for digital objects. They can be general such as Dublin Core, or more specialised, but they are normally extensions to the Dublin Core schema. Our review of best practice revealed a strong emphasis on geospatial metadata standards and the adoption of particular ISO standards to achieve this. By adopting an ISO standard users

are able to know what to look for in the schema and are then better able to use the data, understanding its suitability and possible restriction.

The ISO 19100 is a series of standards for defining, describing, and managing geographic information.

Standardization of geographic information can best be served by a set of standards that integrates a detailed description of the concepts of geographic information with the concepts of information technology. A goal of this standardisation effort is to facilitate interoperability of geographic information systems, including interoperability in distributed computing environments. From this series one particular ISO metadata standard appeared to be almost universally recognized and adopted. ISO 19115 defines the schema for describing geographical information and associated services, including contents, spatial-temporal purchases, data quality, access and rights to use. The standard defines more than 400 metadata elements, 20 core elements. The ISO standards are revised and modified on a regular basis, ISO 19115:2003 is the current version.

INSPIRE Directive

Accepting the value of and necessity for ISO 19115 compliant metadata has become more significant following the implementation of the European Union INSPIRE directive.

The INSPIRE Directive sets out to improve the efficiency and effectiveness of public services – those associated with European environmental policy in the first instance – through the provision of a European spatial data infrastructure. INSPIRE is a directive which mandates member states to provide their public authority datasets and services so that they can more easily be used by other public organizations in the country concerned, in adjacent countries if required, and by the EC itself for policy making, reporting and monitoring. It is a set of principles and rules that each country must now choose how to implement - it will not necessarily need legislation.¹⁹

¹⁷ Dublin Core Metadata Element Set, Version 1.1, <http://dublincore.org/documents/dces/> (accessed May 11, 2009).

¹⁸ GIS *Guide to Good Practice*, Section 5: Documenting your GIS Data set, ADS <http://ads.ahds.ac.uk/project/goodguides/gis/sect54.html> (accessed May 11, 2009).

¹⁹ INSPIRE: you will be affected; you can help, The Association for Geographic Information (AGI), [http://www.agi.org.uk/SITE/UPLOAD/DOCUMENT/P olicy/INSPIRE_Vision.pdf](http://www.agi.org.uk/SITE/UPLOAD/DOCUMENT/Policy/INSPIRE_Vision.pdf) (accessed May 7, 2009).

| CATEGORY | ELEMENT | SHORT DESCRIPTION |
|-------------------------------|-----------------------------------|---|
| IDENTIFICATION | Resource title | characteristic and often unique name |
| | Resource abstract | brief summary of the content of the resource |
| | Resource type | type of resource being described |
| | Resource locator | link to additional information |
| | Unique resource identifier | value uniquely identifying resource |
| | Coupled resource | Identifies the target spatial data sets |
| | Resource language | the language(s) used within the resource |
| CLASSIFICATION | Topic category | high level to assist in grouping and topic based searching |
| | Spatial data service type | to assist in the search of spatial data services |
| KEYWORD | Keyword value | commonly used word to describe the subject |
| | Originating controlled vocabulary | the citation for the controlled vocabulary |
| GEOGRAPHIC LOCATION | Geographic bounding box | the extent of the resource in geographic space |
| TEMPORAL REFERENCE | Temporal extent | time period covered by resource |
| | Date of publication | publication or entry date – could be both |
| | Date of last revision | date resource last revised, if ever |
| | Date of creation | date of creation of the resource |
| QUALITY & VALIDITY | Lineage | statement on process history / quality of data set |
| | Spatial resolution | level of detail of the data set |
| CONFORMITY | Specification | citation of implementing rules to which data conforms |
| | Degree | degree of conformity of the resource |
| CONSTRAINTS | Conditions of access & use | free text description |
| | Limitations on public access | free text – if none then entered as text anyway |
| ORGANIZATIONS | Responsible party | organisation responsible for establishment, management etc |
| | Responsible party role | the role of the responsible organization |
| METADATA | Metadata point of contact | Organization responsible for creating/ maintaining metadata |
| | Metadata date | when the metadata record was created or updated |
| | Metadata language | language in which the metadata elements are expressed |

Table 2. The INSPIRE metadata elements, grouped by category.

The INSPIRE metadata schema is compliant with ISO 19115 / 19119 containing 27 elements grouped into 10 broad categories, see Table 2.

Cultural Heritage Inclusion in Metadata - Controlled Vocabularies / Thesauri

Thesauri, or controlled vocabularies can be added to the Keyword component of the metadata schema. Controlling how the cultural component is described using these resources enhances the ability of users to search and retrieve our data in intelligent ways. More than one thesaurus can be defined within a schema and our research identified a number which could be adopted:-

- Getty Thesaurus of Geographic Names Online (TGN)²⁰ - This identifies ‘place’ based on hierarchal relationships, with the superordinate ‘whole’ and its subordinate ‘members’ or ‘parts’.
- Getty Art & Architecture Thesaurus (AAT) - This is a controlled vocabulary used for describing items of art, architecture, and material culture. This thesaurus is compliant with two further ISO standards: - ISO 2788 & ISO 5964 – both provide guidelines for establishing and developing monolingual thesauri.
- Irish Cultural Heritage Content - The use of international thesauri provides a good

²⁰ Getty Thesaurus of Geographic Names Online, http://www.getty.edu/research/conducting_research/vocabularies/tgn/ (accessed May 7, 2009).

standardised approach but this need to be supplemented to take account of the Irish context of the datasets. For this some de facto standards do exist which could be adopted such as the DoEHLG monuments database which contains terms for describing archaeological monuments.

CIDOC Conceptual Reference Model CRM

CRM provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation, to promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to. It is intended to be a common language for domain experts and implementers to formulate requirements for information systems and to serve as a guide for good practice of conceptual modelling. In this way, it can provide the "semantic glue" needed to mediate between different sources of cultural heritage information, such as that published by museums, libraries and archives.²¹

CRM has been accepted as ISO standard 21127, guidelines for the exchange of information between cultural heritage institutions. In simple terms this can be defined as the curated knowledge of museums.²²

CIDOC CRM is an extremely complex model for cultural objects and maybe something which could be adopted at a later stage. Initially this would be too complex to incorporate into a proposed metadata schema, which should be kept relatively simple, if we want to ensure it is completed by users.

6 OAIS CASE STUDY

The discussion of the OAIS model, archive standards and metadata schema can seem somewhat complex and academic but presenting a simple example, in this case a small magnetometry survey, helps to clarify the issues.

The submission information package (SIP) is what the producer presents to our repository. In this case it includes:-

- The raw proprietary data files (GEOPLOT)
- The detailed report of the survey, in pdf format, created as a licence requirement
- And additional files including georeferencing information for the survey data and jpeg images

From this somewhat unstructured set of information an archival information package (AIP) needs to be generated. This is where data is migrated into formats which are appropriate for long term archival purposes. At present, best practice points to ASCII format for data files and TIFF format for images. An important component of the AIP is the generation of formal structured metadata – the xml file conforming to the designated schema, in this example ISO19115 and INSPIRE. This information is extracted from the pdf report submitted with the data files.

The dissemination information package can then be generated from the archive on request in the format required. Currently this could be GIS layers, or mapping service output which would be supplied with the associated metadata xml document.

7 COST MODELS

It should be clear that archiving data involves costs; from the data preparation and ingest stage, through to the long term costs of the digital archive lifecycle.

²¹ The CIDOC Conceptual Reference Model (CRM) home page, <http://cidoc.ics.forth.gr/> (accessed May 7, 2009).

²² International Organization for Standardization (ISO), ISO 21127:2006, Information and documentation http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=34424 (accessed May 7, 2009).

| RETENTION PERIOD | COST | COST (pence per MB) | CUMULATIVE (pence per MB) |
|------------------|------------------|---------------------|---------------------------|
| 5 years | R + E | 9 + 4 = 13 | 13 |
| 10 years | R – DR + E – DE | 9 - 3 + 4 – 1 = 9 | 22 |
| 15 years | R – DR + E – DE | 9 - 6 + 4 – 2 = 5 | 22 |
| 20 years | R – 3DR + E – DE | 9 – 9 + 4 – 3 = 4 | 27 |
| ongoing | | | 30 |

Table 3. Retention cost model where R = refreshment cost, DR = decreasing cost of refreshment, E = cost of physical equipment, DE = decreasing cost of equipment (adapted from ADS).

Two costing models from separate organizations were investigated:-

Archaeology Data Service (ADS) Cost Model

Archiving costs are calculated on the basis of 4 key elements:-

- *Management and Administration* – i.e. negotiating with depositor, processing the deposit, licences. This normally involves 2 - 3 days of effort.
- *Ingest* – migrating data to preferred formats, creation of metadata, and entry of data to system. Cost dependent on number and complexity of files.
- *Dissemination* – basic data delivery via simple file download is included in the price of data ingest, but special interfaces such as searchable databases or interactive maps may cost up to €15000 depending on functionality.
- *Storage* – (this includes the ongoing periodic process of data refreshments) Archives have to periodically upgrade systems - hardware and software - to take advantage of technological advances. (ADS have progressed through 3 generations of equipment during 10 years).²³

The ADS has developed formulae to estimate the cost of archiving data over variable time periods, which include the costs of refreshing data, costs of physical equipment, and factor in decreases in these costs over time, shown in Table 3.

As the table shows, the conclusion from the ADS project was that a cost of (applying figures from

the Big Data project) a one of charge of 30p per megabyte would cover ongoing preservation costs beyond 20 years. However, no account is taken of the number of files to be archived; e.g. 1 large file of 1GB size would involve significantly less effort than archiving 1000 smaller files of 1MB, although the total file size would be the same. Some adjustment to this model to balance volume and number of files would be an improvement. Applied to a small geophysical survey undertaken recently by the Discovery Programme which has generated 97MB of archiving data, the cost for preservation is around €30.

Life Cycle Information For E-Literature (LIFE)

The LIFE Project has developed a methodology to model the digital lifecycle and calculate the costs of preserving digital information for the next 5, 10 or 100 years.²⁴ There are 6 main lifecycle elements which are broken down further into lifecycle elements, similar to the OAIS functions, as shown in Table 4.

The LIFE model elements defined are not compulsory, but provide a framework within which to work that will be applicable to most situations. The accuracy of the output however is dependent on the sub layers and customisation added alongside the amount of real data that you have to put into the calculator. The more data you collect or have, the more accurate the model becomes.

²³ Archaeological Data Service: Charging Policy, 4th Edition
<http://ads.ahds.ac.uk/project/userinfo/charging.html>
 (accessed 7 May, 2009).

²⁴ Life Cycle Information for E-Literature (LIFE) homepage
<http://www.life.ac.uk/> (accessed 7 May 2009).

| LIFECYCLE CATEGORIES | LIFECYCLE ELEMENTS |
|-------------------------------------|---|
| Acquisition (Aq) (or pre-ingest) | Selection IPR Licensing Ordering and invoicing Obtaining Check-in |
| Ingest (I) | Quality assurance Deposit Holdings Update |
| Metadata (M) | Characterization Descriptive Administrative |
| Access (Ac) | Adding / maintaining links User support Access mechanism |
| Storage (S) | Bit-stream storage costs |
| Preservation (P) | Technology watch Preservation tool cost Preservation metadata Preservation action Quality assurance |

Table 4. Breakdown of the elements in the LIFE model.

From Figure 3 it can be seen that apart from the data acquisition costs all the other categories involve ongoing costs throughout the complete lifecycle. In terms of the share-IT project this is an important observation, which has to be understood in the context of identifying an appropriate hosting organization.

$$L_T = Aq + I_T + M_T + Ac_T + S_T + P_T$$

Figure 3. L is the complete lifecycle cost over time 0 to T. (after Lifecycle Information for E-literature).

This improved understanding of the cost of archiving, and the models to help calculate these costs suggest it may be appropriate for projects to include this as a component in future grant applications. This would see the digital archiving

of research assets become an integral part of overall project design and budget.

8 RIGHTS, ACCESS & SUBMISSION

One of the aims of the share-IT project is the dissemination of geo-spatial data, therefore our objective has to be to maximise use of the system. However we need to consider the intellectual rights and copyright implications of making data available via a webGIS system.

Copyright

A range of approaches to the issue of copyright were noted during the review of best practice. MIDA (the Marine Irish Digital Atlas) confronts this issue by way of a “*Memorandum of Understanding (MOU) for its data contributors*”.

The precise specification is adjusted to meet the needs of the supplier, creating a document in which the conditions that govern data supply, access and exploitation are fully laid out. Typical principles include:-

- The spatial dataset provided by a data owner may be displayed in the web-based GIS. This will be displayed as the data owner provides them, or generalised in a way that the data owner and the Coastal & Marine Resource Centre (CMRC) agree upon.
- Spatial dataset cannot be downloaded from the web-based GIS unless the owner has given prior consent.
- The contact details of the data owner will be provided in the metadata and therefore will be available over the Internet to atlas users who are interested in acquiring a copy of the spatial dataset.²⁵

The ADS requires users to accept both a “Copyright and Liability Statement” and a “Common Access Agreement” before accessing its *ArchSearch + Data* resource.

The OAIS model recognises the importance of copyright, “An archive will honour all applicable legal restrictions. These issues occur when the OAIS acts as a custodian. An OAIS should understand the copyright concepts and applicable laws prior to accepting copyright materials into the OAIS. It can establish guidelines for ingestion of information and rules for dissemination and duplication of the information when necessary.”

The large part of archaeological activity in Ireland is undertaken under licenses issued by the Archaeological Licensing Section of the National Monuments Service at the DoEHLG. The submission of the results in the form of a report is a condition of the license, and as such the results are in the public domain. Whether this system could be extended to include the data files which support the published reports is something which needs further consideration.

Currently a vast amount of archaeological work is being undertaken in advance of infrastructural projects, commissioned by state bodies such as the National Roads Authority. This data is being paid

for by the state and it would seem appropriate that it be made available once the planning process has been passed, and the project completed.

To comply with the schema, metadata must clearly state the conditions attached to access and copyright, and must deal with the issue of quality assurance.

User Community

The OAIS model identifies the “Designated Community” as the set of consumers who should be able to understand the preserved information. It also emphasizes that this community will evolve or change over time.

Archives can allow different access to information or data depending on the user status. It may be that general open access is only given to basic levels of data and simple viewing tools, with different access and functions such as downloading facilities available to those registered or even paying subscription. This was noted with the SAFER-Data web-based interface of the EPA (U.S. Environmental Protection Agency). They identified three categories of user, controlled by a registration and login system:-

- *Public Users* - those interested in finding out about environmental research, exploring data, and possibly downloading data and/or reports to their own computers for further studies.
- *Researchers* - both researchers looking for data and information about other research projects, and also researchers uploading their environmental data and information for archival on the Secure Archive For Environmental Research Data System.
- *EPA Users* - interested in exploring information about environmental research currently being carried out and results of research projects which have concluded.

Promotion of Digital Archiving

The success of a webGIS such as that proposed by the share-IT project is dependent on high volumes of data being submitted. This is a reason in itself for the share-IT project to actively promote the value of digital archiving.

The ADS express the view “that there is little point in preserving data unless it is reused” and actively promote the dissemination of data through its web interface. Options range from pages with

²⁵ The Marine Irish Digital Atlas (MIDA): Data Supply, Access and Exploitation Principles <http://mida.ucc.ie/pages/dataPrinciples.htm> (accessed 7 May 2009).

downloadable files to interactive maps and searchable online interfaces.

9 CONCLUSIONS

This paper has concentrated on the major issues that arose out of the best practice review of archiving digital data. The research quickly showed that our ambition to share and open access to data could only be achieved by ensuring data is archived to the highest standards. These standards are well defined in the narrow realm of our selected data types, geospatial data, through international organizations, in particular, ISO 19115 and INSPIRE.

To progress from the ‘proof of concept’ phase to a fully functioning system would require securing funding to ensure an OAIS compliant archive be put in place, with long term financial support. This is now being actively pursued both domestically in Ireland, and with potential EU partners. The full ShareIT project reported in December 2008, with a report submitted to the Heritage Council (Ireland).

Expanding the scope to less well defined archaeological themes such as excavation data and reports would present new challenges in defining appropriate standards. However, with adherence to the OAIS principles the archiving and sharing of diverse heritage and archaeological digital data assets would be a realistic goal.

Acknowledgements

The authors would like to acknowledge the Heritage Council (Ireland) for the funding of this project through the INSTAR grants program of 2008. Also we would like to thank the members of staff at ADS, University of York, in particular Prof. Julian Richards and Dr. Stuart Jeffrey, for their considerable advice and encouragement.

Bibliography

Archaeological Data Service: Charging Policy, 4th Edition <http://ads.ahds.ac.uk/project/userinfo/charging.html> (accessed 7 May, 2009).

A Review of Research Needs in Irish Archaeology, (Kilkenny: The Heritage Council, 2007).

Reference Model for an Open Archival Information System, OAIS, (Washington DC: Consultative Committee for Space Data Systems, CCSDS 650.0-B-1 Blue Book, 2002).

Common Lidar Data Exchange Format - .LAS Industry Initiative, in American Society for Photogrammetry & Remote Sensing Online http://www.asprs.org/society/committees/lidar/lidar_format.html (accessed May 6, 2009).

Cooney, Gabriel. *Archaeology in Ireland: A Vision for the Future*, ed. (Dublin: Royal Irish Academy, 2006).

Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), in Official Journal of the European Union, http://www.ec-gis.org/inspire/directive/l_10820070425en00010014.pdf (accessed May 6, 2009).

Dublin Core Metadata Element Set, Version 1.1, <http://dublincore.org/documents/dces/> (accessed May 11, 2009).

Getty Thesaurus of Geographic Names Online, http://www.getty.edu/research/conducting_research/vocabularies/tgn/ (accessed May 7, 2009).

GIS *Guide to Good Practice*, Section 5: Documenting your GIS Data set, ADS <http://ads.ahds.ac.uk/project/goodguides/gis/sect54.html> (accessed May 11, 2009).

Harvey, Alison. *The Heritage Council Strategic Plan 2007 – 2013 Consultation Document*, (Kilkenny: The Heritage Council, 2006).

INSPIRE: you will be affected; you can help, The Association for Geographic Information (AGI),
http://www.agi.org.uk/SITE/UPLOAD/DOCUMENT/Policy/INSPIRE_Vision.pdf (accessed May 7, 2009).

Lake, Ron, David Burggraf, Martin Kyle, Sean Forde, *GML in JPEG 2000 for Geographic Imagery (GMLJP2), Implementation Specification*, (Open Geospatial Consortium Inc. 2005).

Life Cycle Information for E-Literature (LIFE) homepage
<http://www.life.ac.uk/> (accessed 7 May 2009).

Marcellin, Michael W., Michael J. Gormish, Ali Bilgin, Martin P. Boliek, "An Overview of JPEG-2000," *Proc. of IEEE Data Compression Conference* (2000), pp. 523-541.

Preservation and Management Strategies for Exceptionally Large Data Formats: 'Big Data', Archaeology Data Service,
<http://ads.ahds.ac.uk/project/bigdata/> (accessed 11 May 2009).

Reeners, Roberta, ed., *Archaeology 2020. Repositioning Irish Archaeology in the Knowledge Society*, (Dublin: University College Dublin, 2006).

The CIDOC Conceptual Reference Model (CRM) home page,
<http://cidoc.ics.forth.gr/> (accessed May 7, 2009).

The Marine Irish Digital Atlas (MIDA): Data Supply, Access and Exploitation Principles
<http://mida.ucc.ie/pages/dataPrinciples.htm> (accessed 7 May 2009).