Irish Archaeological Data: Towards a framework

Anthony Corns
The Discovery Programme
63 Merrion Square
Dublin 2
Anthony@discoveryprogramme.ie

Louise Kennedy
The Discovery Programme
63 Merrion Square
Dublin 2
louise@discoveryprogramme.ie

ABSTRACT
The EU co-funded project ARIADNE (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe) aims to integrate data and services across Europe for the archaeological research community. As part of this project, The Discovery Programme investigated data holdings and data management practices in Irish archaeology. The research sought to document what information is available in physical and digital forms, what approaches the key aggregators of archaeological data took to the preservation and accessibility of their data and the challenges to and opportunities for dataset networking with Irish archaeological data. In addition to one of the largest State commissioners of development led archaeology, representatives of the State organisations which accession the results of licensed archaeological activity (objects, data, and information) were interviewed. Awareness of approaches to digital preservation varied: metadata capture was focused on discovery and/or digital asset management; preservation of data was focused in the business continuity and risk mitigation sphere rather than on preservation of data for research re-use in the long term; choice of file formats was as or more likely to be driven by business need rather than sustainability. The availability of both capital and human resources affected the nature and extent of work carried out with data, whether the task was digitisation or the realisation of further value in digitisation effort carried out previously. While significant work will be required from the sector to enable the accessibility expected by data consumers now and in the future, there is also an opportunity to approach the challenges of digital preservation as a community, by building digital preservation requirements into the workflows of future digitisation and data creation projects.

Keywords
Archaeology, dataset networking, research infrastructure, Linked Data.

1. INTRODUCTION
The ARIADNE project is focused on delivering open and integrated services to the research community and beyond. The primary means of delivering access to datasets is through the ARIADNE portal which is the discovery layer for the project registry [1]. The architecture assumes that archaeological research data generated by projects or groups is stored in institutional repositories, data centres, portals or other national or international aggregators, from where it is integrated within ARIADNE. National data centres represented in the project include the Archaeology Data Service (ADS) [2] in the UK, Data Archiving and Networked Services (DANS) [3] in the Netherlands and the Swedish National Data Service [4]. Subject repositories include the Arachne object database at the German Archaeological Institute [5] and Fasti Online [6], a database of the International Association of classical Archaeology (AIAC) and the Centre for the Study of Ancient Italy (CSAI) excavations since 2000. The Discovery Programme, as the Irish ARIADNE partner, sought to understand archaeological data in the absence of infrastructures like those mentioned above in order to ensure the representation of Irish archaeological data in Europe and globally.

2. SURVEY
An interview based survey aimed to identify the nature and range of Irish archaeological datasets, the practices used to manage them and the potential for and challenges to integrating datasets. The interviews also captured the experiences and concerns of key stakeholders in Irish archaeology around the subjects of digital data management and data sharing.

Semi-structured interviews were conducted with participants from each of the three State organisations with responsibility for archaeology. Background information, theme guides and consent forms were circulated to each participant in advance of the interview. The interviews investigated the nature and range of datasets created or curated, data archiving practice, use of standards, policies and practices for access and re-use and priorities for growth.

3. DATA PRACTICES
3.1 Nature and range of data
Datasets held represent the State's record of licensed archaeological investigation and protection of the physical archaeological record within the landscape. The organisations hold hundreds of thousands of archaeological records, crucial to research and cultural heritage because of the adopted strategy of preservation by record, of which ten to twenty five per cent are held as digital records. Where access to records is only through hard copy, there were concerns about the preservation of the physical record in light of the level of demand and use.

3.2 Data archiving
All of the organisations interviewed regularly backed up their data, usually driven by IT requirements. Cloud storage is in use, and data was also backed up to tape in one case. Participants valued cataloguing work and understood the need to include archiving tasks as a component of project work, but expressed concerns about resources. Therefore while data is being maintained, the kind of approach to descriptive information demanded by an OAIS model [7] of digital preservation is not being implemented or considered, and mechanisms to identify and indicate bit rot [8] are not in place.

The concepts most relied on in Irish archaeological metadata for discovery were excavation or activity licence numbers, spatial concepts such as townland or site names, National Monument numbers (an alpha-numeric unique identifier) and the names of excavators or companies undertaking licensed activity. Excavation licence numbers were used in file naming conventions in two organisations.

3.3 Data quality and standards

Interview participants raised concerns about data quality, particularly the impact of transformations of digitised data, and their consequences for user perception of the data. With the exception of one flat classification scheme implemented in a web service, controlled vocabularies were not in use for data or metadata. One organisation was structuring and releasing data in accordance with the INSPIRE directive [9]. Lack of thesauri will have a significant impact on the ability to integrate datasets and provide Linked Open data for example, for new analyses of (and within) the archaeology sector.

The most common formats for data were word processor files, spreadsheets and PDF documents, some of which were PDF images of text documents and therefore not searchable. Jpeg and TIFF were most common for image files, with RAW also utilised by the photographic survey sections. The lack of standardization in metadata capture at the point of data creation and through the lifecycle presents a barrier to the re-use, integration and preservation of data, one which will require specific and sustainable resources to rectify if the archaeological community is to participate in the Semantic Web. In the long term, influencing the culture of data creation and capture towards proactive recordkeeping [10] can ensure capture of the necessary metadata to fulfil purposes of discovery, re-use, integration and preservation efficiently as part of common workflows.

3.4 Data access and re-use

For a number of participants, Irish archaeological data is considered to be open: it has been created through publicly funded research and/or through the licensing system, it forms part of the public record and can be accessed for research purposes as physical records, or in certain cases, through an intranet. For some records, consulting the hard copy is the only means of access. All of the participants saw improved digital access to data as a long term aim, and resources were a significant barrier to realizing that aim. Lack of resources resulted in data not being digitised, access tools were not launched, growth of online resources were curtailed or publication online was limited. Potential misinterpretations by remote end-users of data created in analogue media and presented in digital media was also a concern.

All of the organisations were open to public re-use of metadata. Attitudes to data re-use varied according to data type: rights in images were more carefully considered, protected and respected than rights in other data, due to the potential revenue generation from the commercial re-use of images.

3.5 Growth: priorities and barriers

The publication of primary and/or secondary data is a priority in the sector, with the use of online media a key element of the process.

Reductions in numbers or staff and/or recruitment restrictions were cited by all participants as significant barriers to growth and planning. Frequent changes in organisation management structure were also a barrier to medium-long term planning, leading to uncertainty about the sustainability of currently active projects.

Participants had concerns around depositing data with external repositories, around the sustainability of providing accurate and timely updates to the data and managing rights and responsibilities of re-use where data creation has been complex.

4. CULTURAL DATA FRAMEWORK

ARIADNE activities in Ireland has led to the formation of a formal group of cultural heritage stakeholders addressing the issues highlighted in the survey together, through a Cultural Data Framework. The Framework has provided a forum for communication of shared needs and goals around archaeology and cultural heritage data and insights from the development of archaeological data projects such as the Dublin City and County Archaeological GIS [11].

The National Roads Authority's (NRA) grey literature excavation reports are the test case for ingesting archaeological records into the Digital Repository of Ireland (DRI). This collection of over two thousand reports detail the archaeological findings of road scheme developments all over Ireland. Metadata is manually captured from the content within the PDF reports received by the NRA from excavators and complies with the DRI guidelines [12] for qualified Dublin Core. Spatial coverage metadata includes townland URIs from Logainm [13]. The NRA specified a term list for temporal terms which, though not implemented uniformly by authors, has been helpful for capturing temporal site classification in the metadata. Following ingestion into the DRI digital object identifiers (DOI) will be minted for each excavation report, providing a persistent reference for citation.

In addition, work has begun on a thesaurus of Irish archaeological objects, based on previous work on terms and concepts. The thesaurus will utilise the expertise available within ARIADNE in the areas of thesaurus construction and SKOS [14]. The thesaurus will be subject to expert community review and will expand with the support of the Royal Irish Academy Committee for Archaeology.

5. DATA AND THE ARCHAEOLOGICAL RECORD IN IRELAND

The system of licensed archaeology results in the fragmentation of the archaeological record. Licensed excavators or surveyors are required to submit and publish final reports of the activity undertaken. The supporting data which these reports synthesize is not required to be submitted to the National Monuments Service or the National Museum under the current conditions of the licence. Given the large volume of development led archaeological activity between 2000 and 2008 [15] and the subsequent closure of many commercial archaeology companies, there is a significant risk of loss of primary data in the sector, which in the context of preservation by record, raises ethical issues for policy makers and questions around data management culture and practice for the archaeology sector. While efforts have been made to capture records of licensed activity from closed or closing companies, much of what has been captured exists only in physical form, with basic metadata and without the possibility of digitisation within current resources, representing a further barrier to the development of archaeological research in Ireland.

The archaeology and wider cultural heritage sector in Ireland has a crucial task ahead; to build data management practices which, from the conception of projects, influence workflows to include practices which can support OAIS models of digital preservation for appropriate data. The sector requires appropriate support and resources to achieve this, in addition to sustainable and trusted infrastructure to provide preservation and access.

6. ACKNOWLEDGMENTS

Many thanks to the interview participants from the National Monuments Service of the Department of Arts, Heritage and the Gaeltacht, the National Roads Authority and the National Museum of Ireland. ARIADNE is funded by the European Commission under the Community's Seventh Framework Programme.

7. REFERENCES

[1] ARIADNE, 2014. *The Way Forward to Digital Archaeology in Europe*. http://www.ariadne-infrastructure.eu/content/download/4569/26666/version/2/file/Ariadne+Booklet.pdf.

[2] Hardman, C. and Green, K. 2013. Preserving Our Digital Heritage: The Work of the Archaeology Data Service at the University of York. *Forum Yorkshire* 2, 115-120. http://www.cba-yorkshire.org.uk/forum

[3] Guéret, C. and Cudré-Mauroux, P. 2014. The Entity Registry System: Publishing and Consuming Linked Data in Poorly Connected Environments. In: *ERCIM-News*, 97 (Apr. 2014), 54-55

[4] Carlhed, C. and Alfredsson, I. 2009. Swedish National Data Service's Strategy for Sharing and Mediating Data: Practices of Open Access to and Reuse of Research Data *IASSIST Quarterly 32* (Spring/Winter 2008), 30-39. http://snd.gu.se/sites/snd.gu.se/files/Open_Access_to_and_Reuse_of_Research_Data_SWE.pdf

[5] Arachne object database http://arachne.uni-koeln.de/drupal/?q=en

[6] Fasti Online http://www.fastionline.org/index.php

[7] The OCLC/RLG Working Group on Preservation Metadata. 2002. *A Metadata Framework to Support the Preservation of Digital Objects*. http://www.oclc.org/research/pmwg/

[8] De Stefano, P., Fleischauer, C., Goethals, A., Kjörling, M., Krabbenhoeft, N., Lacinak, C., Mandelbaum, J., McCarthy, K., Murray, K., Navale, V., Rice, D., Ruggaber, R., Owens, T., Zwaard, K. 2014. *What is Fixity, and When Should I be Checking it?* National Digital Stewardship Alliance (2014) : http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA-Fixity-Guidance-Report-final100214.pdf

[9] Directive 2007/2/EC of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community Directive *Official Journal of the European Union* L108/1(Mar. 2007) http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32007L0002

[10] Upward, F, 1996. Structuring the records continuum (Series of two parts) Part 1: post custodial principles and properties. *Archives and Manuscripts* 24, 2 (Nov. 1996), 268-285. http://www.infotech.monash.edu.au/research/groups/rcrg/publications/recordscontinuum-fupp1.html

[11] Duggan, C and Johnson, R. 2013. Dublin City Archaeology GIS Project: Phase 2. *Newsletter of the Institute of Archaeologists of Ireland* (Winter 2013) 5-6. http://iai.ie/wp-content/uploads/2014/09/Winter-2013-18-12-13-high-res.pdf

[12] Bustillo, M., Collins, S., Gallagher, D., Grant, R., Harrower, N., Kenny, S., Martínez García, A., Ní Cholla, R., O'Carroll, A., Redmond, S. and Webb, S. 2015. *Qualified Dublin Core and the Digital Repository of Ireland* (R. Grant, Ed.). Maynooth: Maynooth University; Dublin: Trinity College Dublin; Dublin: Royal Irish Academy; Galway: National University of Ireland, Galway. DOI= http://dx.doi.org/10.3318/DRI.2015.3

[13] Placenames Database of Ireland http://www.logainm.ie/en/

[14] Vlachidis A, Tudhope D. 2012. A pilot investigation of information extraction in the semantic annotation of archaeological reports. *International Journal of Metadata, Semantics and Ontologies*, 7,3 (2012) 222-235. DOI= 10.1504/IJMSO.2012.050183

[15] Cooney, G. 2006. Key Recommendations from the Royal Irish Academy Forum Archaeology in Ireland: A Vision for the Future (Sep. 2006). https://www.ria.ie/getmedia/98ddf707-06b6-47ec-9040-d59ae607c2c5/Archaeology-Forum-Recommendations-(Sept-2006).pdf