Impact case study (REF3b)

Institution: University of South Wales

Unit of Assessment: B11

Title of case study: Linking Archaeological Data - enabling semantic infrastructure in the digital archaeology domain

1. Summary of the impact Our research has enabled archaeological professional and commercial organisations to integrate diverse archaeology excavation datasets and significantly develop working practices. Commercial archaeological datasets are usually created on a per-site basis structured via differing schema and vocabularies. These isolated information silos hinder meaningful cross search and comparison. As the only record of unrepeatable fieldwork, it is essential that these data are made available for re-use and re-interpretation. As a result of the research, the Archaeology Data Service, English Heritage, the Royal Commissions on the Ancient and Historical Monuments of Scotland and Wales have published as Linked Data important excavation datasets and national vocabularies that can act as hubs in the web of archaeological data.

2. Underpinning research *Numerical references with prefix I, P refer to elements of Impact and Publications in later sections of the case study.* Over the last 20 years, our research has investigated semantic interoperability in cultural heritage and in particular archaeological datasets. Commercial archaeological data depositors form a large group with varied working practices. Datasets are often created on a per-site basis structured according to differing schema and employing different vocabularies. Cross searching, comparing or reusing the data in any meaningful way remains difficult. This hinders the reassessment of original archaeological findings and interpretations in the light of new data or broader research questions. The work on semantic metadata originated in a PhD project using a small dataset from a local history museum, where the research prototype had connections between items based on measures of semantic closeness rather than explicitly authored links (P1). EPSRC funded research extended this approach to the Science Museum's collections database and the widely used Getty Art and Architecture Thesaurus. The Demonstrator's API facilitated use of the thesaurus in dynamic user interface elements (P2). The generalisation of this approach beyond a single organisation's collection to multiple data schemas and thesauri required an integrating conceptual framework and the CIDOC CRM (ISO 21127) was selected as a standard in the cultural heritage domain. Diverse data structures and schemas are integrated when datasets are mapped to the CRM. However the terminology problem remains - different words can mean the same thing; the same word can mean different things. Our work on terminology services and knowledge organization systems was an early contributor of tools and applications for the W3C SKOS standard for machine readable vocabularies (P4, I5). This exploration of the complementary use of formal ontologies and information science vocabularies (eg thesauri) for semantic interoperability has underpinned subsequent research (P3, P6) and our contributions to the new ISO 254964 thesaurus standard (I6). Work on a major application of the approach to the archaeological domain began in 2000 with the AHRC funded STAR project (P3). Data was mapped to an English Heritage archaeological extension of the CRM, which was expressed in the standard semantic representation RDF (I3). The resulting STAR Demonstrator provides cross search at a conceptual level over diverse archaeological datasets from different organisations and associated grey literature excavation reports (P5). The second (STELLAR) phase of the research extended this approach beyond the development team and to the production of Linked Data. It produced tools and guidelines to streamline the process of mapping/extracting data to the (CRM) ontology and reduce the need for specialist knowledge of the particular ontology or semantic techniques in general; mapping and extraction can be performed by archaeological data curators or providers, rather than semantic web developers (P5, I1, I4).

STELLAR tools are capable of delivering thesaurus identifiers (URIs) - however, there was a lack of key thesauri published as Linked Data. Building on the work with terminology services and SKOS/RDF, the SENESCHAL AHRC funded knowledge exchange collaboration with major UK heritage vocabulary providers has resulted in the publication of national archaeological thesauri as Linked Data (I2). This allows concepts for different kinds of monuments and time periods to be directly referenced via identifiers and thus act as the much needed vocabulary hubs in the web of archaeology data.

Research Team
Professor Douglas Tudhope 1993 -
RA Ceri Binding 2000 - 2005; 2007 -
RS/RA (gained PhD) Andreas Vlachidis 2007 –

3. References to the research
* references selected as best indicating the quality of the underpinning research

P1*. Tudhope D., Taylor C. 1997. Navigation via Similarity: automatic linking based on semantic closeness. Information Processing and Management, 33(2), 233-242. Elsevier Science. doi:10.1016/S0306-4573(96)00067-2

P2. Binding C., Tudhope D. 2004. KOS at your Service: Programmatic access to knowledge organisation systems. Journal of Digital Information, 4 (4), Open access, http://journals.tdl.org/jodi/article/view/110/109 *(RAE2008 output)*

P3*. Binding C. 2010. Implementing archaeological time periods using CIDOC CRM and SKOS. Proceedings 7th Extended Semantic Web Conference, Heraklion, L. Aroyo et al. (Eds.): ESWC 2010, Part I, Lecture Notes in Computer Science, 6088, 273–287, Springer-Verlag
Conference had 21% acceptance full papers.

P4. Binding C., Tudhope D. 2010. Terminology web services. Knowledge Organization, 37(4), 287-298. *(REF output)*

P5*. Tudhope D, May K, Binding C, Vlachidis A. 2011. Connecting archaeological data and grey literature via semantic cross search. Internet Archaeology, 30, Open access, http://intarch.ac.uk/journal/issue30/tudhope_index.html doi:10.11141/ia.30.5
(REF output) Early work published in ECDL 2008 (Springer LNCS, 28% acceptance full papers). The hypertext format of Internet Archeology (published by Council for British Archaeology and dating to 1996) permitted an extended treatment (>14,000 words) and the linking of use scenario discussions to a live demonstrator of archaeological cross search.

P6. Souza R, Tudhope D, Almeida M. 2012. Towards a taxonomy of KOS: dimensions for classifying knowledge organization systems, Knowledge Organization, 39(3), 179-192.

**Selection of Relevant Research Funding,
with grant values as awarded to University of South Wales (Glamorgan)**

EPSRC. Thesaurus-based access to multimedia collections: faceted retrieval tools (FACET), Tudhope PI, 3 years, 2000-2003, £121k.

AHRC. Semantic Tools for Archaeological Resources (STAR). Tudhope PI, 3 years, 2007–2010, £222k

AHRC. Semantic Technologies Enhancing Links and Linked data for Archaeological Resources (STELLAR), with co-investigators ADS. Tudhope PI, 1 year, 2010–2011, £73k.

AHRC. Semantic ENrichment Enabling Sustainability of arCHAeological Links (SENESCHAL). Tudhope PI, 1 year, 2013-2014, £50k.

IMSL/JISC/AHRC/ESRC Digging into Data Challenge (transatlantic). Digging into Metadata, Tudhope Co-I, 2 years, 2012-2014, £75k.

FP7 Infrastructures Grant. Advanced Research Infrastructure for Archaeological Dataset Networking in Europe (ARIADNE), Tudhope WP Leader, 4 years, 2013 – 2017, €205k.

5 JISC Projects between 2006 and 2010 in collaboration with UKOLN (University of Bath) and MIMAS (University of Manchester), Tudhope PI or Co-I, over £155k to USW/Glamorgan. Terminology Services review, semantic interoperability demonstrator, (semantic) tag suggestion service, etc.

4. Details of the impact
There are two primary and four secondary impact items.

I1. *2011.* The research on semantic data integration (STELLAR) provided tools and techniques that enabled the Archaeology Data Service (ADS http://archaeologydataservice.ac.uk/) to extract and publish Linked Data from major commercial archaeology units' excavation datasets, integrated semantically via mapping to the CIDOC CRM ontology. It is envisaged this will serve as a catalyst for further production of archaeological Linked Data by ADS and others. Building on this work, we are leading the FP7 ARIADNE archaeology e-infrastructure Work Package, Linking Archaeology Data.

The research enabled ADS (non-specialists in semantic technologies) first foray into Linked Data and represents a major development in practice and capability by ADS and in UK archaeological data publication. It has generated considerable attention - from June 2012 over roughly 12 months, 41,110 requests were made to the SPARQL endpoint to the Linked Data which averages approximately 3425 requests per month. Lee (2012) positively refers to non-specialist STELLAR tools in an English Heritage Practitioner article. The *significance* also derives from the importance of the published datasets and the exemplar. The Linked Data includes datasets drawn the Channel Tunnel Rail Link and the Aggregates Levy Sustainability Fund, major archaeological programmes with excavations undertaken by two of the largest commercial units in England (Oxford Archaeology Ltd and Wessex Archaeology Ltd). Other datasets included an excavation database with details of the earliest ironworking yet known in Britain. As the only record of unrepeatable fieldwork, it is essential that these data are preserved and made available for re-use and re-interpretation.

Commercial archaeology units benefit from wider exposure of their data and the ability to cross search across different datasets from different units and for reuse of data (with potential economic benefit). The *reach* is amplified by the key strategic role played by the ADS nationally and internationally. The ADS is a national repository for digital data from the UK historic environment sector, crosscutting the academic and public and private sectors. It provides online access to over one million metadata records on behalf of national government agencies, local government Historic Environment Records, and amenity and period societies and other specialist databases. The ADS user community includes national and local government archaeologists and cultural heritage managers, museums and commercial archaeologists and members of the public.

I2. July *2013.* We (and the vocabulary partners in the SENESCHAL project) published as (SKOS) Linked Data the nationally recognised cultural heritage thesauri standards from English Heritage, the Royal Commission on the Ancient and Historical Monuments of Scotland and the Royal Commission on the Ancient and Historical Monuments of Wales. This includes concepts widely used for indexing relating to monument types, archaeological events and time periods.

The *significance* is that previously the vocabulary providers lacked the ability to facilitate uniquely identified semantic indexing of data. Major thesauri can act as vocabulary hubs for the Web of

Data (as suggested by W3C Library Linked Data Incubator Group). For example, the availability of the Thesaurus of Monument Terms in this way is seen as a major development for the ADS *archive metadata* Linked Data (ADS Blog). This Linked Data publication of the English Heritage thesauri is a significant development in their vocabulary standards practice and their information access strategy.

The potential *reach* is wide since it is a core activity of ADS, English Heritage, The Royal Commissions on the Ancient and Historical Monuments of Scotland/Wales to promote and disseminate best practice to the heritage sectors, as well as providing guidance on appropriate data standards including thesauri. The linked data vocabularies and web services will be integrated into the widely used ADS reporting/archiving tool, OASIS, which is in near universal use by commercial and local government archaeologists. Adoption of linked data based vocabulary management in this tool will immediately affect how all sectors engage in archaeological field practice and development control planning.

I3. *2010.* We represented the English Heritage archaeological extension to the CRM ontology in RDF and as Linked Data. This allowed it to be a key ontology hub in the ADS archaeology Linked Data [I1]. This is another important step in English Heritage's strategic plans for information standards.

I4. *2012.* The Deutsches Archäologisches Institut have used STELLAR research tools to make a SKOS version of the DAI Archaeological Thesaurus (German language).

I5. *2011.* We supplied the UK Data Archive (Essex) with their first SKOS version of HASSET, the pre-eminent social science thesaurus. This was very helpful for their development of HASSET.

I6. *March 2013.* ISO 25964-2 published - Tudhope was a member of the ISO Working Group. Building on previous work for the British Thesaurus Standard BS 8723, the research focus on complementary use of thesauri and ontologies has contributed to the chapter on thesaurus – ontology interoperability for the new thesaurus standard, ISO 25964 (part 2). The ISO thesaurus standard has been widely influential and the potential audience includes all commercial information system designers and developers making of thesauri or related vocabularies.

5. Sources to corroborate the impact

I1. Archaeology Data Service Linked Data http://data.archaeologydataservice.ac.uk/
 I1 and I2 *Corroborator:* ADS Lead Applications Developer.
 English Heritage **Practitioner article**. Everything We Know Informs Everything We Do: A Vision for Historic Environment Sector Knowledge and Information Management. http://dx.doi.org/10.1179/1756750512Z.0000000006
I2. SENESCHAL **Archaeological Vocabulary Linked Data** http://www.heritagedata.org/
 ADS blog http://archaeologydataservice.ac.uk/blog/2013/07/seneschal-value-to-the-ads/
 I2 and I3 *Corroborator*: English Heritage, Head of Heritage Data Management
I3. CRM-EH archaeological extension of CIDOC CRM ontology
 http://hypermedia.research.glam.ac.uk/resources/crm/ (follow link to RDF document).
I4. DAI blog http://c4tc.wordpress.com/2012/10/08/skosifying-an-archaeological-thesaurus/
I5. *Corroborator:* UK Data Archive, Management Information Manager
I6. ISO 25964. Thesauri and interoperability with other vocabularies: Part 2: Interoperability with other vocabularies (Ch21 Ontologies). http://www.niso.org/schemas/iso25964/