Template Based Semantic Integration: From Legacy Archaeological Datasets to Linked Data

Ceri Binding, University of South Wales, Pontypridd, UK
Michael Charno, Archaeology Data Service, York, UK
Stuart Jeffrey, Glasgow School of Art, Glasgow, UK
Keith May, Historic England, Portsmouth, UK
Douglas Tudhope, University of South Wales, Pontypridd, UK

ABSTRACT

The online dissemination of datasets is becoming common practice within the archaeology domain. Since the legacy database schemas involved are often created on a per-site basis, cross searching or reusing this data remains difficult. Employing an integrating ontology, such as the CIDOC CRM, is one step towards resolving these issues. However, this has tended to require computing specialists with detailed knowledge of the ontologies involved. Results are presented from a collaborative project between computer scientists and archaeologists that created lightweight tools to make it easier for non-specialists to publish Linked Data. Archaeologists used the STELLAR project tools to publish major excavation datasets as Linked Data, conforming to the CIDOC CRM ontology. The template-based Extract Transform Load method is described. Reflections on the experience of using the template-based tools are discussed, together with practical issues including the need for terminology alignment and licensing considerations.

Keywords: CIDOC CRM, Data Integration, Digital Archaeology, Linked Data, Ontology, Semantic Interoperability

1. INTRODUCTION

Linked Data can be seen as a step towards the Semantic Web vision of creating a globally accessible web of data. In this context there has been much interest in exposing cultural heritage data online to encourage interoperability and reuse (Bizer, Heath & Berners-Lee, 2009; Linked Data). In practice, this has tended to require specialists in semantic technologies and detailed

DOI: 10.4018/IJSWIS.2015010101
knowledge of the ontologies involved. This paper presents results from a collaborative project between computer scientists and archaeologists, where a key aim was to make it easier for archaeologists new to semantic technologies to create and publish Linked Data.

Archaeology has seen an increasing use of the Web in recent years for dissemination of datasets describing the results of archaeological interventions. Archaeology datasets are disseminated in a platform neutral format as delimited text files, enabling import and manipulation by a wide range of tools. Most of the excavation fieldwork datasets in the UK are produced by commercial archaeology units. However there are many hundreds of these archaeological contractors who vary in their working practices. Datasets are often created on a per-site basis structured according to differing schema and employing different vocabularies, and as a consequence cross search, comparison or other reuse of the data in any meaningful way remains difficult. This hinders the reassessment of the original archaeological findings and reinterpretation in the light of evolving research questions.

The use of an integrating framework, such as the CIDOC Conceptual Reference Model (CIDOC CRM; Doerr 2003), is seen as one step towards resolving these issues. However in practice this activity requires an understanding of the source dataset schema, together with specialist knowledge of the target ontological model and the techniques required for expressing mappings. In many organisations a single person does not possess all of the required skills; as a result the overall process can be resource intensive and error prone. There is a need for tools and approaches to assist the creation of Linked Data by people other than experts in semantic technologies. This general point is also emphasised by Shakya et al. (2009), although their approach makes use of social platforms to create very informal ontologies, which in turn drive community based Linked Data. Addressing similar general goals by different methods, the work presented here investigates the use of lightweight techniques and tools to map and extract archaeological data conforming to a formal ontology to be published as Linked Data.

1.1. Background

This paper draws on work by the authors on use of semantic technologies in the archaeology domain over the period 2007 to 2012 and which is still continuing. The paper largely draws on two research projects (STAR followed by STELLAR') mainly the latter phase.

The collaborators for the research are the Archaeology Data Service (ADS) hosted by the Department of Archaeology at the University of York, and English Heritage (EH). The ADS undertakes archival and preservation of a wide range of digital data from work funded by various UK research councils and other organizations. It acts as a bridge between commercial archaeological contractors and specialists and the academic and public research communities. In addition to ‘grey literature’ (unpublished fieldwork reports), ADS also make available fieldwork datasets underpinning the findings described in the grey literature. EH advises the UK government and local authorities on the management of nationally important parts of England’s cultural heritage and provides research resources, including new methodologies for information management.

The ADS hold over 400 archival collections of archaeological data representing thousands of archaeological interventions and excavations in the last two decades. Two major archived research programmes were selected for the research discussed in this paper, the Channel Tunnel Rail Link (Foreman, 2004) representing over 100 excavations along the line of the rail link from Kent to Central London, and the Aggregates Levy Sustainability Fund (ALSF) which funds excavations relating to the aggregates extraction industry in the UK. Both these programmes offered a broad range of datasets containing excavation databases with a variation in structure and are typical of archaeological archives, particularly excavation databases, held by the ADS.
Challenges on Semantic Web Services
www.igi-global.com/chapter/challenges-semantic-web-services/35721?camid=4v1a