

# International Journal on Semantic Web and Information Systems

January-March 2015, Vol. 11, No. 1

## Table of Contents

### RESEARCH ARTICLES

- 1 **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*
  
- 30 **PatchR: A Framework for Linked Data Change Requests**  
*Magnus Knuth, Hasso Plattner Institute for Software Systems Engineering, University of Potsdam, Potsdam, Germany*  
*Harald Sack, Hasso Plattner Institute for Software Systems Engineering, University of Potsdam, Potsdam, Germany*
  
- 46 **Complex Role Inclusions with Role Chains on the Right are Expressible in SROIQ**  
*Michael Compton, CSIRO, Cygnet, Australia*

### Copyright

The **International Journal on Semantic Web and Information Systems (IJSWIS)** (ISSN 1552-6283; eISSN 1552-6291), Copyright © 2015 IGI Global. All rights, including translation into other languages reserved by the publisher. No part of this journal may be reproduced or used in any form or by any means without written permission from the publisher, except for noncommercial, educational use including classroom teaching purposes. Product or company names used in this journal are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI Global of the trademark or registered trademark. The views expressed in this journal are those of the authors but not necessarily of IGI Global.

The *International Journal on Semantic Web and Information Systems* is indexed or listed in the following: ACM Digital Library; Bacon's Media Directory; Burrelle's Media Directory; Cabell's Directories; Compendex (Elsevier Engineering Index); CSA Illumina; Current Contents®/Engineering, Computing, & Technology; DBLP; DEST Register of Refereed Journals; Gale Directory of Publications & Broadcast Media; GetCited; Google Scholar; INSPEC; Journal Citation Reports/Science Edition; JournalTOCs; Library & Information Science Abstracts (LISA); MediaFinder; Norwegian Social Science Data Services (NSD); Science Citation Index Expanded (SciSearch®); SCOPUS; The Index of Information Systems Journals; The Standard Periodical Directory; Thomson Reuters; Ulrich's Periodicals Directory; Web of Science

# 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<sup>1</sup>) 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.

Datasets made available by ADS typically consist of a collection of delimited text files - each file representing a database table, each delimited row representing a fielded database record. Some datasets are accompanied by limited schema documentation – usually taking the form of a diagram or a table description. The data files may contain a header row of column names but this convention is not consistently practised for all datasets.

Given that there is no common schema in use in the archaeological sector and there is extensive variability in the terminology, normal usage of these datasets requires analysis to take place on a site by site basis. Cross-search is extremely limited. Site metadata may allow search at broad location or major time period level. However it is almost impossible to search across datasets directly for, say, examples of a particular type of artefact from a particular period occurring in a particular type of context (e.g. *Roman pottery* found in *early medieval middens*). Datasets are increasingly available online but effectively isolated from each other and also with no connection to grey literature (unpublished excavation reports), for example from the ADS digital library. These isolated resources do not support research inquiries that depend on semantic interoperability between differing database structures and terminology, even on such fundamental questions as finding all hearths (Richards and Hardman 2008).

In order to connect these disparate resources, the CIDOC CRM ontology was used as an integrating conceptual framework. Data was mapped to an archaeological extension of the CRM (CRM-EH) developed by EH and converted to Resource Description Framework (RDF) format. Natural Language Processing techniques were developed to extract key information from grey literature and represent it in the same CRM based RDF format. The first phase of the research resulted in the STAR Web Demonstrator, which allows cross search at a conceptual level over five archaeological datasets and archaeological grey literature reports. Tudhope et al. (2011) provide an overview of the Demonstrator for an archaeological audience, discussing some of the archaeological modelling issues and various detailed search scenarios.

The initial requirements gathering involved discussions with immediate collaborators and two workshops with archaeologists, particularly those involved with major UK excavation database systems. This resulted in the underlying use case of cross search over different archaeological datasets (and grey literature) at a meaningful level of detail. This could potentially address key themes for archaeological research inquiry. It would also offer the possibility of resource discovery of a dataset for subsequent detailed local investigation.

It was decided to focus on the archaeological notions of contexts (an identified unit of excavation defined by a physical space), groups (interpretive higher level groupings of contexts), finds and samples. These archaeological concepts are discussed further in section 3.2.4, together with associated attributes and properties. Thus the scope of the work focused on mapping excavation data records to the corresponding ontology classes, together with their associated properties and attributes. This meant that elements of the datasets corresponding to, for example, administrative issues or detailed procedural aspects of recording practices were not included.

The decision to extract a (major) selection from the datasets was a judgement following discussion with the archaeological collaborators that the elements selected afforded answers to common research questions at the inter-site level. An example of the general cross search use case might involve an archaeology researcher or postgraduate student who encounters a result in an excavation report concerning the finding of a metal tool within a hearth context, and wishes to discover whether any other grey literature report or any published excavation dataset has a similar or closely related finding. Similarly the field archaeologist may wish to discover whether an unexpected dating of an object found within a particular context in their excavation has been replicated in any datasets from other parts of the country, or in any grey literature report of a dataset not yet available.

A set of search statements typical of the use case might include:

- Search for Contexts of type “*post-hole*”;
- Search for Contexts of type “*hearth*” containing a Find;
- Search for Contexts of type “*hearth*” containing a Find of type “*coin*”;
- Search for Finds of type “*coin*” located in Contexts of type “*hearth*”;
- Search for Contexts of type “*corn mill*” where a Sample was taken.

This paper goes on to discuss the generalisation of the data extraction methods employed in the first phase of the research, where data mapping and extraction was assisted by the use of a project specific purpose built data extraction tool (Binding & Tudhope, 2008) and the mapping and extraction was performed by the development team. The second (STELLAR) phase extended this approach beyond the development team and to the production of Linked Data. It provided tools and guidelines to streamline the process and reduce the potential for errors and inconsistency, thus allowing the mapping and extraction work to be performed by archaeological data curators or providers, rather than Semantic Web developers. Some database proficiency is required but specialist knowledge of ontologies or the CIDOC CRM is not a prerequisite.

The next section reviews related work, including methods for mapping between relational and RDF models, ontology based mapping/extraction generally and with specific reference to the CIDOC CRM. Section 3 describes the methods employed, including the underlying template-based approach and the resulting tools. A case study of the ADS experience in applying the STELLAR tools to extract and publish archaeological linked data is discussed in Section 4, while Section 5 discusses use of the published data, including SPARQL querying. Section 6 presents conclusions and future work.

## 2. RELATED WORK

In the conversion from relational model to event-based ontology, mappings between relational fields and ontological properties may not be a straightforward 1:1 relationship (Barrasa, Corcho & Gómez-Pérez, 2004; Binding & Tudhope, 2008; Kondylakis, Doerr & Plexousakis, 2006). A data element may map to several interrelated ontology elements and conversely a single ontology element may require a composition of data elements. When mapping to an event based model such as CIDOC CRM information concerning the events may be only implicit in the original dataset and needs to be made explicit, usually resulting in a chain of entities and properties. There can also be conditional mappings which are dependent on particular data values. These issues are discussed further with an accompanying practical example in section 1.1.1. There are two principal approaches to the creation of mappings between relational and RDF models – automatic generation or domain ontology mapping.

### 2.1. Automatic Generation of Mappings

Generally this means mirroring a relational schema as RDF. Tables become classes, records become nodes, columns become predicates, and cells become values (Berners-Lee, 1998). Examples include Relational.OWL (Pérez de Laborda & Conrad, 2005), DB2OWL (Cullot, Ghawi & Yétongnon, 2007) and RDBToOnto (Cerbah, 2008) which all facilitate automatic generation of RDF based on a relational database schema. The advantage is that coverage is complete – all data becomes RDF, although Byrne (2008) highlights the potential for generation of a large

number of unnecessary additional triples using this method. There are minimal gains in terms of semantic interoperability and further work would still be required to relate the generated RDF data to a domain specific ontological model.

## 2.2. Domain Ontology Mappings

This approach involves the creation of custom mappings between the relational database schema and a pre-defined domain ontology. It is seen as having advantages over the automatic approach described in section 0, as it offers the possibility to model the domain semantics more explicitly. Domain ontology mapping may also be more selective, not necessarily encompassing all the data held in the original database (conversely the ontology may exceed the scope of the original database). Mappings may be described in terms of a formal declarative language, processing tools utilising these mappings in data retrieval or conversion.

### 2.2.1. Languages for Expressing Domain Ontology Mappings

A number of languages have been developed to formally express domain ontology mappings. For example, D2RQ is an intermediary language expressed in RDF used for on-demand mapping between a database schema and RDF (D2R Server). The R2O language (Barrasa, Corcho & Gómez-Pérez, 2004) is expressed in XML and is processed by the ODEMapster plug-in (Barrasa & Gómez-Pérez, 2006; Priyatna et al., 2011); R2O was intended to extend the mapping description capabilities of D2RQ to tackle a perceived lack of expressiveness. Virtuoso uses a declarative meta schema language to map a relational database schema to an RDF schema or OWL ontology, an approach referred to as Linked Data Views (Erling & Mikhailov, 2009).

Kondylakis, Doerr & Plexousakis (2006) proposed a mapping language represented as an XML DTD for formally declaring mappings between an XML data structure and CIDOC CRM. They highlighted many higher level issues inherent in the mapping process; particularly that additional data may be required to complete the ontological representation.

Hert, Reif & Gall (2011) provided a useful comparison of RDB-to-RDF mapping languages. In the light of a potential proliferation of languages and approaches the W3C RDB2RDF Incubator Group was set up to assess the various existing approaches to mapping relational data into RDF. Tools and techniques for creating RDF representations of relational data were summarised in a survey report (Sahoo et al., 2009), and the group's findings led to the establishment of the W3C RDB2RDF Working Group, which aimed to converge on a standardized language for mapping relational data and relational database schema to RDF and OWL. The resultant mapping language (R2RML) became a W3C Recommendation in September 2012.

The STELLAR approach effectively performs domain ontology mapping, although it will be seen that it employs predefined 'templates' rather than a formal mapping language, and as a result hides much of the inherent complexity from the end user. Within domain ontology mapping there are two main approaches to making data available, on-demand mapping, and extract-transform-load.

### 2.2.2. On-Demand Mapping

On-demand mapping translates queries 'on the fly' to SQL, using mapping languages as described in Section 2.2.1. For example, D2R Server provides an on-demand RDF interface to existing relational databases by converting RDF queries to SQL, using D2RQ as the mapping language. The NeOn toolkit included the ODEMapster tool (Priyatna et al., 2011), an RDB to RDF engine



using the R2O mapping language. ODEMapster was subsequently renamed to morph-RDB, and upgraded to support the R2RML mapping language.

The on-demand mapping approach can be advantageous in cases where the underlying data is frequently changing as the translated data remains up to date, rather than being a snapshot. This of course requires the existence of (and access to) a relational database. When repurposing an existing database the indexes present may not be appropriate for the task and parallel on-demand mapping queries can adversely affect the performance of an operational database system. Optimisation can be problematic in cases where the original database schema itself is not well normalised - as can be the case in archaeology datasets, which are often produced on a per-site basis and evolve as the project proceeds.

### **2.2.3. Extract Transform Load (ETL)**

A term originating from data warehousing, this defines the process of extracting and integrating data from multiple sources. ETL is appropriate for heterogeneous data sources requiring more complex application logic to transform the mapped data and for cases where reasoning may be subsequently performed. ETL may be achieved using mapping languages. Alternatively, a more lightweight approach to the translation from relational to ontological models involves the creation of bespoke queries and scripts or templates to achieve the desired transformation of the data, for which various tools exist. Open Refine (formerly Google Refine) is a general application for importing, cleaning and transforming data; RDF Refine is then an extension containing a graph template for exporting RDF format output (Open Refine). TRIPLIFY (Auer et al. 2009) is an application for selective ETL conversion of relational data and for subsequent Linked Data publication. Rather than formally expressing mappings in a declarative language the application uses a configuration file (a PHP script) to perform a series of user-defined SQL queries against a target database. The queries are written to output tabular data with URIs as column names, which can then be converted directly to RDF triples. The configuration requires the user to have some knowledge of PHP, namespaces and RDF predicates.

The STELLAR applications exhibit some commonality with aspects of the TRIPLIFY approach, performing selective ETL extraction of relational data via user-defined SQL queries, thus avoiding the need for a declarative mapping language. The key difference is that for STELLAR the output of the query becomes the input to one or more templates rather than being directly transformed to RDF. It is assumed that the user has knowledge of their domain and experience of querying their own datasets - but perhaps less detailed knowledge of RDF syntax, namespaces, integrating ontologies or mapping languages. Thus the user is offered a set of pre-defined templates specifying the input columns required to achieve a particular output. The process of mapping and data extraction then becomes a matter of the user creating a suitable query on their own dataset to produce tabular output having specific column names recognised by the chosen template. The template approach has the advantage that the output does not have to be XML, RDF or OWL; rather it can be *any* textual format and multiple templates may be applied to the same underlying data to produce different output.

## **2.3. CIDOC CRM Mapping**

Nußbaumer & Haslhofer (2007) discuss problematic issues that emerged during the course of the BRICKS FP6 Integrated Project when defining mappings from two archaeological databases to the CIDOC CRM, with the mappings performed by different domain experts. Alternative valid mapping chains were sometimes created for the same underlying semantics in the different databases. Additionally, similar but different concepts were assigned the same mapping chain,

due to the high level nature of the CRM and the absence of implementation guidelines. A lack of guidance for the mapping process allowed different personnel involved in mapping to focus on different aspects of the database schema and the ontology, or possibly different levels of generality.

Nußbaumer, Haslhofer and Klas (2010) argue that the CRM has need both of standards for technical representation and sets of mapping guidelines for particular domains. They discuss various examples, including whether:

- A material (e.g. *gold*) should be associated with a production activity or an object;
- A method of manufacture (e.g. *hammered*) should be considered a type of production or a procedure appellation;
- Man-made objects should directly be given an identifier or this connection made via the identifier of a document describing the objects.

Since the CRM is fundamentally event-based there is also the issue of when it is appropriate to assert an assignment event when assigning an attribute to an object. Within the frame of the CRM this decision rests upon whether the attribute assignment is a judgement appropriate to record, e.g. where provenance metadata might be appropriate, or the judgement might be revisited subsequently in light of new knowledge.

The CRM standard documentation does not specify implementation details. This is in part due to the longevity of the CRM. The collaborative effort that resulted in the CIDOC CRM began in 1996 to consider an object oriented approach to museum information interchange, '*in order to benefit from its expressive power and extensibility for dealing with the necessary diversity and complexity of data structures in the domain*' (CIDOC CRM SIG). The first complete publication of a CIDOC CRM version was in 1999 - predating the Semantic Web. Some older applications of the CRM have employed it as an intellectual resource rather than as a formal implementation.

In theory, different practice in mapping and different granularity of modelling detail could be accommodated by software systems capable of traversing the resulting CRM graph network, and in the future this may be the case. In operational practice today this can be hard to achieve and different mappings can thwart the goal of semantic interoperability. This proved problematic for the BRICKS project discussed above, which required the addition of an intermediate mapping indexing which itself served as the integrating layer rather than the CRM. Nußbaumer, Haslhofer and Klas (2010) propose a general mapping methodology that combines the choice of the most specific CRM class or property with domain specific guidelines, thus leaving significant responsibility to the implementation team or a domain mapping standards effort. Since various mappings are potentially possible from the same data elements, this issue is not trivial and is not confined to the CRM. Any general core ontology will tend to permit different mappings from the same set of data elements, depending on the purpose of the mapping exercise and the precise aspects of the data that it is desired to capture. However, the potential divergence of mapping practice poses challenges for implementations and the final applications relying on such data.

These considerations influenced the approach followed in this paper; there is a need to communicate the purpose of any common mapping exercise and to make available to data providers a choice of what might loosely be termed *mapping patterns* for their domain. In the archaeology domain, an example might be data relevant to the identification, typology and dating of particular 'Finds' objects (STELLARb). Working from established RDF expressions improves the prospects for interoperability and offers a practical implementation route for non-specialists. This approach is particularly appropriate for the commonly encountered case in archaeology of legacy delimited data files, where the original database may no longer exist or be easily accessible.



## 2.4. Pattern Based Approaches

Pattern based approaches have emerged within Linked Data (Dodds & Davis, 2012) and have been applied for some time in computer science generally with differing degrees of formality. Gangemi (2005) introduced the notion of Ontology Design Patterns that can be employed throughout the ontology lifecycle, encompassing requirements gathering, modelling, and consuming applications. Conceptual patterns are able to generalise specific ontology design elements from different domains and platforms. The approach is being taken forward by a community based around the Ontology Design Patterns portal (ODP) with the aim of formalizing and communicating recurring ontological modelling components to support reusability on the design side. Useful patterns encoding best practices are categorised and communicated using OWL, UML diagrams and textual descriptions.

The work described in this paper has an affinity with this approach (the term *mapping patterns* intended to indicate an informal pattern based approach). However, rather than a concern with ontology design (building a new ontology or applying one to solve a domain modelling problem), there is a specific use case of mapping and extracting archaeological datasets to conform to an *existing* core ontology, the CIDOC CRM. The templates and associated applications identify and document particular domain specific ontological modelling patterns and encode a practical implementation mechanism to enable transformation of data *instances* to conform to the identified patterns.

Templates were created to enable the bulk transformation of tabular relational data, taking care of necessary lower level RDF syntax matters and ensuring consistency and repeatability of the output. The STELLAR templates do not correspond to any existing ODP, which tend to document ontology design elements rather than instance data corresponding to a given ontology pattern. Thus templates allow a correspondence to be made between elements of source datasets and selected ontology entities, allowing a STELLAR application to transform the source data to instance data conforming to the ontology. Given that the use case is pre-defined, the documentation of the existing templates contains many of the ODP elements required for ontology design patterns, demonstrating the use of each template field via graph diagrams, descriptions, example data and corresponding RDF output (STELLARb, STELLARc). If in future a large number of templates become available then some form of template database with metadata assigned to templates for discovery purposes might be considered and for this an approach similar to ODP could then be helpful.

The modelling is based on an established ISO standard ontological model (CIDOC CRM), which is focused at the fairly abstract level of “*integration of heterogeneous cultural heritage data*”, while allowing for extension for more specific modelling in particular domains. The English Heritage CRM extension covering the archaeological excavation process (CRM-EH) was considered to represent documented best practice for the domain and facilitate more specific use cases. We did not aim to design or create new models; instead we would identify and document discrete subsets of an established larger model as patterns. The CRM-EH extension was achieved via sub-classing, thus preserving interoperability with the CIDOC CRM at more abstract levels, while allowing specific archaeological mappings (and queries). Encapsulating suitable mapping patterns within reusable templates removes the onus from individual data providers to make potentially ambiguous modelling decisions. Choosing a template corresponds to making a mapping to the CRM and CRM-EH entities associated with the template. The user provides the input required for the chosen template, choosing which of the optional columns to supply. The approach was favoured by the general use case of inter site cross search with a

focus on key archaeological concepts. It is also possible to define new mappings, although that requires more technical expertise.

Contributors to a discussion on ‘Museums and the Machine Processable Web’ (Linking Museums III) voiced similar concerns, expressing the need to move towards agreed patterns and templates for the encoding of historical information to be exposed as Linked Data. Also following an informal pattern based approach, Kurtz et al. (2009) aggregated data on Classical Art objects from multiple data providers via the CRM (CLAROS). The mapping work was delegated to the data providers by specifying CRM-based patterns for Objects, Places, Periods & People, as part of the input workflow. As part of the work discussed in this paper, a specific CLAROS template was developed (see section 3.2.1) following interest from that project.

### 3. TEMPLATE-BASED MAPPING AND EXTRACTION

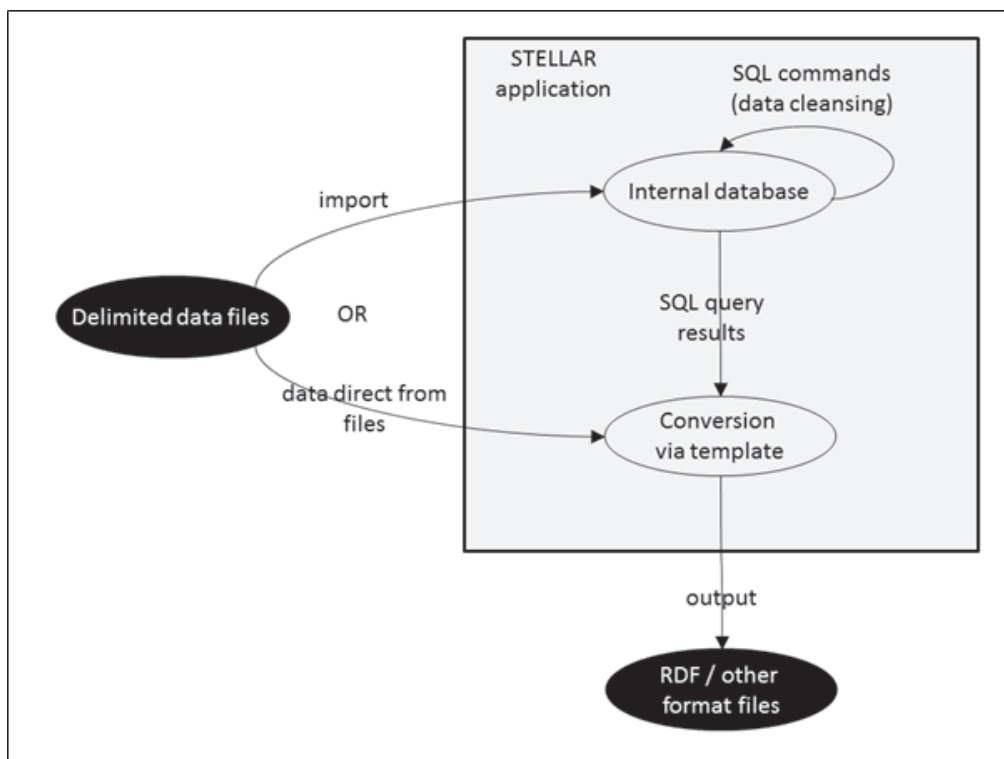
The aim is the development of applications and templates that might offer a consistent and repeatable process to convert a volume of data from delimited files via templates to RDF. Users must identify an appropriate template for each pattern to be expressed as CRM compliant RDF. Queries on the datasets or a suitable delimited data file can produce input for the templates. The applications and templates can then transform a set of archaeological datasets to a CRM (or CRM-EH) compatible RDF format. If existing templates do not cover the desired mapping pattern then a user-defined template for the pattern may be developed. Several templates corresponding to the cross search use case described in Section 1.1 are available online with accompanying tutorials (STELLARa). The general approach is described in an introduction document (STELLARb) and the current templates are detailed in a manual (STELLARc).

#### 3.1. Applications

The STELLAR applications perform a variety of data manipulation and conversion tasks (STELLARa—see Tools). Figure 1 illustrates the data conversion functionality; files of delimited tabular data can be imported and consolidated to an internal database; the data may then be manipulated, cleansed, modified, and enriched using a succession of SQL UPDATE commands. The tabular output of a SQL SELECT query on this database is then fed through a template (either pre-defined or user-defined, see section 3.2), processing each input row in sequence to convert to a chosen textual output format. A tabular delimited data file may also be used as input directly, for cases where the data is already in a suitable state to be converted without requiring data cleansing or other pre-processing.

*STELLAR.Console* is a command line utility application supporting batch processing of a series of commands so the whole import-manipulation-conversion process for an entire dataset may be consistently repeated and also stored for future reference.

*STELLAR.Web* is an online browser-based application using the same pre-defined internal templates. It provides less flexible functionality than the command line utility but has a simpler interface. The conversion process works directly on data files uploaded to the server and the resulting output data can be retrieved via a generated link to a downloadable file. Statistics for the RDF output are displayed, such as a count of unique resource URIs and literal values generated, a count of instances for each entity and the number of statements using each distinct property. These statistics can help in assessing whether the RDF data being created matches what was expected from the conversion process.

Figure 1. *STELLAR* data conversion functionality

*STELLAR.Win* is a Windows application processing delimited text files via user-defined templates. This application performs a subset of the functionality of *STELLAR.Console* - providing a simpler user interface.

### 3.2. Templates

Templates express RDF patterns with placeholders for instance data, addressing the necessary ‘boilerplate’ syntax issues required for repeatable and consistent ETL data conversion – such as namespaces, character encoding, cardinality of entities, properties and attributes, URI structure, conformance to a specific predefined ontological model (frequently there is no simple 1:1 correspondence between data fields and ontology entities) etc. The STELLAR applications can perform bulk data conversion operations using these templates, repeatedly creating the same pattern for each row in the dataset with the data values from named fields inserted into embedded placeholders. Each record in the tabular input data is thus converted via the template to conform to a chosen output format (in this case conforming to the CRM-EH ontological model for archaeological data, expressed in RDF/XML). A concise database element may result in a chain of CRM relationships and so templates can generate intermediate ‘virtual’ entities to model events, which are core to the CIDOC CRM but often implicit within the data itself, together with the automatic generation of inverse properties.

### 3.2.1. Pre-Defined Templates

A number of internal pre-defined templates were initially created (implemented as compiled XSL transformations). The rationale for pre-defined templates is that the resultant output remains consistent for all users and is always valid RDF data conforming to a specific ontological model. As the immediate objective of the project was the conversion of legacy archaeological datasets to RDF conforming to the CRM-EH model, the majority of the pre-defined templates deal with this - so there are templates covering archaeological projects, groups, contexts, finds, samples and measurements. However there are also other templates available, including a template used for creating SKOS thesauri from tabular data, and an experimental template for the CLAROS project used to convert museum collections data to CIDOC CRM.

### 3.2.2. User-Defined Templates

Responding to feedback from a workshop with archaeologists, the STELLAR.Console application was upgraded to facilitate the creation of *user-defined* templates. An alternative to XSL was explored based on an external template engine called StringTemplate (Parr). User-defined templates take the form of simple text files having embedded placeholders for data values. Modifying a user-defined template does not require rebuilding the STELLAR application, as was the case for the built in pre-defined (compiled XSLT) templates. This option added more flexibility to the existing template approach, and enhanced the scope of the application, since it enabled the transformation of data to *any* chosen textual output format. Although the primary objective of the research was conversion of archaeology datasets to RDF and the initial user-defined templates produced RDF/XML format output, user-defined templates are format agnostic and other templates may be defined to produce JSON-LD, HTML, text reports, etc. from the same tabular input data. A number of example user-defined templates were bundled with the STELLAR.Console application installation, and were also made available via the project website together with tutorials (STELLARa).

The rationale for investigating the StringTemplate engine as an alternative to XSL transformation was to offer a simpler way for end users to express their custom templates. In hindsight, creating completely new templates to express RDF/XML data conforming to a large ontological model such as CIDOC CRM is inherently complex, although tailoring of an existing template may be easier. Thus the anticipated advantages of the alternative to XSL were not necessarily realised and both implementation methods appear equally valid. The source code for the STELLAR applications is available as open source. For specific details on the two template approaches, see the respective tutorials (STELLARa).

### 3.2.3. User-Defined Template Syntax and Structure

The required syntax for user-defined templates for use with the StringTemplate engine is as defined in (Parr). Delimited template placeholders (e.g. \$data.id\$, \$data.note\$) are replaced at run time by values from correspondingly named input data fields and the resultant text is written to the output. During execution the STELLAR application will look for 3 key named templates (all optional) contained within the specified StringTemplateGroup (\*.stg) file:

- **HEADER (*options*):** If this template is present it is used once at the start of processing - *options* will contain any options passed in as configuration values (e.g. the base namespace to use);

- **RECORD (*options, data*):** If this template is present it is used once for every record in the input tabular dataset. *data* is passed in by the STELLAR application, representing a single row having a series of fields with names derived from the original column names – any template placeholders present having these names are then replaced by the applicable data values at runtime;
- **FOOTER (*options*):** If this template is present it is used once at the end of processing.

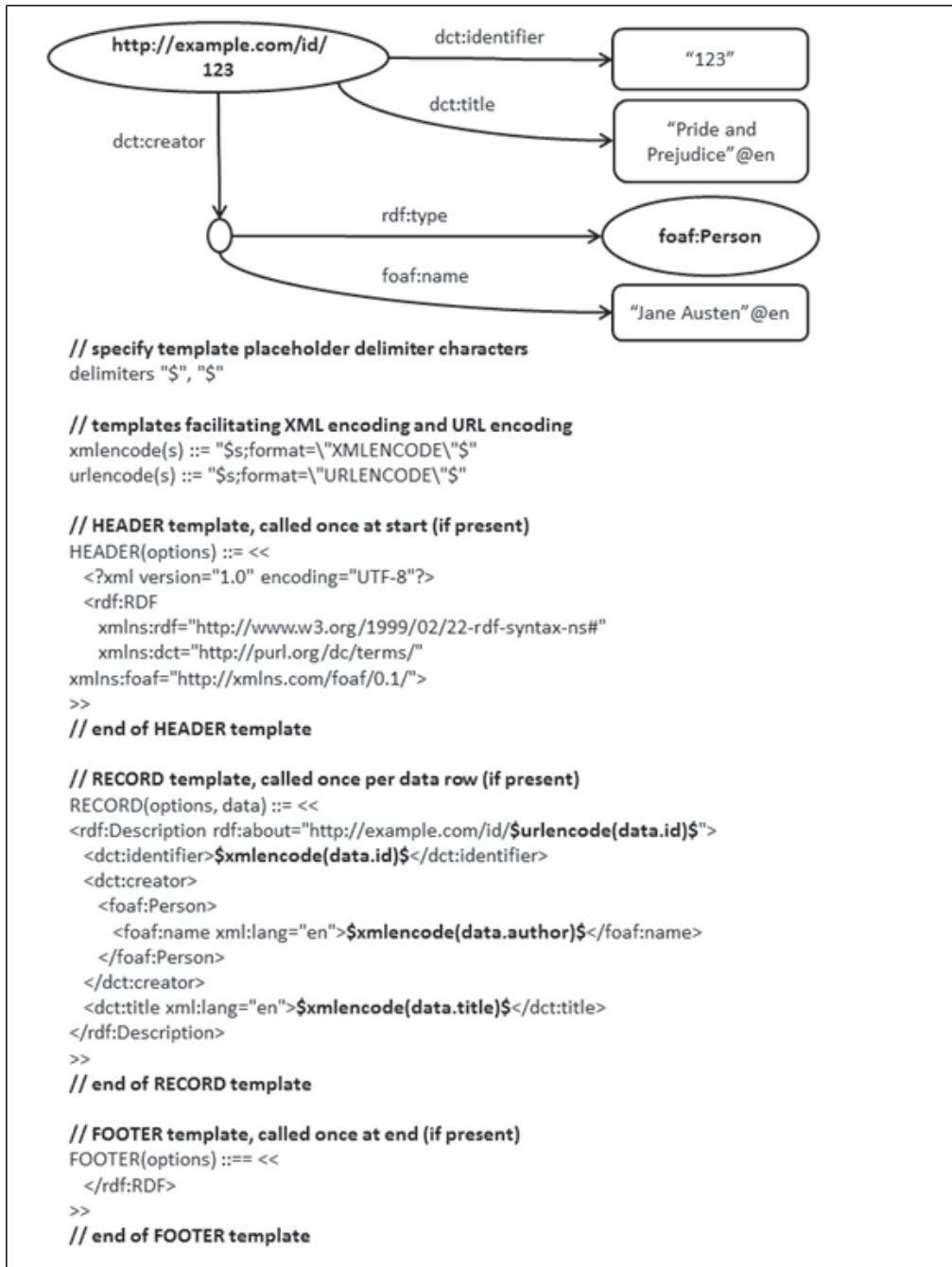
Figure 2 is a simple self-contained example showing the syntax of a StringTemplateGroup containing each of these templates. This illustrative example will create RDF/XML formatted bibliographic data conforming to the model fragment as illustrated. Figure 3 then shows some example tabular input and the resultant RDF/XML output. The templates perform XML-encoding and URL-encoding of data values to prevent the potential generation of invalid output, especially important where input data values are being used in the formation of a URI. This is undertaken within the template itself rather than by the STELLAR application because the required output of a user-defined template may be a format other than XML. This simple example template group could then be repeatedly reused for any tabular data having named columns *id*, *author* and *title* - to produce consistently formatted bibliographic Dublin Core metadata expressed as valid RDF/XML. Ontological modelling patterns can of course be far more complex than a simple list of resources with properties and so a more detailed example is described in Section 1.1.1 - converting data to conform to a subset of the CRM-EH ontological model.

### 3.2.4. CRM-EH Template Usage Example: Archaeological Stratigraphy

A key data unit encountered in archaeology datasets is the *context*, a unit of excavation that represents a discrete physical element of the site. Relationships between contexts may be either spatial (physical relationships encompassing adjacency, intersection and containment) or temporal (stratigraphic relationships). Context numbers (identifiers) are fundamental to the recording systems used by archaeologists to record each identified individual unit of stratigraphy. The archaeological stratigraphic relationship is defined by the archaeological excavator. It is recorded in the Single Context Recording methodology by use of the Harris Matrix (Harris 1989). Crucially, the task of the excavator is to record the precise relative temporal sequence in which the deposits were laid down, by identifying during excavation any related sequence of later human disturbance, movement or any subsequent natural disturbance or inverting of the layers. Thus the stratigraphic relation '*stratigraphically below*' is a record that refers to the earlier unit of stratigraphy being placed 'below' the later unit of stratigraphy in the resulting Harris matrix record. The '*stratigraphically below*' relationship does not necessarily mean it is always physically 'below' that stratigraphic unit in the ground. For example, a deep pit cut down through natural strata will have man-made deposits in the bottom which are physically below 'natural' bedrock, but the natural bedrock deposits will still be '*stratigraphically below*' the (temporally later) man-made intrusions. Sometimes this relationship is expressed as '*stratigraphically before*' - to clarify that it is a temporal rather than physical relationship.

Where templates produce RDF data in a CRM-compatible form, a single named column may implement a chain of entities and properties. This section discusses stratigraphic relationships as an illustrative example of one of the more complex data patterns encountered in the CRM-EH model. The model inherits the event based nature of its parent ontology CRM, so a direct stratigraphic relationship between archaeological contexts is expressed in a somewhat verbose fashion by relating context entities to context formation/deposition events (where each context has exactly one associated formation/deposition event). These events are then related

Figure 2. Simple example of a StringTemplateGroup file - describing some Dublin Core meta-data for resources





*Figure 3. Example tabular input for the template group - and resultant (RDF/XML) output*

<b>Tabular input:</b>		
id	author	title
123	Jane Austen	Pride and Prejudice
234	William Shakespeare	Twelfth Night
345	Charles Dickens	David Copperfield

<b>RDF/XML output:</b>		
<pre> &lt;?xml version="1.0" encoding="UTF-8"?&gt; &lt;rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"   xmlns:dct="http://purl.org/dc/terms/" xmlns:foaf="http://xmlns.com/foaf/0.1/"&gt;    &lt;rdf:Description rdf:about="http://example.com/id/123"&gt;     &lt;dct:identifier&gt;123&lt;/dct:identifier&gt;     &lt;dct:creator&gt;       &lt;foaf:Person&gt;         &lt;foaf:name xml:lang="en"&gt;Jane Austen&lt;/foaf:name&gt;       &lt;/foaf:Person&gt;     &lt;/dct:creator&gt;     &lt;dct:title xml:lang="en"&gt;Pride and Prejudice&lt;/dct:title&gt;   &lt;/rdf:Description&gt;    &lt;rdf:Description rdf:about="http://example.com/id/234"&gt;     &lt;dct:identifier&gt;234&lt;/dct:identifier&gt;     &lt;dct:creator&gt;       &lt;foaf:Person&gt;         &lt;foaf:name xml:lang="en"&gt;William Shakespeare&lt;/foaf:name&gt;       &lt;/foaf:Person&gt;     &lt;/dct:creator&gt;     &lt;dct:title xml:lang="en"&gt;Twelfth Night&lt;/dct:title&gt;   &lt;/rdf:Description&gt;    &lt;rdf:Description rdf:about="http://example.com/id/345"&gt;     &lt;dct:identifier&gt;345&lt;/dct:identifier&gt;     &lt;dct:creator&gt;       &lt;foaf:Person&gt;         &lt;foaf:name xml:lang="en"&gt;Charles Dickens&lt;/foaf:name&gt;       &lt;/foaf:Person&gt;     &lt;/dct:creator&gt;     &lt;dct:title xml:lang="en"&gt;David Copperfield&lt;/dct:title&gt;   &lt;/rdf:Description&gt;  &lt;/rdf:RDF&gt; </pre>		

to each other using a subset of temporal relationships (*before/after/equals*) represented using CRM properties. The transformation of relational data to an event based model often requires the creation of implicit event entities - for example the relationship [Context 456] *stratigraphically before* [Context 123] would be represented by a chain of entities and relationships as shown in Figure 4. Reciprocal relationships may not actually be stated in the originating datasets. Some systems may employ deductive reasoning to automatically generate the implied inverse relationship [Context 123] *stratigraphically after* [Context 456] - but rather than presuming that end user applications possess any inbuilt reasoning capability the corresponding inverse path is also automatically generated by the template.

Figure 4 illustrates the entire chain of RDF entities & relationships (including inverse relationships) for modelling a stratigraphic relationship between two contexts, conforming to the CRM-EH model. This arrangement would be as generated by the relevant template from a tabular input record containing the *context\_id* and *strat\_lower\_id* named columns. It depicts how the intermediate entities in the chain (with suitable identifiers) are generated automatically to ensure consistent and valid data conforming to a recognised ontological model by English Heritage representing current best practice guidance for representing archaeological contexts and stratigraphic relationships between them. See the manual for further details (STELLARc – section 3.2) or the template itself (STELLARb –CRMEH\_CONTEXTS.stg).

Figure 4. Chain of entities and relationships generated by the STELLAR CRMEH\_CONTEXTS template

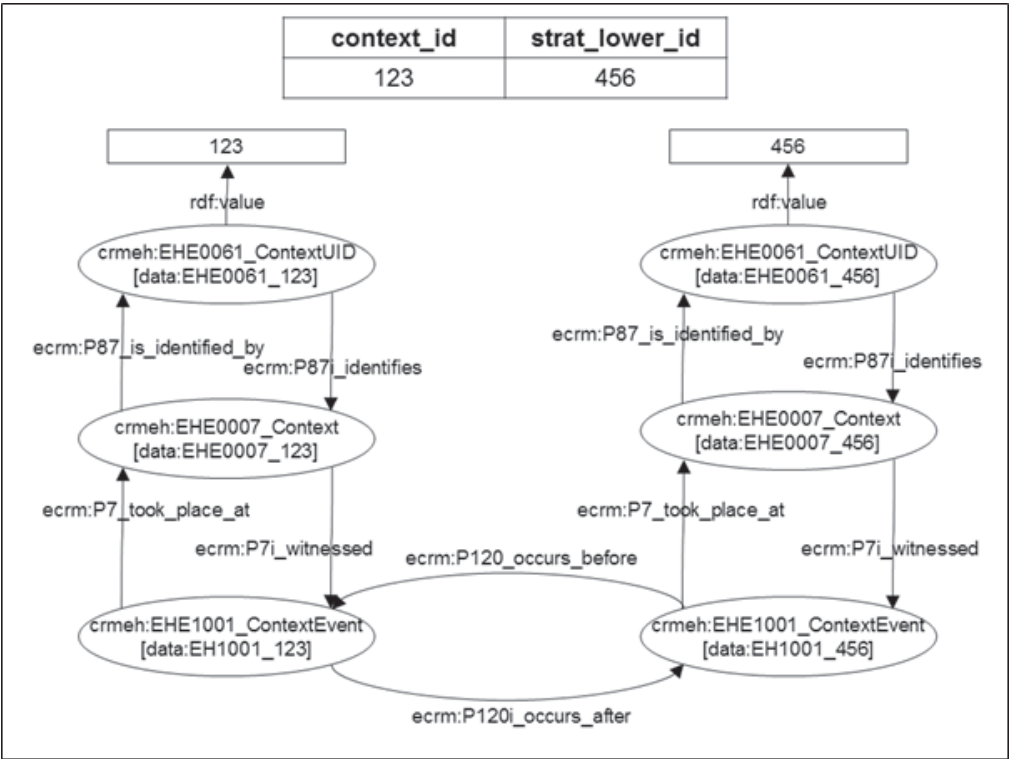


Figure 5 shows the resultant formatted RDF/XML data output. This data can be imported directly into an RDF aware application or a triple store for further processing, and for subsequent exposure as Linked Data<sup>2</sup>.

## 4. MAPPING, EXTRACTION AND LINKED DATA PUBLICATION AT ADS

As a case study of the tools and techniques developed for the project, data mapping and extraction was performed by the ADS, with the resulting output published as Linked Data. The process is described here, together with a discussion of issues encountered and reflections. The ADS staff engaged in the case study have archaeological experience and, although computer literate and familiar with databases and GIS applications, had no specialist knowledge nor previous experience with Linked Data or semantic technologies.

The case study employed 10 different datasets drawn from the two major archaeological programmes mentioned in Section 1.1, the Channel Tunnel Rail Link (CTRL) and the Aggregates Levy Sustainability Fund (ALSF). These datasets comprised excavation databases containing details of 18,619 archaeological contexts, 5,563 archaeological finds, and 234 environmental samples. The following datasets were downloaded from the CTRL excavation archive (doi:10.5284/1000230):

- Cobham Golf Course, Cobham, Kent (Museum of London Archaeology Service);
- Cuxton, Kent (Museum of London Archaeology Service);
- Eythorne Street, Hollingbourne, Kent (Oxford Archaeology / Wessex Archaeology);
- Saltwood Tunnel, Kent (Oxford Archaeology / Canterbury Archaeological Trust / Wessex Archaeology);
- West of Sittingbourne Road, Boxley, Kent (Oxford Archaeological Unit);
- Thurnham Villa, Kent (Oxford Archaeological Unit);
- Tutt Hill, Westwell, Kent (Museum of London Archaeology Service).

These datasets were chosen as they include good representative examples of databases produced for excavations undertaken by two of the largest commercial units in England (Oxford Archaeology and Wessex Archaeology). All of these databases included information typical of an excavation archive—contexts, stratigraphy, small finds and environmental sampling information. In addition, three ALSF datasets were also downloaded:

- Hartshill (doi:10.5284/1000365): an excavation database that included details of the earliest ironworking yet known in Britain. (Cotswold Archaeology);
- St Peter's Church, Barton-upon-Humber (doi:10.5284/1000389): A post-excavation database with extensive details of excavation of a medieval/post-medieval cemetery;
- Wellington Quarry, Worcestershire (doi:10.5284/1000392): a database from extensive excavations of a multi-phase prehistoric/Romano-British settlement and associated cemetery (Worcestershire Historic Environment and Archaeology Service).

### 4.1. Data Mapping and Extraction

Initially the test data was downloaded as delimited text (CSV) files and loaded into the STELLAR.Console application. Simple SQL queries aligned the existing data field names with column names expected by the existing templates, as shown in Figure 6.

Figure 5. Resultant RDF describing stratigraphic relationships between contexts

```

<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:ecrm="http://erlangen-crm.org/101001/"
  xmlns:crmeh="http://purl.org/crmeh#">
  <crmeh:EHE0007_Context rdf:about="http://stellar/test/test/EHE0007_123">
    <rdfs:label>123</rdfs:label>
  </crmeh:EHE0007_Context>
  <crmeh:EHE0061_ContextUID rdf:about="http://stellar/test/EHE0061_123">
    <rdf:value>123</rdf:value>
  </crmeh:EHE0061_ContextUID>
  <rdf:Description rdf:about="http://stellar/test/EHE0007_123">
    <ecrm:P87_is_identified_by rdf:resource="http://stellar/test/EHE0061_123"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://stellar/test/EHE0061_123">
    <ecrm:P87i_identifies rdf:resource="http://stellar/test/EHE0007_123"/>
  </rdf:Description>
  <crmeh:EHE1001_ContextEvent rdf:about="http://stellar/test/EHE1001_123"/>
  <crmeh:EHE1001_ContextEvent rdf:about="http://stellar/test/EHE1001_456"/>
  <rdf:Description rdf:about="http://stellar/test/EHE1001_123">
    <ecrm:P7_took_place_at rdf:resource="http://stellar/test/EHE0007_123"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://stellar/test/EHE0007_123">
    <ecrm:P7i_witnessed rdf:resource="http://stellar/test/EHE1001_123"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://stellar/test/EHE1001_456">
    <ecrm:P7_took_place_at rdf:resource="http://stellar/test/EHE0007_456"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://stellar/test/EHE0007_456">
    <ecrm:P7i_witnessed rdf:resource="http://stellar/test/EHE1001_456"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://stellar/test/EHE1001_456">
    <ecrm:P120_occurs_before rdf:resource="http://stellar/test/EHE1001_123"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://stellar/test/EHE1001_123">
    <ecrm:P120i_occurs_after rdf:resource="http://stellar/test/EHE1001_456"/>
  </rdf:Description>
</rdf:RDF>

```

Selectively extracting data via queries gives the flexibility to work with more manageable modular subsets of the data, although possibly introducing a potential risk whereby data may be unintentionally omitted or altered. This is typical of many ETL processes (even when using the entire dataset) - pointing to the need for assessment and validation of any output data to ensure that what is being produced is as originally intended.

*Figure 6. Example SQL query to align existing data fields with named template fields*

```

SELECT DISTINCT
  small_find_no AS find_id,
  description AS find_type,
  material AS find_material,
  context AS within_context_id,
  note AS find_note,
  period AS production_period
FROM Finds

```

There was a need for careful analysis of the original dataset to ensure that table fields were mapped to appropriate template fields. A number of the original datasets contained fields named 'Finds' which in reality related to 'Bulk Samples', an archaeological field recording process. The actual 'Finds' in the sense meant by the template (and by the CRM-EH) were in fields named 'Small Finds'. Another issue was the need to concatenate field values within, and sometimes across different tables. An example of this might be where the original data has multiple fields for 'period' say, 'start\_date' and 'end\_date' which need to be concatenated into a single field for mapping to the template 'production\_period' field.

Sometimes pre-processing SQL queries were employed to create intermediate tables appropriate for mapping to the template fields. For example, in some datasets multiple tables are used for small finds, say 'Coins', 'Ceramic Building Material', 'Animal Bone'. Or some datasets hold multiple 'description' fields, often with no clear motive (e.g. 'description1', 'description2', 'description3'), which map to the template 'notes' field.

The naming convention for resource identifiers comprised a base domain URI followed by a unique identifier string built from a combination of existing data values. However when combining datasets originating from multiple archaeological interventions the programme, site and project identifiers used by the various original data depositors were not necessarily unique. The ADS operate a Digital Object Identifier (DOI) allocation policy under the auspices of the DataCite project at the British Library and each archive is allocated a DOI within the ADS (e.g. 10.5284/1000365). Since DOIs are guaranteed to be unique the value could be incorporated forming a suitable unique base URI for each dataset:

- <http://data.archaeologydataservice.ac.uk/10.5284/1000365>
- <http://data.archaeologydataservice.ac.uk/10.5284/1000389>
- <http://data.archaeologydataservice.ac.uk/10.5284/1000392>

The CTRL datasets however shared a DOI (10.5284/1000230) as they were sub-projects of the overarching CTRL programme, so in this case the base URI was further extended by including the individual site names to ensure unique identification:

- <http://data.archaeologydataservice.ac.uk/10.5284/1000230/cobham/>
- <http://data.archaeologydataservice.ac.uk/10.5284/1000230/cuxton/>
- <http://data.archaeologydataservice.ac.uk/10.5284/1000230/eyhorne/>

## 4.2. Linked Data Publication

The RDF outputs from the STELLAR tools (644,114 triples) were imported into an Allegro-Graph (Franz Inc.) triple store. The ingest proved trivial, the RDF being correctly formatted by the STELLAR.Web tool. The Pubby open source tool was then used to publish the Linked Data (Cyganiak & Bizer, 2011). The Linked Data can be navigated via <http://data.archaeologydataservice.ac.uk/> and also queried from the SPARQL endpoint.

A number of minor technical issues were addressed when initially setting up the ADS Linked Data. AllegroGraph required the Linux 64 bit platform, necessitating a dedicated server at the ADS. The SPARQL endpoint and its management tool (AG WebView) operate on a specific port which could not be opened externally from the corporate University of York firewall. In order to enable external access to the SPARQL endpoint, a reverse proxy was configured which allowed access via the ADS domain.

Another minor hurdle was in ensuring URL safe characters were used in the Linked Data nodes. STELLAR employs existing data values in the construction of unique resource identifiers. As Linked Data identifiers specifically take the form of HTTP URIs, it was necessary to encode any 'unsafe' characters (e.g. spaces) in the original data to ensure that valid URIs were produced. Unfortunately pre-encoded identifiers were then being unencoded during Pubby queries – leading to requests for resources that did not exist. Two potential solutions were investigated: (i) removing problematic characters at the source and (ii) revising the STELLAR application to perform hash encryption of data values as an alternative to URI encoding that avoids identifier decoding problems (at the expense of more opaque URIs). The first approach was adopted in the case study – the existing data was pre-processed to replace any potentially problematic characters. The STELLAR application was however also revised to allow the *option* of producing hashed identifiers.

## 5. USING THE DATA

The final stage of the case study demonstrates that the exercise of mapping, extracting and publishing Linked Data addresses the goal of semantic interoperability by affording meaningful cross search over disparate datasets from different archaeological organisations with different excavation data recording systems. While detailed consideration of possible consuming applications for Linked Data is beyond the scope of the paper, this section first discusses querying of the SPARQL endpoint and then goes on to briefly consider possibilities for higher level applications or APIs that can offer semantic search but do not expose the complexity of the ontology or require SPARQL querying.

### 5.1. SPARQL Endpoint Querying

The ADS exposes a test page including some predefined SPARQL queries to illustrate syntax and to test the endpoint. Using the SPARQL endpoint, with some knowledge of SPARQL and the ontology, queries based on the use cases as shown in Figure 7 are possible (conducted during June 2013).

The results shown in Figure 7 are two contexts originating from the CTRL Thurnham dataset. This result could probably have been achieved without any semantic integration using the native database system. Consider however the results from a similar query shown in Figure 8.

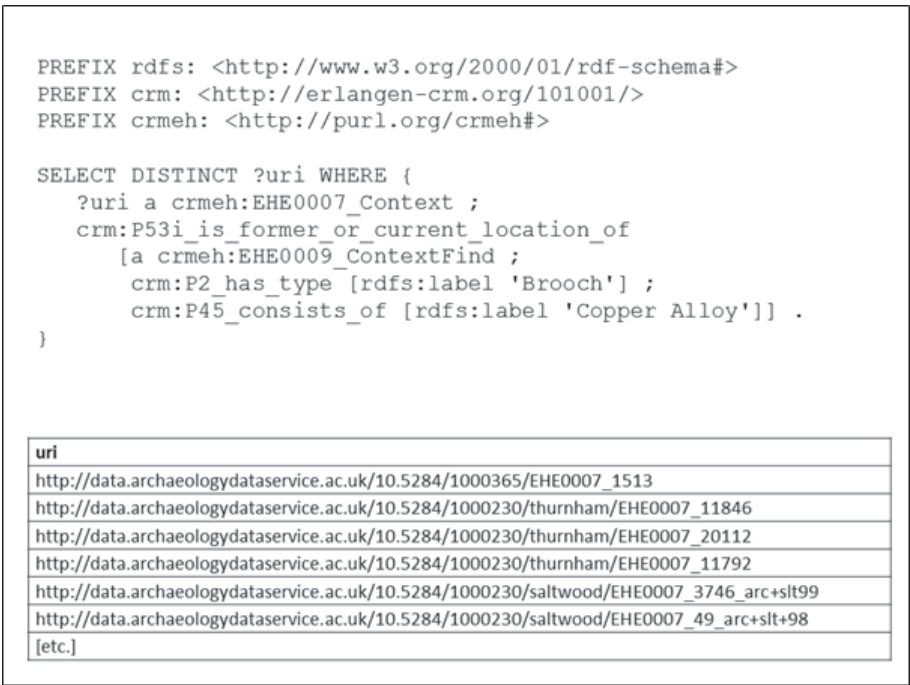
The result items this time originate from three separate datasets:



Figure 7. SPARQL query and results for contexts containing counters made of bone



Figure 8. SPARQL query and results for contexts containing copper alloy brooches



- ALSF Hartshill project (Cotswold Archaeology);
- CTRL Thurnham (Oxford Archaeology);
- CTRL Saltwood (Oxford Archaeology / Canterbury Archaeological Trust / Wessex Archaeology).

This demonstrates that cross search using a single query across multiple archaeological datasets originating from multiple data providers has been achieved. There are however a number of issues associated with use of a SPARQL endpoint to facilitate querying:

- Users need to know details of the syntax for constructing SPARQL queries;
- Users need to know the underlying ontological schema and associated namespaces in order to formulate their own queries;
- It is easy to ‘trip up’ the endpoint by submitting a badly formed query, or to submit a query that would result in an excessive amount of data being returned.

A user with sufficient knowledge could manually run a series of queries to incrementally determine the underlying schema, exploring the general structure of the data exposed. However applications, services, agents, tools and techniques can make the usage of endpoints easier. Formulating SPARQL queries utilizing the graphical user interface of the STAR project query builder was a simpler process, as discussed in the next section.

## 5.2. Hiding Complexity via a Higher Level Query

Figure 9 illustrates an equivalent query to Figure 7 using the STAR query builder, part of the Demonstrator from the earlier phase of the research (see Section 1.1), which operates over a similar set of excavation datasets to those discussed in Section 1. The Demonstrator seeks to hide the complexity of the underlying ontology and (like the templates) is based on the key archaeological concepts of Samples, Finds, Contexts or interpretive Groups (of contexts). We use it here as an example of possibilities with future APIs or higher level applications. A SPARQL query is automatically constructed based on the relevant CRM (and CRM-EH) ontological elements. The Demonstrator query builder is a JavaScript application which generates SPARQL queries, executing them on the server via a series of AJAX calls to a SOAP Web service.

This example STAR query has returned context records originating from different datasets (here LEAP and MoLAS) demonstrating cross search on multiple archaeological datasets. Similarly Figure 10 illustrates a similar query to the basic SPARQL query in Figure 8.

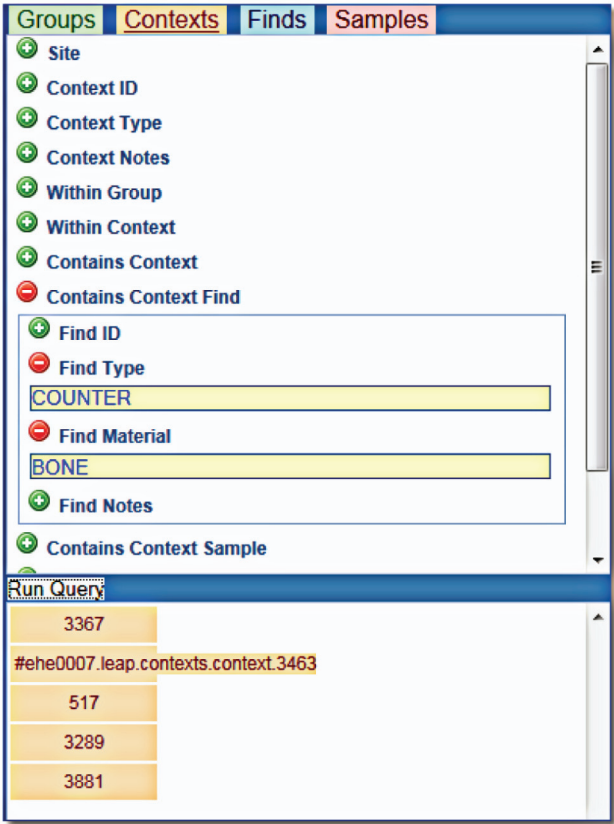
Figure 10 shows that in this case the query has returned a single context record originating from the MOLAS (Museum of London) dataset.

The basic SPARQL queries shown in Figure 7 and Figure 8, together with their associated results demonstrate that semantic cross search has been achieved using the approach described. The queries of the STAR Demonstrator in Figure 9 and Figure 10 illustrate what is possible with a higher level user interface and also suggest what might be possible using an API with data elements similar to those employed in the STAR query builder.

## 5.3. External Linking

Hand crafted triples to the Ordnance Survey (OS) Linked Data associated with ADS “project” nodes were the first step - linking the spatial coverage of the excavation to the definitive source of information about UK places. The direct link is an improvement on previous ADS practice. For

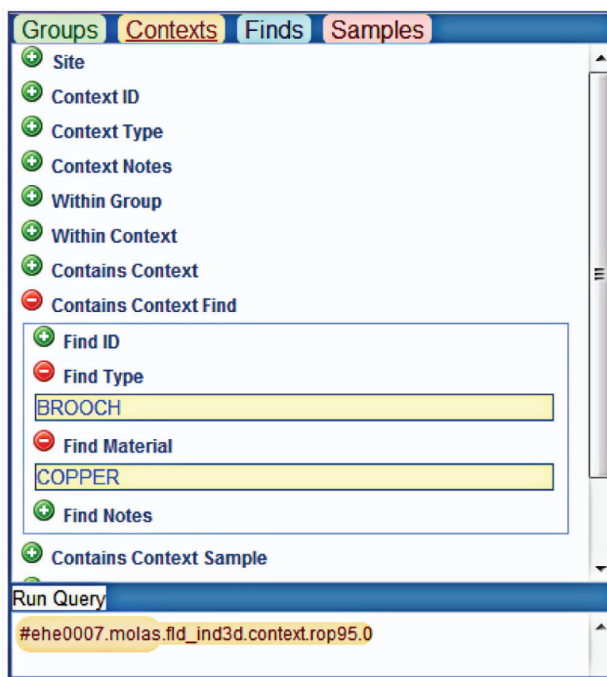
Figure 9. Query on STAR project query builder



example, each of the excavation datasets described comes from an ADS archive, where management and resource discovery metadata is available in the archive record for a dataset. However, an ADS archive with a location property of *Swindon* retains a level of ambiguity - since there is a Swindon in Wiltshire, Gloucestershire and Staffordshire. To associate the archive metadata (and therefore the excavation data) with the ‘correct’ Swindon, the ADS had explicitly listed the whole place name hierarchy in the indexing of each record (e.g. Europe, United Kingdom, England, Gloucestershire, Swindon). In contrast Linked Data URIs can be used to unambiguously identify a particular place; the OS Linked Data for Swindon in Gloucestershire (<http://data.ordnancesurvey.co.uk/id/7000000000020349>) also features other properties such as centroid coordinates, extent geometry (GML), hierarchical containment and adjacency.

Although the STELLAR tools are capable of generating controlled types of monuments, finds or materials with unique (URI) identifiers, standard external URIs representing the domain thesaurus concepts were not available when the Linked Data was created. This issue has been a key feature of our subsequent work (SENESCHAL), which recently published Linked Data SKOS representations of national archaeological vocabularies, including concepts for various types of monuments, archaeological objects and time periods. As ongoing work, the ADS are incorporating external links to these thesauri and other resources. For example, the ADS has begun exposing its archive level metadata as Linked Data and aligning this metadata to other

Figure 10. Query on STAR project query builder



external vocabularies and authorities. In addition to the Ordnance Survey and SENESCHAL vocabularies, links have been made to GeoNames, the Library of Congress Subject Headings, the Natural Environment Research Council Subject Categories, DBpedia concepts, and Lexvo.org. The internally developed tool used to align its archive level metadata with these authorities is reusable and has been integrated into the ADS Collections Management System, in order to enable the efficient alignment of metadata within existing workflows.

## 6. CONCLUSION

The STELLAR applications enable a consistent process for extracting a volume of data from raw delimited text files and converting to RDF conforming to existing ontological models (CRM and CRM-EH). The tools were employed by non-specialist (in semantic technologies) archaeology users to extract and convert Linked Data from major excavation datasets. The example queries from the ADS SPARQL endpoint show that the semantic integration necessary for cross search of different datasets was achieved.

The higher level query examples point to possible avenues for domain specific applications that leverage Linked Data publications. The main disadvantage of direct interaction with the SPARQL endpoint is the prerequisite knowledge of the SPARQL language and the underlying schemas within the triple store, neither of which are currently well understood within the archaeological domain. Exposing the triple store as Linked Data is seen as a good first step to developing understanding within the domain. In addition, an intermediate data access layer that exposed a particular set of query capabilities (such as a RESTful API) could provide easier access

to the underlying data, provided that it corresponds to common use cases, such as cross search. This approach may reduce some flexibility for querying, but it removes the need for end users to learn SPARQL syntax or ontology details and prevents the issuing of badly formed, inefficient or ineffective queries by encapsulating and hiding the detail of the underlying implementation.

Templates serve to document the mapping decisions and present a simplified interface to the underlying target model, providing a navigable (to non-specialists) bridge from relational to ontological structure without involving an intermediary mapping language. Declarative mapping languages have their own particular advantages and may simplify the process of creating a complex, new mapping, assuming the required technical expertise is available. On the other hand, a template-based approach may offer a gentler learning curve for non-technical users, especially when templates for key use cases are provided. The comparison is complicated in that approaches employing mapping languages may additionally hold a library of mapping examples, which could be presented as templates to data providers and thus in practice the distinction may be blurred. One point of difference remains as to whether the aim is to map and extract a complete dataset, or whether it is sufficient to focus on a subset of mapping patterns for the use cases envisaged.

In our view, the pattern-based ETL approach lends itself to the relatively common situation of legacy datasets that may not be well structured and that may require significant data cleansing. The process can begin from a set of delimited text files. Usage guidelines can be expressed in terms of the underlying domain use cases. The need for specialised knowledge of the ontology in question and semantic techniques generally is reduced. The templates take care of low level details such as creation of bidirectional links, RDF namespaces, implicit event entities, identifier formats, maintaining general consistency etc. The ability to create a sequence of commands in a batch file facilitates building up a repeatable processing chain for an entire dataset (or for a collection of datasets).

The resulting Linked Open Data has been published online by ADS and it is envisaged this will serve as a catalyst for further developments in archaeological Linked Open Data. Although this initial offering was regarded as an experimental foray into Linked Open Data production the resulting dataset is considered a permanent ADS resource. It has set in motion a further project whereby English Heritage excavation records held in a proprietary object oriented database system will be imported to the ADS Linked Open Data store after conversion using the STELLAR tools. This will represent a significant extension of the data published via this route and it is hoped that it will form the core of a resource large enough that the benefits of publishing as Linked Open Data become apparent to the broader archaeological community and that this can become a standard option for the deposit of excavation databases with ADS in future. This has the potential to create a rich, cross-searchable dataset directly amenable to answering more profound research questions than the current usual archival arrangements allow.

The repurposing and potential for subsequent reuse of Linked Open Data may necessitate a re-evaluation of existing ADS licensing arrangements particularly for legacy datasets. The ADS operate a licensing regime that does not imply transfer of copyright. In essence the ADS has a non-exclusive license to distribute material but without the right to distribute data for commercial re-use. In order for this to be considered legally enforceable, use of the data takes place only after the user has actively clicked to accept the ADS terms of use and access. This access model does not suit the Linked Open Data approach of directly accessing material via a SPARQL end point. Therefore changes need to be made to the terms and mode of access if not the nature of the deposit licence. The original licence arrangement did not envisage the direct access to datasets that the Linked Open Data architecture facilitates. There are a number of alternative ways to licence data (e.g. Creative Commons), but data already deposited with the ADS under the existing licence would need additional permissions from the original depositors

prior to publication in this form. If Linked Open Data publication is anticipated then datasets should be deposited with a modified licence that does not require an interactive user acceptance of the terms and conditions of access.

## 6.1. Future Work

The adoption of a common integrating ontological structure is only a partial solution towards full interoperability. It is also necessary to align terminology where possible. Alignment to and possibly between known controlled vocabularies and authorities (such as glossaries and thesauri) would significantly increase the potential for practical wider interoperability between datasets. Major archaeological controlled vocabularies have recently been exposed online as Linked Open Data. Alignment of the published Linked Data to these new resources is an area of ongoing current work within ADS (see Section 5.3).

Accompanying metadata expressing the status and provenance of the derived data would serve to clarify the scope for repurposing and trusting of such data. Archaeological datasets can reflect different stages of the excavation recording, analysis and publication process. This should be captured in the provenance metadata. There is also a need to accommodate the modelling of uncertainty in order to more accurately reflect the realities of the data encountered.

In other work, STELLAR tools were used as part of PhD research by an ADS colleague. Elements of two additional excavation datasets were extracted and aligned to the CRM-EH using the STELLAR Web application. Wright (2011) discusses how tools like STELLAR make it easier for archaeologist users to perform the technical tasks associated with semantic mapping, extraction and publication as Linked Data.

The STELLAR templates express common mapping patterns that have been applied to the archaeology domain in this work. However, the approach has potential in other areas where there is a need to consistently transform large volumes of tabular data to other structured formats. While selecting a mapping from a template might be seen as a reduction of flexibility in mapping from dataset to ontology, there is also the option of creating a new user-defined template. Creating a new template does require more programming expertise unless it is a simple tailoring of an existing template. With user-defined templates, the onus falls on end users to agree on and create the templates, and to ensure the validity of the output. However user-defined templates are text files that may be freely shared and edited. This may assist the evolution of (and convergence on) common approaches to mapping within user communities, particularly where alternative valid mappings are possible, as discussed in Section 2.3.

The current set of templates largely correspond to the general aim of cross searching excavation datasets for inter-site analysis and comparison. Different templates drawing on other areas of the ontology could be designed for purposes such as project management or detailed intra-site analysis. For example, the recent creation of Linked Data for the digitisation of a Palaeolithic archive, commissioned by English Heritage, involved a new set of user-defined templates, involving other areas of the CIDOC CRM (Cripps 2014). The CLAROS project templates mentioned in Section 2.4 are a step towards application of the techniques to a different domain (Classical Art). Current work is exploring the application of the approach to non-English datasets and vocabularies in a collaborative European project on archaeological research infrastructure (ARIADNE).

## ACKNOWLEDGMENT

The STELLAR project was supported by the UK Arts and Humanities Research Council [grant number AH/H037357/1]. Thanks are also due to the participants of the STELLAR workshops.



## REFERENCES

- ALSF. *Aggregates Levy Sustainability Fund Programme Overview*. Archaeology Data Service Archive. Retrieved May 1<sup>st</sup> 2014, from <http://archaeologydataservice.ac.uk/archives/view/alsf/>
- Archaeology Data Service. *Linked Data*. Retrieved May 1<sup>st</sup> 2014, from <http://data.archaeologydataservice.ac.uk/>
- ARIADNE FP7 Project. *Advanced Research Infrastructure for Archaeological Dataset Networking in Europe*. Retrieved May 1<sup>st</sup> 2014, from <http://www.ariadne-infrastructure.eu/>
- Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., & Aumüller, D. (2009). Triplify – Light-Weight Linked Data Publication from Relational Databases. In *Proceedings 18th international conference on World Wide Web (WWW '09)*, (pp. 621-630). ACM Press. doi:10.1145/1526709.1526793
- Barrasa, J., Corcho, Ó., & Gómez-Pérez, A. (2004). R2O, an Extensible and Semantically Based Database-to-Ontology Mapping Language. In *Proceedings 2nd International Workshop on Semantic Web and Databases (SWDB 2004)*, Lecture Notes in Computer Science, 3372, (pp. 1069-1070). Springer
- Barrasa, J., & Gómez-Pérez, A. (2006). Upgrading relational legacy data to the semantic web. In *Proceedings 15th International Conference on World Wide Web (WWW'06)*, (pp. 1069 – 1070). ACM Press.
- Berners-Lee, T. (1998). *Relational Databases on the Semantic Web*. Retrieved May 1<sup>st</sup> 2014, from <http://www.w3.org/DesignIssues/RDB-RDF.html>
- Binding, C., & Tudhope, D. (2008). Using Terminology Web Services for the Archaeological Domain. In *Proceedings 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008)*. Lecture Notes in Computer Science, 5173, (pp. 392-393). Springer. doi:10.1007/978-3-540-87599-4\_42
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 1–22. doi:10.4018/jswis.2009081901
- Byrne, K. (2008). Having Triplets - Holding Cultural Data as RDF. In M. Larson et al. (Eds.), *Proceedings of the ECDL 2008 Workshop on Information Access to Cultural Heritage*.
- Cerbah, F. (2008). Learning Highly Structured Semantic Repositories from Relational Databases - RD-BtoOnto Tool, In *Proceedings 5th European Semantic Web Conference (ESWC 2008)*. Lecture Notes in Computer Science, 5021, (pp. 777-781). Springer. CIDOC CRM: *CIDOC Conceptual Reference Model*. Heraklion, Crete: Institute of Computer Science, Foundation for Research and Technology. Retrieved May 1<sup>st</sup> 2014, from <http://www.cidoc-crm.org/>
- CIDOC CRM Special Interest Group (CRM SIG). Retrieved May 1<sup>st</sup> 2014, from <http://network.icom.museum/cidoc/working-groups/crm-special-interest-group/>
- CLAROS. *The world of art on the semantic web*. Retrieved May 1<sup>st</sup> 2014, from <http://www.clarosnet.org/>
- Cripps, P. (2014). *Colonisation of Britain Project*. Retrieved May 30<sup>th</sup> 2014, from <http://www.archaeogeomancy.net/2014/05/colonisation-of-britain>
- Cripps, P., Greenhalgh, A., Fellows, D., May, K., & Robinson, D. (2004). *Ontological modelling of the work of the Centre for Archaeology* (CIDOC CRM Technical Paper). Retrieved May 1<sup>st</sup> 2014, from [http://www.cidoc-crm.org/technical\\_papers.html](http://www.cidoc-crm.org/technical_papers.html)
- CRM-EH. *English Heritage Extension to CRM for the archaeology domain*, Retrieved May 1<sup>st</sup> 2014, from <http://hypermedia.research.southwales.ac.uk/resources/crm/>
- Cullot, N., Ghawi, R., & Yétongnon, K. (2007). DB2OWL: A tool for automatic database-to-ontology mapping. In M. Ceci et al. (Eds.): *Proceedings 15th Italian Symposium on Advanced Database Systems (SEBD 2007)*. Retrieved May 1<sup>st</sup> 2014, from <http://le2i.cnrs.fr/le2i/IMG/publications/DB2OWL1.pdf>
- Cyganiak, R., & Bizer, C. (2011). *Pubby: a linked data front end for SPARQL endpoints*. Freie Universität, Berlin. Retrieved May 1<sup>st</sup> 2014, from <http://www4.wiwiwiss.fu-berlin.de/pubby/>

D2R Server: *Accessing databases with SPARQL and as Linked Data*. Retrieved May 1<sup>st</sup> 2014, from <http://www4.wiwi.fu-berlin.de/bizer/d2r-server/>

D2RQ: *Accessing Relational Databases as Virtual RDF Graphs*. Retrieved May 1<sup>st</sup> 2014, from <http://d2rq.org/>

Dodds, L., & Davis, I. (2012). *Linked Data Patterns – A pattern catalogue for modelling, publishing and Consuming Linked Data*. Retrieved May 1<sup>st</sup> 2014, from <http://patterns.dataincubator.org/book/>

Doerr, M. (2003). The CIDOC conceptual reference model: An ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3), 75–92.

Doerr, M., Gradmann, S., Henniecke, S., Isaac, A., Meghini, C., & Van de Sompel, H. (2010). The Europeana Data Model (EDM). In *Proceedings 76th IFLA General Conference*. Retrieved December 9<sup>th</sup> 2014, from <http://conference.ifla.org/past-wlic/2010/149-doerr-en.pdf>

Eide, Ø., Felicetti, A., Ore, C. E., D'Andrea, A., & Holmen, J. (2008). Encoding Cultural Heritage Information for the Semantic Web - Procedures for Data Integration through CIDOC-CRM Mapping. In *Proceedings EPOCH Conference on Open Digital Cultural Heritage Systems*, (pp. 1–7). Retrieved May 1<sup>st</sup> 2014, from <http://public-repository.epoch-net.org/rome/05%20Procedures%20Data%20Integration.pdf>

Erling, O., & Mikhailov, I. (2009). *Mapping relational data to RDF in Virtuoso*. White paper. Retrieved May 1<sup>st</sup> 2014, from <http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/VOSSQLRDF>

Foreman, S. (2004). *Channel Tunnel Rail Link Section 1*. Archaeology Data Service Archive. Retrieved May 1<sup>st</sup> 2014, from 10.5284/1000230

Franz Inc. *AllegroGraph server*. Retrieved May 1<sup>st</sup> 2014, from <http://www.franz.com/agraph/allegrograph/>

Gangemi, A. (2005). Ontology Design Patterns for Semantic Web Content. In: Y. Gil et al. (Eds.): *Proc. International Semantic Web Conference (ISWC)*, LNCS 3729, (pp. 262 – 276). Springer. *Gruff: a grapher-based triple-store browser for AllegroGraph*. Retrieved May 1<sup>st</sup> 2014, from <http://www.franz.com/agraph/gruff/>

Harris, E. (1989). *Principles of archaeological stratigraphy*. Academic Press Inc; 2nd Revised edition.

Hert, M., Reif, G., & Gall, H. C. (2011). A Comparison of RDB-to-RDF Mapping Languages. In *Proceedings 7th International Conference on Semantic Systems (I-SEMANTICS 2011)*, (pp 25-32). ACM Press.

Isaac, A. (2011). *Europeana Data Model Primer*. Retrieved May 1<sup>st</sup> 2014, from <http://pro.europeana.eu/edm-documentation>

Kondylakis, H., Doerr, M., & Plexousakis, D. (2006). *Mapping Language for Information Integration* (Technical Report 385). Heraklion: Institute of Computer Science, Foundation for Research and Technology. Retrieved May 1<sup>st</sup> 2014, from [http://www.cidoc-crm.org/docs/Mapping\\_TR385\\_December06.pdf](http://www.cidoc-crm.org/docs/Mapping_TR385_December06.pdf)

Kurtz, D., Parker, G., Shotton, D., Klyne, G., Schroff, F., Zisserman, A., & Wilks, Y. (2009). CLAROS - Bringing Classical Art to a Global Public. In *Proceedings 5th IEEE International Conference on e-Science*, (pp. 20-27). IEEE. *Linked Data – Connect Distributed Data Across The Web*. Retrieved May 1<sup>st</sup> 2014, from <http://linkeddata.org/>

Linking Museums, I. I. I. *Museums and the Machine Processable Web*. Retrieved May 1<sup>st</sup> 2014, from <http://museum-api.pbworks.com/w/page/33068521/Linking%20Museums%20III%3A%20'people'%20records>

May, K., Binding, C., Tudhope, D., & Jeffrey, S. (2011). Semantic Technologies Enhancing Links and Linked Data for Archaeological Resources, In M. Zhou et al. (Eds.), *Proceedings 39th Conference on Computer Applications and Quantitative Methods in Archaeology (CAA2011)*, (pp. 261-272). Amsterdam University Press *NeOn Toolkit*. Retrieved May 1<sup>st</sup> 2014, from <http://neon-toolkit.org/>

Nußbaumer, P., & Haslhofer, B. (2007). *Putting the CIDOC CRM into Practice – Experiences and Challenges*. (Technical Report TR-200). University of Vienna. Retrieved May 1<sup>st</sup> 2014, from <https://eprints.cs.univie.ac.at/404/>

Nußbaumer, P., Haslhofer, B., & Klas, W. (2010). *Towards Model Implementation Guidelines for the CIDOC Conceptual Reference Model*. Technical Report TR-201. University of Vienna. Retrieved May 1<sup>st</sup> 2014, from <http://eprints.cs.univie.ac.at/58/>

ODP - Ontology Design Patterns. Semantic Web portal dedicated to ontology design patterns. Retrieved May 1<sup>st</sup> 2014, from <http://ontologydesignpatterns.org>

OpenRefine. Data transformation tool (formerly Google Refine). Retrieved May 1<sup>st</sup> 2014, from <http://openrefine.org/>

Ordnance Survey Linked Data Platform. Retrieved May 1<sup>st</sup> 2014, from <http://data.ordnancesurvey.co.uk>

Parr, T. *StringTemplate template engine*. University of San Francisco. Retrieved May 1<sup>st</sup> 2014, from <http://www.stringtemplate.org/>

Pérez de Laborda, C., & Conrad, S. (2005). Relational.OWL - A Data and Schema Representation Format Based on OWL. In *Proceedings Conceptual Modelling 2005, Second Asia-Pacific Conference on Conceptual Modelling (APCCM2005)*, (pp. 89-96). Retrieved May 1<sup>st</sup> 2014, from <http://crpit.com/confpapers/CRPITV43deLaborda.pdf>

Priyatna, F., Villazón Terrazas, B. M., Barrasa, J., & Schulte, J. (2011). *ODEMapster plugin for the NeOn toolkit*. Retrieved May 1<sup>st</sup> 2014, from <http://neon-toolkit.org/wiki/ODEMapster>

R2RML: *RDB to RDF mapping language*. Retrieved May 1<sup>st</sup> 2014, from <http://www.w3.org/TR/r2rml/>

RDF Refine, Extension to OpenRefine tool for exporting RDF. Retrieved May 1<sup>st</sup> 2014, from <http://refine.deri.ie/>

Richards, J., & Hardman, C. (2008). Stepping Back from the Trench Edge. In: (Eds Greengrass, Hughes) *The Virtual Representation of the Past*. Ashgate, (pp. 167-168). Retrieved May 1<sup>st</sup> 2014, from <http://eprints.whiterose.ac.uk/7795/>

Sahoo, S. S., Halb, W., Hellman, S., Idehen, K., Thibodeau, T., Auer, S., . . . Ezzat, A. (2009). *A Survey of Current Approaches for Mapping of Relational Databases to RDF*. W3C RDB2RDF Incubator Group. Retrieved May 1<sup>st</sup> 2014, from [http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF\\_SurveyReport.pdf](http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf)

Schiemann, B., Oischinger, M., & Görz, G. (2012). *Erlangen CRM/OWL*. Friedrich-Alexander University of Erlangen-Nuremberg. Retrieved May 1<sup>st</sup> 2014, from <http://erlangen-crm.org/>

SENESCHAL Project. *Semantic ENrichment Enabling Sustainability of arCHaeological Links*. University of South Wales: Hypermedia Research Group. Retrieved May 1<sup>st</sup> 2014, from <http://hypermedia.research.southwales.ac.uk/kos/seneschal/>

Shakya, S., Takeda, H., & Wuwongse, V. (2009). Community-Driven Linked Data: Authoring and Production of Consolidated Linked Data. *International Journal on Semantic Web and Information Systems*, 5(3), 23–48. doi:10.4018/jswis.2009081902

SKOS. *Simple Knowledge Organization System*. W3C. Retrieved May 1<sup>st</sup> 2014, from <http://www.w3.org/2004/02/skos/>

STAR Project. *Semantic Technologies for Archaeological Resources*. University of South Wales: Hypermedia Research Group. Retrieved May 1<sup>st</sup> 2014, from <http://hypermedia.research.southwales.ac.uk/kos/star/>

STELLARa. STELLAR Project and applications. *Semantic Technologies Enhancing Links and Linked Data for Archaeological Resources*. University of South Wales: Hypermedia Research Group. Retrieved May 1<sup>st</sup> 2014, from <http://hypermedia.research.southwales.ac.uk/resources/STELLAR-applications/>

STELLARb. STELLAR Introduction (pdf). University of South Wales: Hypermedia Research Group. Retrieved May 1<sup>st</sup> 2014, from <http://hypermedia.research.southwales.ac.uk/media/files/documents/2014-05-01/STELLAR.Introduction.pdf>

STELLARc. STELLAR Manual (pdf). University of South Wales: Hypermedia Research Group. Retrieved May 1<sup>st</sup> 2014, from <http://hypermedia.research.southwales.ac.uk/media/files/documents/2014-05-01/STELLAR.Applications.pdf>

*Triplify – expose semantics!* Retrieved May 1<sup>st</sup> 2014, from <http://triplify.org>

Tudhope, D., May, K., Binding, C., & Vlachidis, A. (2011). Connecting archaeological data and grey literature via semantic cross search. *Internet Archaeology*, 30, Open access. Retrieved May 1<sup>st</sup> 2014, from 10.11141/ia.30.5

Wright, H. (2011). *Seeing Triple: Archaeology, Field Drawing and the Semantic Web*. PhD thesis, University of York. Retrieved May 1<sup>st</sup> 2014, from <http://etheses.whiterose.ac.uk/2194/>

## ENDNOTES

- <sup>1</sup> The research projects (STAR, STELLAR) were funded by the UK Arts and Humanities Research Council and coordinated by the Hypermedia Research Group at the University of South Wales (formerly University of Glamorgan). They investigated research questions concerning semantic interoperability in the archaeology domain.
- <sup>2</sup> The work presented in this paper preceded the establishment of a fixed reference RDF namespace for CRM entities and properties. The namespace visible in the examples refers to the Erlangen OWL implementation of CRM (Schiemann, Oischinger & Görz, 2012) and could be revised to the newer reference namespace <http://www.cidoc-crm.org/cidoc-crm/> in future work.

# CALL FOR ARTICLES

## International Journal on Semantic Web and Information Systems

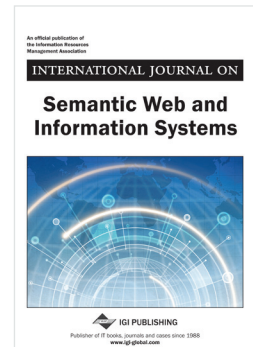
*An official publication of the Information Resources Management Association*

### MISSION:

The **International Journal on Semantic Web and Information Systems (IJSWIS)** is an archival journal that publishes high quality original manuscripts in all aspects of Semantic Web that are relevant to computer science and information systems communities. IJSWIS is an open forum aiming to cultivate the Semantic Web vision within the information systems research community. The main focus is on information systems discipline and working towards the delivery of the main implications that the Semantic Web brings to information systems and the information/knowledge society.

### COVERAGE/MAJOR TOPICS:

- Beyond Semantic Web (e.g., extending meaning with perception and experience)
- Enterprise application integration
- From e-government to e-democracy
- Integration with other disciplines
- Intelligent Systems
- Metadata-driven (bottom-up) versus ontology-driven (top-down) SW development
- New Semantic Web enabled business models
- New Semantic Web enabled information systems
- New Semantic Web enabled tools for the citizen/ learner/ organization/ business
- Ontologies, folksonomies, and associated knowledge representation issues
- Real world applications towards the development of the Knowledge society
- Semantic enabled business intelligence
- Semantic Web applications on the Web, enterprises, desktops, personal and mobile devices, e-science and e-government applications, and associated issues of provenance, trust, privacy, security, quality, scalability, and performance
- Semantic Web data management
- Semantic Web issues, challenges, and implications in each of the IS research streams
- Semantics and human computer interfaces including visualization and mashups
- Semantics in business processes and distributed computing and services
- Social Semantic Web and people Web
- Standards



ISSN 1552-6283  
eISSN 1552-6291  
Published quarterly

All inquiries regarding IJSWIS should be directed to the attention of:  
Sören Auer, Editor-in-Chief  
[ijswis@igi-global.com](mailto:ijswis@igi-global.com)

All manuscript submissions to IJSWIS should be sent through the online submission system:  
<http://www.igi-global.com/authorseditors/titlesubmission/newproject.aspx>

Ideas for Special Theme Issues may be submitted to the Editor-in-Chief.

**Please recommend this publication to your librarian. For a convenient easy-to-use library recommendation form, please visit:**  
**<http://www.igi-global.com/IJSWIS>**