Deciphering the code: implementing the CIDOC CRM at Priniatikos Pyrgos, Crete

Frank Lynam

Zusammenfassung – Dieser Beitrag beschreibt den Prozess der Verwendung von Ontologien, um digitale archäologische Datensätze für die Veröffentlichung im Semantic Web zu strukturieren. Zunächst werden die Begriffe und Themen Open Data, Linked Data und Semantic Web vorgestellt und erläutert. Anschließend legt der Aufsatz dar, wie die Erweiterungen von archäologischen Daten durch CIDOC CRM und English Heritage verwendet werden, um die Daten des Projekts Priniatikos Pyrgos im Osten von Kreta zu modellieren.

Schlüsselwörter – Archäologie, Big Data, Open Data, Linked Data, Semantic Web, RDF, Ontologie, Wortschatz

Abstract – This paper describes the process of using ontologies to structure digital archaeological datasets for publication on the Semantic Web. Open Data, Linked Data and the Semantic Web are all introduced and explained before the paper proceeds to discuss how the CIDOC CRM and its English Heritage archaeological data extension were used to model the data of the Priniatikos Pyrgos archaeological project in East Crete.

Key words – archaeology, Big Data, open data, linked data, semantic web, RDF, ontology, vocabulary

Introduction

„It is nevertheless the map that precedes the territory – precession of simulacra“
*(*Baudrillard 1994, 1*)*

The archaeological discipline in the sense that we understand it today is a relatively new entrant to the academy. It was really only in the excavations carried out by Augustus Pitt Rivers at Cranborne Chase in the 1880s and 1890s that we can begin to identify the methodological rigour and intellectual coherency that we associate with the workings of an independent field of academic research (Renfrew & Bahn 2004, 31). This period, however short, has nonetheless seen much intellectual and methodological change, reflecting both impulses originating from within archaeology and from without.

In recent times, we have witnessed the beginning of what might well prove with hindsight to be the most revolutionary of changes. The emergence of digital culture has challenged the central tenets of the archaeological praxis itself. It is difficult to find a single archaeological activity that has not witnessed some degree of transformation as a result of the digital arrival. Where would 21st century archaeological planning be without the scale and multiple viewpoints afforded by Geographical Information Systems? Is it possible to imagine a pre-Social Web archaeological knowledge dissemination system now that it is so deeply embedded in the proliferation of archaeological thought? The digital idea has been embraced by the majority of archaeologists. Its employment is seen throughout the discipline and the epistemological implications of this shift are profound and will take many years for the philosophers and historians of the field to untangle.

One aspect of the digital revolution that is particularly interesting is the field of data. At a prosaic level, archaeologists are now producing quantities of information that would have been simply unimaginable even ten years ago. In the wider world, these collections of information are being referred to as Big Data and this form and scale of information analysis is being heralded by some commentators as the future of industry and science (Kitchin 2014). Big Data also promises a new approach to the way that humanities and, specifically archaeological, research is conducted, but this paper argues that archaeology is still a long way off the point at which this promise can be delivered upon. The challenges are great. Big Data demands the investment of significant technological resources at a minimum and we will discuss in some detail Linked Open Data within this context. But often the greatest impediments to radical infrastructural change such as this are less technological and more sociological in character. Archaeologists will need to adopt new perspectives and forego many traditional norms if Big Data archaeology is to succeed in any meaningful way.

This paper addresses one of the key Big Data challenges from the perspective of a small to medium scale archaeological project. It considers the process of mapping the digital dataset of the site of Priniatikos Pyrgos to the CIDOC CRM ontology, so that it can be published to the Semantic Web as Linked Open Data.

Received: 16 Nov 2014
accepted: 29 March 2015
published online: 3 June 2015

The Priniatikos Pyrgos case study

Research, geographical and historical overview of the site

Priniatikos Pyrgos is a small[1] site (c1.26 ha area) located on the shores of the Mirabello Bay in East Crete **(Figure 1)**. It is situated on a slight rise in the landscape on a rocky limestone promontory on the outskirts of the town of Istron/Kalo Chorio. At this point in the landscape the coast is separated from the mountains by a thin strip of plain that over many millennia has functioned as the primary east-west communications corridor along the northern half of the island.

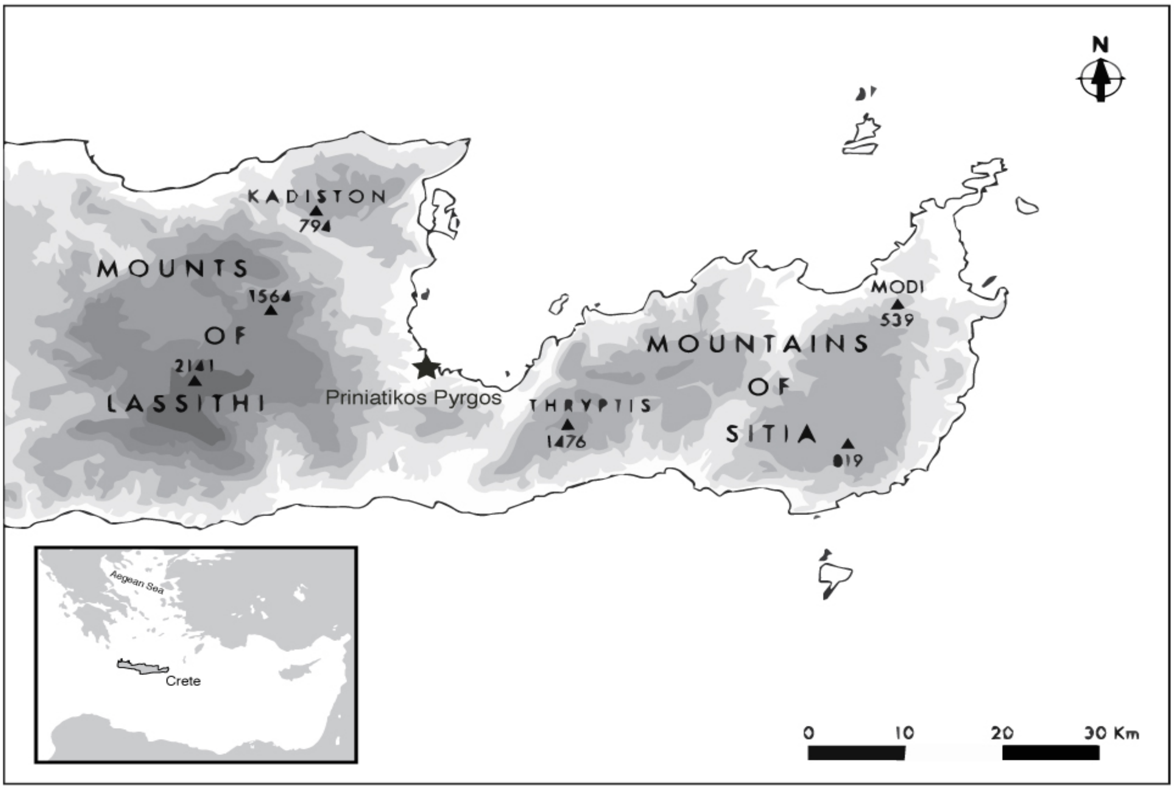


Fig. 1 The location of Priniatikos Pyrgos (after Molloy et al. 2014)

The site first entered the archaeological consciousness in the early 20th century when the pioneering American archaeologist, Edith Hall, sunk a test trench somewhere within its vicinity (Betancourt 2014, 11)[2]. It lay undisturbed for more than a century until 2005 when Dr Metaxia Tsipopoulou and Dr Barbara Hayden began a rescue excavation at the site (Hayden 2014, 15). This project was backed by the American School at Athens and continued for a further season ending in 2006. The following year the site's operation was handed over to the Irish Institute of Hellenic Studies at Athens and excavation continued until the final season in 2010.[3] Remains of human occupation spanning over 5,000 years were uncovered at the site, beginning in the Early Bronze Age (c.3100-2650 BCE) (Molloy et al. 2014, 1). For most of its occupation, the area functioned as an industrial quarter, probably on the edge of a larger settlement located to the south in the area that is now covered by the Istron River's floodplain. However, it became more monumental in the Byzantine period when a number of substantial buildings were erected on its summit.

Recording Priniatikos Pyrgos

On archaeological recording systems

The 2005-2007 rescue project decided to record its activities using the locus-pail system in keeping with the approach taken by the majority of North American-influenced projects operating in the Eastern Mediterranean region (Morgan 2010). Generally speaking, locus numbers are assigned to spatial areas that are delineated by architectural boundaries, most commonly walls, or an absence of cultural material. Pails subdivide loci into administrative and to a lesser extent stratigraphic units (Pavel 2012, 49–50).

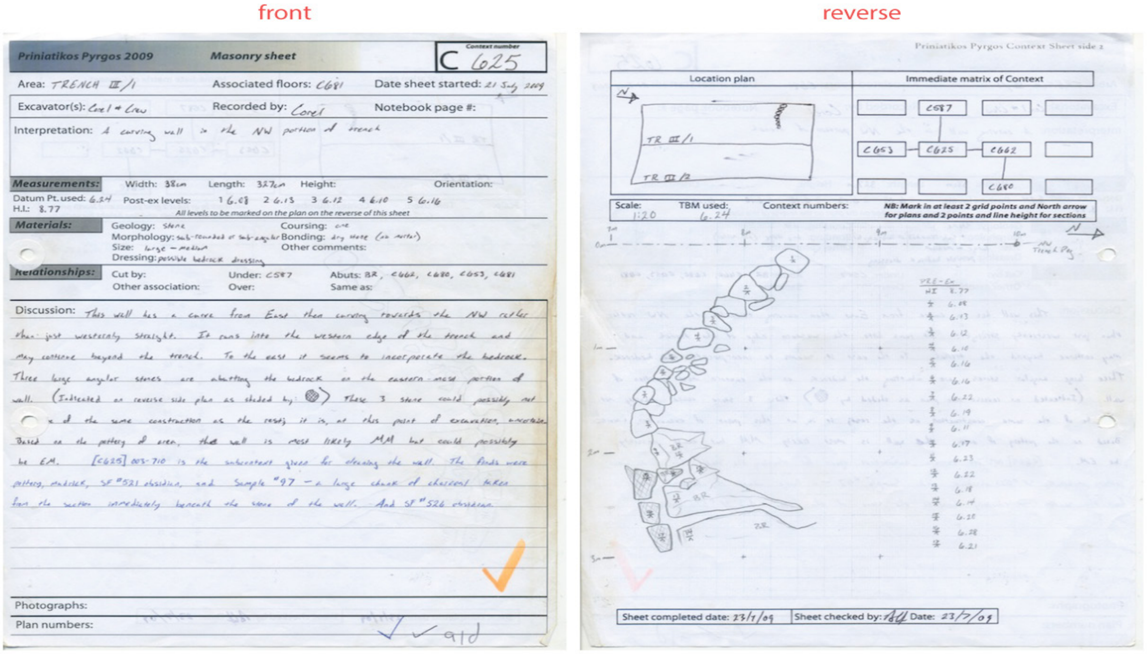
In 2007 the new project management decreed that there be a change in the recording system used at the site. Single Context recording is used by the majority of commercial archaeological units in the UK and Ireland as well as in pockets across the globe. The SC method states that "any single action, whether it leaves a positive or a negative record within the sequence, is known as a 'context'" (Museum of London 1994, sec. 1.2). As a field method, it goes hand-in-hand with the processes of planning and stratigraphic analysis, so much so in fact, that in the Museum of London's SC guide (which is almost universally acknowledged as the authority on SC), it is referred to as the 'Single Context Planning System'.

The Priniatikos Pyrgos data

The core of the Priniatikos Pyrgos dataset is, therefore, structured around the foundational units of the locus and the context. This is only the beginning of the site's data story, however. Archaeologists produce many different forms of information (excavation observations, quantitative data, plans, photographs) when they excavate a site or indeed survey an area, and the Priniatikos Pyrgos project is no different in this regard. Each of these data categories is tied to a particular activity or more often a range of activities that are conducted by the project's archaeologists. These records can be viewed as informational correlates of the physical activity that brought them into being.

The first point to make is that all of these records, excepting the photographs, which were captured using digital cameras, were created using paper, pen and pencil, a practice that is common across the vast majority of archaeological field projects (Ellis & Wallrodt 2011).

For the trenches that were excavated using the SC method, each context was recorded using a pro forma paper sheet that would be familiar to most archaeologists **(Figure 2).** Ostensibly objective fields such as 'context name', 'date of excavation', 'excavators involved' and more discursive fields like 'description' and 'comments' were filled out by the archaeologists on site and occasionally these were updated off site as new information came to light either as a function of reflection or feedback from the post-excavation team.


front
reverse
N
TR III/1
TR III/2
7m
8m
9m
10m
0m
NW Trench Peg
1m
2m
3m
BR

front

Priniatikos Pyrgos 2009 **Masonry sheet**

Context number: C 625

Area: TRENCH III/1 Associated floors: C681 Date sheet started: 21 July 2009

Excavator(s): Carl + Crew Recorded by: Carel Notebook page #:

Interpretation: A curving wall in the NW portion of trench

Measurements: Width: 38cm Length: 327cm Height: Orientation:

Datum Pt. used: 6.24 Post-ex levels: 1 6.08 2 6.13 3 6.12 4 6.10 5 6.16

H.I.: 8.77

All levels to be marked on the plan on the reverse of this sheet

Materials: Geology: stone Coursing: one

Morphology: sub-rounded or sub-angular Bonding: dry stone (no mortar)

Size: large – medium Other comments:

Dressing: possible bedrock dressing

Relationships: Cut by: Under: C587 Abuts: BR, C662, C680, C653, C681

Other association: Over: Same as:

Discussion: This wall has a curve from East then curving towards the NW rather than just westerly straight. It runs into the western edge of the trench and may continue beyond the trench. To the east it seems to incorporate the bedrock. Three large angular stones are abutting the bedrock on the eastern-most portion of wall. (Indicated on reverse side plan as shaded by: ●) These 3 stone could possibly not be of the same construction as the rest; it is, at this point of excavation, uncertain. Based on the pottery of area, the wall is most likely MM but could possibly be EM. [C625] 003-710 is the subcontext given for cleaning the wall. The finds were pottery, mudbrick, SF #521 obsidian, and Sample #97 – a large chunk of charcoal taken from the section immediately beneath the stone of the wall. And SF #526 obsidian.

✓

Photographs:

Plan numbers: ✓ ✓ ald

reverse

Priniatikos Pyrgos Context Sheet side 2

Location plan

Immediate matrix of Context

	C587		
C653	C625	C662	
		C680	

Scale: 1:20 TBM used: 6.24 Context numbers:

NB: Mark in at least 2 grid points and North arrow for plans and 2 points and line height for sections

PRE-Ex	
HI	8.77
1	6.08
2	6.13
3	6.12
4	6.10
5	6.16
6	6.16
7	6.22
8	6.19
9	6.11
10	6.17
11	6.23
12	6.22
13	6.18
14	6.14
15	6.20
16	6.28
17	6.21

Sheet completed date: 23/7/09 Sheet checked by: AH Date: 23/7/09

Fig. 2 A Priniatikos Pyrgos example context sheet.

An idiosyncrasy of the project's implementation of the SC method was its employment of a sub-context record. These allowed for the excavation of a context over a number of days, easing the administration of the material recovered and also protecting against the possibility of a new stratigraphic layer being missed. Sub-context pro forma sheets contained mainly information related to finds. The archaeologist would fill in the amount of ceramic, human bone, animal bone, shell, carbon and other artefact and ecofact categories of material that were recovered during the sub-context's excavation. An identical pro forma record was used to record pails.

The site registers performed an important administrative function on site. Photographs, taken by the archaeologists, were logged in the photo register. The many environmental samples taken were also logged in a separate register before the sample was explained on a pro forma sample sheet. In fact, all sheets created on site were registered using this system.

The post-excavational phase at the project followed a similar record creation scheme. Pottery readings, a key component of any archaeological post-excavational effort, were recorded on their corresponding sub-context or pail sheets. The other major contributor to the record base of the project during this phase was the cataloguing of particular ceramic and other artifactural pieces of interest. Again pro forma records were used to record the interpretations of these items, with one form used to record catalogued ceramic objects and another used to record any other catalogued object. **Figure 3** lists all of these various categories of paper record and it also shows the relationships that link each record type. These relationships were accommodated by the addition of unique IDs for each paper record, e.g. Context 1, Photo 2009-02-0001. While most of these record types will be familiar to the working archaeologist, some will not and it is important to note that no one archaeological project's record structure ever matches another's identically (Ross et al. 2013, 102).

In terms of numbers, the amount of pro forma records created at Priniatikos Pyrgos is not enormous[4] and it would be difficult to argue that it qualifies for the status of Big Data on these terms alone. However, this paper would argue that the definition of Big Data is as much based on the approaches used to interrogate it, as it is on its scale. Data, which is designed to be consumed primarily by machines, is Big Data.

Digitizing the data – phase 1

From as early as 2005 it was understood that the project's paper record would need to be digitized and a FileMaker Pro database system was selected to satisfy this need. FileMaker is a database system that is used by many archaeological

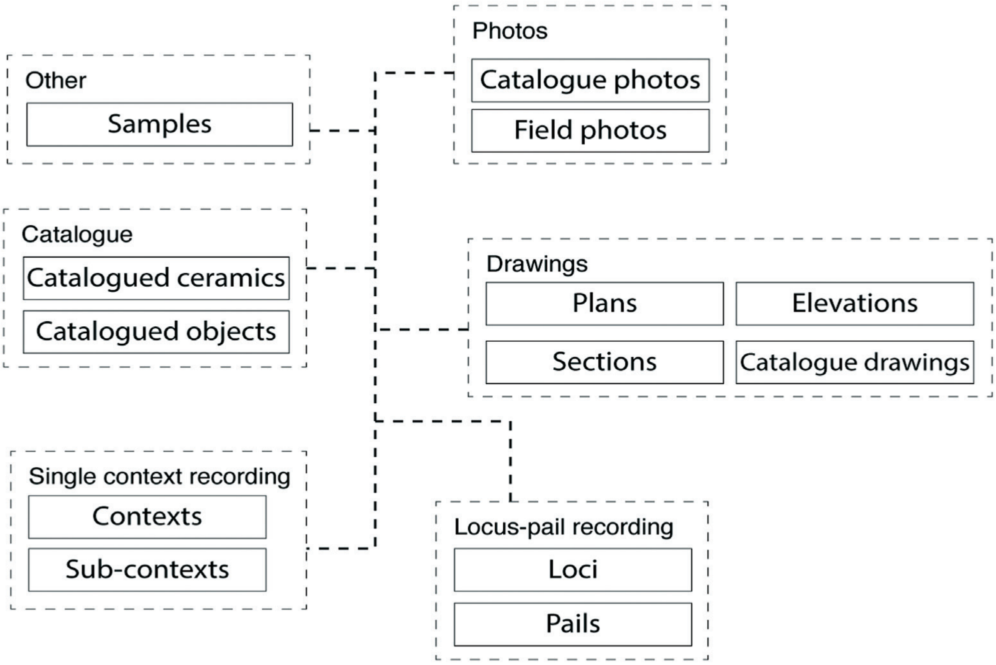


Fig. 3 The types of paper record created by the Priniatikos Pyrgos project.

projects (Motz & Carrier 2012; Wallrodt 2012). It has a number of advantages. It is graphical. It can manage multiple data types and it is collaborative. However, it also suffers from the fact that it is proprietary software and, therefore, restricted in terms of its extensibility. It is also costly and this can be a limiting factor for many archaeological projects. Over the 2009 season the database was redesigned and the process of importing the backlog of paper information began in earnest and has continued to the present day.

This digitisation process resulted in the creation of a FileMaker database that represents an almost complete digital reflection of the corpus of paper records described above. While FileMaker had certainly fulfilled its function in the sense that it allowed the project to install a digital database platform quickly, it soon became apparent that the system's lack of software and data openness was going to be a serious problem going forward. The project needed to be able to publish its raw data to the World Wide Web to allow for further study and as part of a more general dissemination strategy. It was also felt that the FileMaker system was too isolated from other archaeological datasets. For both of these reasons, it was decided to look for alternatives and it was out of this conversation that the linkedarc.net project came into being.

On Open Data, Linked Data and their role in archaeology

Linked Data and Open Data are similar concepts but it would be a mistake to consider them analogous (Kitchin 2014, 49). Open Data is a sub category of Open Access. Whereas Open Access in a general sense advocates the opening up of information to a wider audience, Open Data specifically targets the freeing up of raw data. Raw data in the Open Data model refers to the 'empirical' or 'objective' information on which interpretation is built (Rosenberg 2013, 18) and examples include CSV data, text files and relational databases. The perceived benefit of following Open Data practice to archaeologists is that by not only publishing an interpretation but also its supporting data, the knowledge creation process becomes more transparent and reflexive, and its underlying data more reusable.

In principal, one might imagine that there is a lot to gain from this shift in practice but the transition to Open Data has been anything but straightforward (Mauthner & Parry 2013). Change is a troubling concept in any walk of life and this is certainly the case within the academy in which traditionally a researcher's raw data has been viewed as a personal asset and one that is not without its value (Kitchin 2014, 41). Persuading academics to become part of this Open Data revolution, particularly in the archaeological field, has proved difficult. It has been argued by Kansa (2012) that given the inherently destructive nature of the archaeological process, archaeologists are ethically bound to make their data open and given that such a large proportion of archaeological investigation is funded by the public purse (Schadla-Hall 1999), it is difficult to argue against this position.

Open Data in itself is just an idea, a belief or an aspiration. It needs an implementation to be realised. As a concept, it is fairly simple: make your data open, encourage participation and collaboration around that data (Kitchin 2014, 48). A question that immediately presents itself is how would one go about measuring data openness? For example, if one were to populate an Excel file with the raw data that supports a particular paper and to upload this file to a website, would this make the paper Open Data compliant? Perhaps, but if the file's information was ordered or structured in a way that made sense only to the author and its Internet address went unadvertised, then its status as an Open Data resource would be largely academic. We can, therefore, surmise that Open Data is about more than just making data available. It must also be structured in some way. While the data creator is free to select whichever structure they deem suitable, they must then also make the form of this structure available to the data consumer. This is what is meant when one talks of dataset transparency.

When Tim Berners-Lee outlined his 4 principles of Linked Data (Berners-Lee 2006), he was attempting to solve this problem of investing structure in Open Data so as to maximise its potential for re-use. Structure comes in many forms and for Open Data it is applied at various levels of the conceptual and technological hierarchy or stack. We will discuss higher level conceptual structures below when we talk about data ontologies but in Berners-Lee's 2006 paper he was targeting the more low level data representation and data transport structures or protocols. Essentially, Berners-Lee was trying to uncover the most efficient way of allowing users to publish data online and then to have that data interrogated, read and understood by a second user.

For his first principle, he stated that the names of every digital 'thing' should be in the form of a Uniform Resource Identifier or URI. We are familiar with URLs, as these are the addresses that we use to locate resource on the document web. URLs are a subset of URIs; all URLs are URIs but not all URIs are URLs. The two have the same form but are differentiated by the fact that a URL must be dereferenceable. This means that accessing it using a World Wide Web browser will return some resource. URIs that are not also URLs do not point to any web resource.

Secondly, Berners-Lee said that HTTP should be the protocol used to handle the movement of one data resource from one point to another. HTTP is a simple request-response based protocol and is it noteworthy for the fact that it has powered the movement of resources on the document web for over twenty years. The power of Berners-Lee 2006 Linked Data manifesto is in its appropriation of existing technologies and practices. The Linked Data model demands no change to the underlying transport protocol. Linked Data resources will traverse the Web of Data as any other web resource, such as a HTML page or an image.

The third principle of Linked Data stipulates that standards be used to structure the data's representation and the way that a client interrogates that data. The principal standard that Berners-Lee advocates for the structuring of the data's representation is known as Resource Description Framework or RDF. RDF is built upon the idea that any complex element of information can be broken down into a network or graph of related subject and object pairings that are linked by what are known as predicates (Antoniou 2012, chap. 2). These 'triples' are linkable across a HTTP network such as the World Wide Web because their constituent components can be represented using URIs. Chains of triples form graphs of Linked Data **(Figure 4)**.

Another standard advocated by Berners-Lee relates to the querying of Open Data. Simple Protocol and Query Language or SPARQL (W3C 2013) was designed to allow for the querying of RDF triple data and is highlighted by Berners-Lee as a way of empowering users to sort through vast quantities of data in order to find exactly the data that they are looking for. SPARQL's syntax is designed with the triple in mind. Users ask the SPARQL engine to match triple patterns that they provide. These patterns can be chained together to construct highly complex queries, if required.

Finally, Berners-Lee states that Linked Data datasets should link to other Linked Data datasets. For example, an archaeological dataset that documents the findings of a Bronze Age site in the UK might reference a centralised catalogue of settlement enclosure types. As it happens, it is at this final hurdle towards total Linked Data compliance that most aspiring datasets fall down.

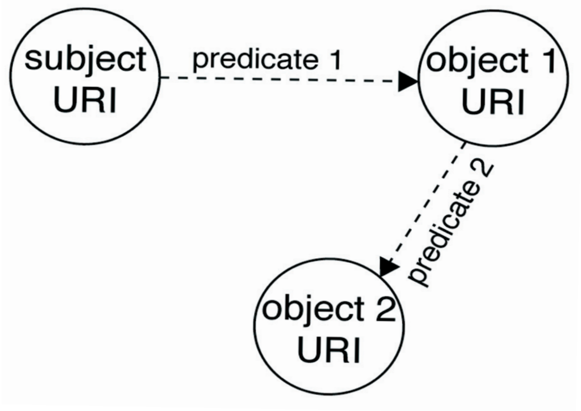


Fig. 4 A chain of RDF triples.

For the purposes of this document, we will refer to data resources, which comply with Open Data and Linked Data practices as being Linked Open Data resources.

Data ontologies, the Semantic Web and controlled vocabularies

We mentioned above how data can be structured at various different levels. This requirement certainly adds work to the data creator but it is absolutely necessary if the intent is to make the information available to as wide a user group as is possible. We will now consider the role that ontologies play in the conceptual structuring of archaeological information and as they are related, we will also introduce the subject of controlled vocabularies or thesauri.

The ontology

Data ontologies are crucially important to the practical workings of the global collection of Linked Open Data resources that we refer to as the Web of Data. Making data accessible and linkable is only one part of the Open Data publication process. The data creator also needs to consider how their information is to be structured conceptually, so that it is semantically coherent from the creator's perspective but also from the consumer's point of view. More often than not, when this topic of data about data is mentioned, the first term that springs to mind is metadata and while metadata is certainly a part of this, the conceptual structuration of a dataset runs deeper than simply applying tags to information. Ontologies provide a means by which an agent, human or machine, can discover how a resource maps onto a conceptual framework, what properties are associated with each of its classes of information and how these classes relate to other internal and external classes (Ceusters & Smith 2011, 123).

An ontology is typically understood from one of two somewhat different perspectives: philosophical or in the sense that Computer Science understands it.[5] In the latter reading, an ontology implies a formal framework in which some form of knowledge can be represented or, as Gruber (1993) puts it, it is "a specification of a conceptualization". The definition of the philosophical ontology is unsurprisingly broader, potentially more complex and without the universal acceptance that its CS equivalent enjoys. Both of these understandings share, however, at their core the principle that classification aids the flow of knowledge from one agent to the next (Hendler 2011, 127).

For the purposes of this paper, we will follow the CS definition, as this best suits our needs. Let us consider an example of an archaeological ontology. In our scenario, we need to create a model that can represent the act of depositing an artefact in the archaeological record. We first need to define a number of classes that we can assign properties to. As the central concept in our narrative is an artefact, we should add a class called Artefact.[6] The reality that we want to model involves agency; the artefact is deposited. So we will need a class that models this deposition event. We can call this DepositonEvent. Finally, the destination of the artefact is in the archaeological record. This is a bit vague, so let us narrow it down slightly and talk of archaeological contexts in which the artefact becomes deposited. As such, our final class will be called Context.

Next, we need to define the properties of these various classes. The Artefact class might have a type property and a set of properties for its dimensions. It also needs to property that links it to the DepositionEvent class. We might call this depositedBy. The DepositonEvent class is different to the Artefact class in the sense that it does not model a material reality. Instead, it models an event, which is located at some point in time and space. It also needs to include links to the destination for whatever artefact is being acted upon by this event. As such, we could give it the following properties: tookPlaceAtTime, tookPlaceAtPlace and destination. Finally, the Context class might have a property that locates it in space. These classes, their properties and relationships are represented schematically in **Figure 5**.

An ontology in itself is conceptual in form. It needs to be realised through the creation of data, which it gives structure to. This realisation happens by the creation of what are called instances of the ontology's classes. In our example, we might instantiate the Artefact class and populate its fields with information relating to a dagger artefact. In order to represent the deposition of this dagger, we would instantiate the DepositionEvent class and the Context class. The DepositionEvent instance would contain information about the specific event in time and space in which the dagger became deposited and the Context instance would contain information about the dagger's archaeological context. This instantiation of our ontology is represented in **Figure 6**.

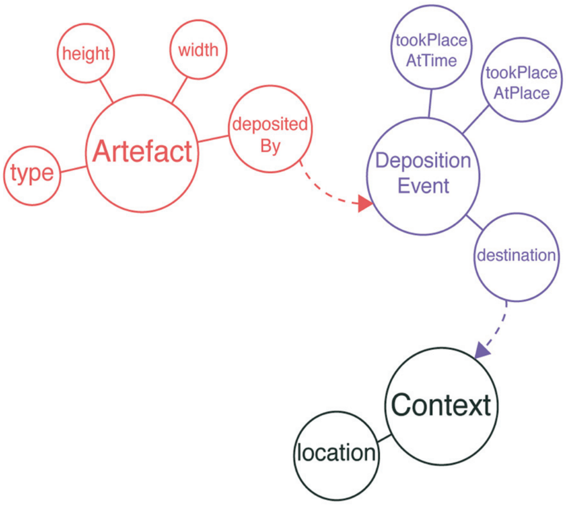


Fig. 5 The artefact deposition ontology.

Now when clients wish to access our dataset, they can query its ontology first in order to understand the structure of the dataset's contents.

The Semantic Web

The study of semantics is the study of meaning (Blackburn 2008) and in our previous discussion on ontologies, we were essentially talking about injecting meaning into data. The Semantic Web is a global and public web of Linked Open Data in which each individual data node sits within

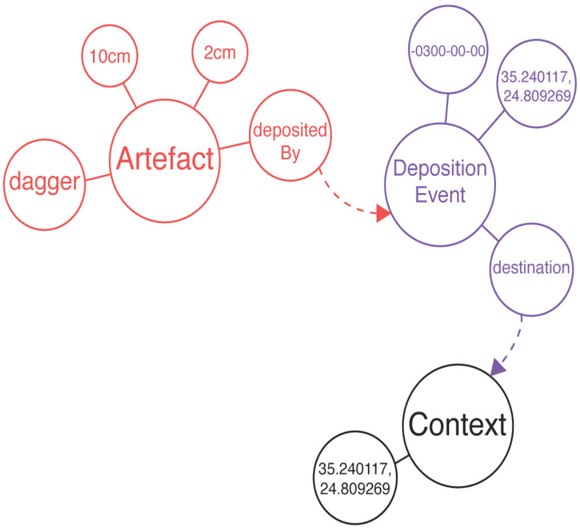


Fig. 6 Instances of the artefact deposition ontology.

its own conceptual structure (Hendler 2001). The Semantic Web cannot contain information, which is contextless or devoid of meaning. While there is some confusion about the relationships between the terms Open Data, Linked Data, the Web of Data and the Semantic Web (Heath 2009), for the purposes of this paper we will view Open Data and Linked Data as being compatible but not dependant approaches to data representation and the Web of Data and the Semantic Web as synonymous networks that are built upon the principles of Open Data and Linked Data practice.

Finally, this paper would also contend that the Semantic Web implies a certain degree of non-human client activity. It perhaps goes too far to label these activities Artificial Intelligence but certainly the Semantic Web, as first imagined by Berners-Lee (Berners-Lee, Hendler, and Lassila 2001), Bizer, Heath (Bizer et al. 2008) and others, involves a degree of automated machine-driven processing, which might be as simple as an online weather data aggregator or as complex as the type of problems being tackled currently by the IBM Watson machine (Ferrucci et al. 2010).

The controlled vocabulary

Complementing the data ontology is the controlled vocabulary. While ontologies provide an overarching conceptual framework for a dataset, controlled vocabularies apply limits to the types of values that class properties can contain. Essentially, vocabularies are domain-specific ranges of values that can be applied to specific class properties (Bojārs et al. 2008).

Take for example, the triple shown in **Figure 7**. The subject is an instance of the Context class referred to in our ontology example. The predicate is rdf:type, which has the meaning 'is of type' (Brickley, Guha, & McBride 2014) and the object contains a type value. The designer of the dataset schema could decide to allow string values to serve as objects for the rdf:type predicate. This is a simple solution and very flexible but at the same time this flexibility undermines the object's indexability. For example, a user might create a Context instance with the type 'wall'. A second user might create another instance with the type 'low wal'. Note that in the second instance, the user has mistyped the entry as 'wal' and he has also added the adjective 'low'. Should these two instances be considered the same? And how do you deal with the misspellings?

By using controlled vocabularies in an RDF system, it is possible to limit the objects that a user can enter to a particular set of URI values, thereby avoiding the problems and potential ambiguities that the example above highlights. This ultimately makes the data more structured, indexable and comprehensible to external agents. On the other hand, using controlled vocabularies does present the very real danger of ending up with overly deterministic (the values that a user can enter are predetermined by the dataset's designer), essentialised (not allowing for free text entry inevitably involves a simplification of the conceptual values that are being represented) and normalised (the objects affected sit within a predefined system of values) datasets (Beall 2010).

Simple Knowledge Organisation System or SKOS has become a very popular framework for delivering controlled vocabularies (Isaac & Summers 2009). The model allows schema designers to build hierarchies of concepts, which can be used to populate controlled vocabulary lists. SKOS allows for the mapping of concepts between one vocabulary and another using the skos:exactMatch predicate and this simple feature allows for the connecting of datasets, which employ different vocabulary lists. Archaeological and related cultural heritage projects, which have relied so heavily on controlled vocabularies in the past, have embraced the SKOS standard. Two initiatives worthy of special note are the Seneschal project, which is making the controlled vocabularies of English, Scottish and Welsh cultural heritage institutions available to the wider archaeological community using SKOS (Charno 2013; May, Binding,

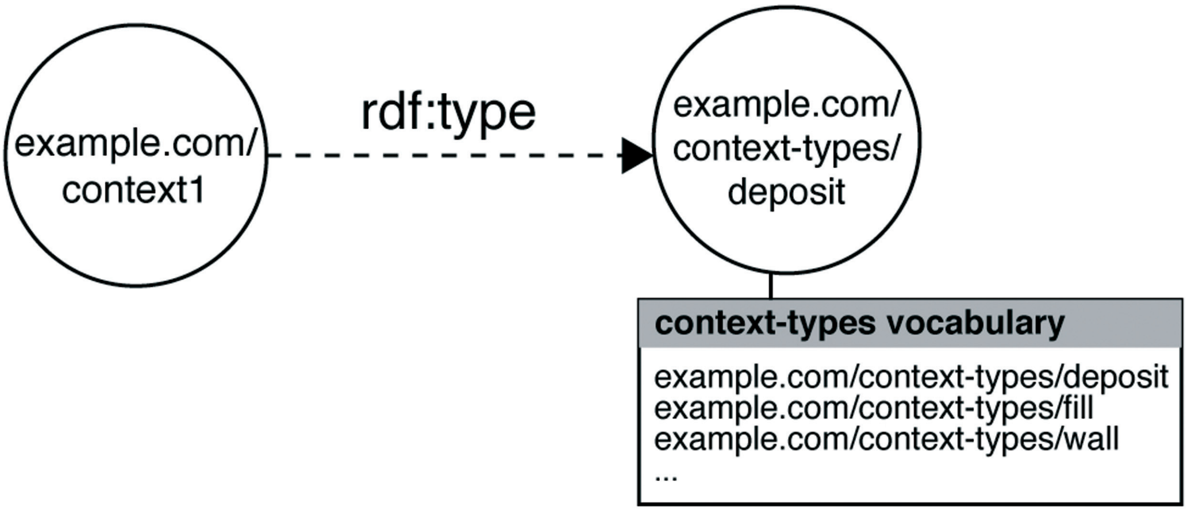

example.com/
context1
rdf:type
example.com/
context-types/
deposit
context-types vocabulary
example.com/context-types/deposit
example.com/context-types/fill
example.com/context-types/wall
...

Fig. 7 A triple‘s object containing vocabulary entries..

& Tudhope 2015) and the Getty Vocabularies as Linked Open Data project (Harpring 2014), which is doing the same for the Getty’s AAT, CONA, TGN and ULAN vocabularies.

Mapping the Priniatikos Pyrgos dataset onto the CIDOC CRM

While building a custom ontology and providing this alongside a set of published data that it models allows the dataset to be understood by external agents, in practice it is much more beneficial to use pre-existing ontologies with active user-bases. This also introduces a degree of data determinism but it also makes the data more comprehensible due to the model’s familiarity. We will now discuss the process of mapping the Priniatikos Pyrgos data onto one of the most popular cultural heritage ontologies currently in use: the CIDOC CRM.

The CIDOC CRM and its English Heritage extension

The CIDOC CRM has come to dominate the ways in which cultural heritage professionals structure their digital data (Oldman & Rahtz 2014)[7]. While it is not prescribed that the CRM be conceptualised as Linked Open Data or serialised using RDF, in practice RDF is used for most CRM implementations and it is also the practice that we will follow here. The CRM was designed primarily as a means of handling museum and archive material but its core set of classes and predicates can be extended and English Heritage have provided an extension that structures archaeological data and it is onto this ontology that the Priniatikos Pyrgos data is mapped (Binding, May, & Tudhope 2008b)[8].

The core CRM is built around the structuring unit of the event (Binding, May & Tudhope 2008a, sec. 2.5). This model states that material objects are transformed when they move from one context to the next. Take for instance, the scenario described above to explain the data ontology. The dagger that becomes deposited in an archaeological context adopts different meanings as it is affected by one event after another. Perhaps after its deposition, the dagger comes to be excavated by an archaeologist and is removed from its context. This event transforms it into an archaeological find. It might then be transported to an archive, where it becomes a collection record. As the chain of events unfolds, the meaning that we associate with the dagger changes as well.

The mapping process

Theoretically, mapping a source dataset onto an ontology is a relatively simple process. Essentially, you decide which source fields align best with the target ontology fields. In practice, however, this process is much more complex and time-consuming than that. The following guide synthesises and necessarily simplifies this process as it was applied for the Priniatikos Pyrgos dataset.

Step 1 – Understanding the source and target ontologies
While the original Priniatikos Pyrgos dataset was not explicitly structured using any one ontology, it was nonetheless built around a framework that was implicitly understood by its designers and users within the project. Essentially, this structure mirrored the SC recording method as described by MoLAS. While the target ontology, the CRM-EH extension, was designed to structure SC data as well, the manner of its interpretation is different to that of the Priniatikos Pyrgos dataset.

Step 2 – CRM-EH mapping strategy
As such, the Priniatikos Pyrgos FileMaker structure did not map directly on to the CRM-EH with a one-to-one correlation. There were of course situations in which this did occur, for example, the 'context name' property of the context FileMaker table mapped directly onto the CRM predicate P87_is_identified_by. However, other aspects of the mapping presented greater challenges. One of the primary reasons for this complexity derived from the event-based structure of the CRM. Each context in the CRM-EH is represented using multiple class instances, with each instance either representing a material object, such as the material excavated from a context (e.g. EHE0008_ContextStuff), or an event (e.g. EHE2001_ContextExcavationEvent, EHE2020_ContextFindMeasurementEvent) taking place during the context's creation or excavation and recording.

The FileMaker database contained a number of different tables each holding data relating to a particular type of project information as we have discussed. Most of these types of data, whether they represent photograph images, catalogue or sample records or plans and section drawings could be mapped onto one or other CRM or CRM-EH class and in doing so become part of the mesh of information surrounding the central EHE0007_Context class. The CRM-EH model is immensely broad, which is a reflection of the comprehensiveness of its modelling of the SC method, from context creation to excavation to study and finally to dissemination. The mapping of the Priniatikos Pyrgos data did not make use of all of these classes. In total, 23 of the 138 CRM-EH classes[9] were used to map the Priniatikos Pyrgos data and a diagram explaining a sample of the relationships used is reproduced in **Figure 8**.

Further difficulties arose with the mapping of conceptual entities that fell outside of the SC model. For example, the sub-context concept, which is unique to the Priniatikos Pyrgos Project, necessitated the creation of a custom ontology to augment the classes and predicates provided for by the core CRM and CRM-EH.[10]

Step 3 – Cleaning the data
Data cleaning is defined as the detection and removal of errors and inconsistencies within a dataset (Rahm & Do 2000). In other words, it is the process of conforming a dataset to a particular array of

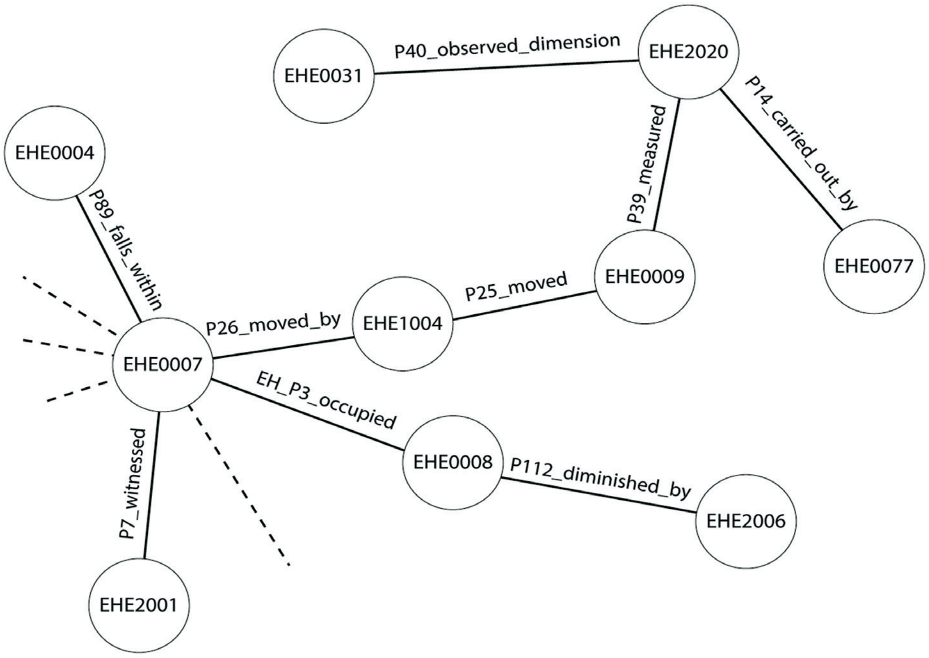


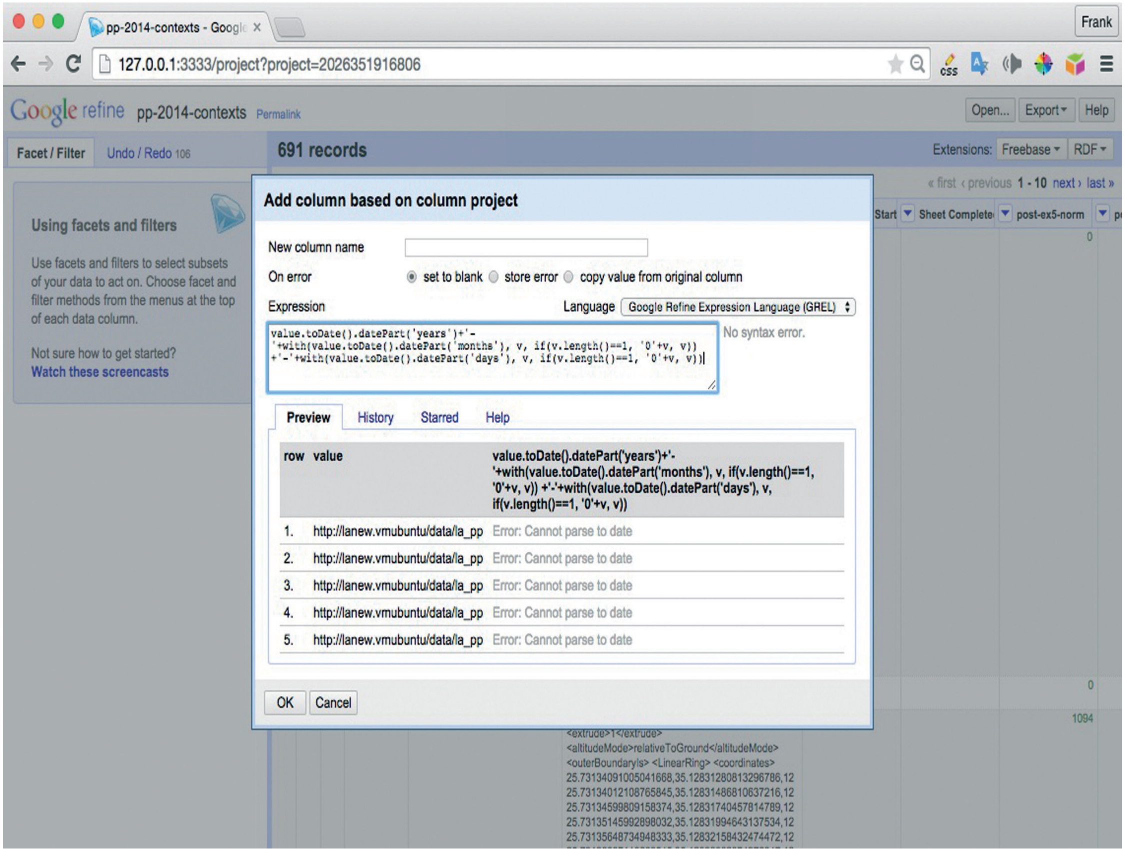
Fig. 8 A sample of the CRM-EH classes used for the mapping of the Priniatikos Pyrgos dataset.

formatting rules, in order to reduce redundancy, which ultimately makes it more indexable and valuable. The Priniatikos Pyrgos FileMaker data required a lot of data cleaning before it could be mapped onto the CRM-EH model. The majority of the many hundreds of fields used in the original FileMaker design required some degree of correction, adding particular value to fields, which contained date, period, place name and type information. Ideally, the FileMaker system would have been designed from the beginning to restrict user input for certain fields by limiting them to selected controlled vocabularies. This process was not put in place for the creation of the Priniatikos Pyrgos data, however, and as a result most of the fields could be entered as free-text and allowing for free-text entries will almost certainly introduce errors in the form of misspellings, inconsistent use of vocabulary and variable formatting for fields such as dates.

Google Refine[11] provides a powerful platform on which to clean data **(Figure 9)**. It can comfortably handle many thousands of rows of CSV[12] table data, quickly checking for patterns in potentially messy data ranges. It provides a reasonably flexible scripting function using the Google Refine Expression Language (GREL), which allows for more sophisticated parsing and fixing of potential errors. Google Refine also provides advanced features such as an ability to make HTTP web service requests, which allows the user to mine data from external web resources using project data as input parameters for these calls.[13]

Step 4 – Implementing the mappings

Google Refine's baseline functionality can be extended using extensions. The RDF Refine extension for Google Refine (DERI 2014) adds RDF mapping and export functionality. This extension proved extremely useful in putting the mapping plans discussed in step 2 into action **(Figure 10)**. For the types of complex multi-level mappings needed to create the CRM-EH data, the RDF Refine extension proved an invaluable resource. The extension allows for the exporting of the CSV input data to RDF/XML or Turtle serialised files, following the rules specified by the mapping 'skeleton'.



Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

Add column based on column project

New column name

On error ● set to blank ○ store error ○ copy value from original column

Expression

Language: Google Refine Expression Language (GREL)

```
value.toDate().datePart('years')+'-
'+with(value.toDate().datePart('months'), v, if(v.length()==1, '0'+v, v))
+'-'+with(value.toDate().datePart('days'), v, if(v.length()==1, '0'+v, v))
```

No syntax error.

Preview | History | Starred | Help

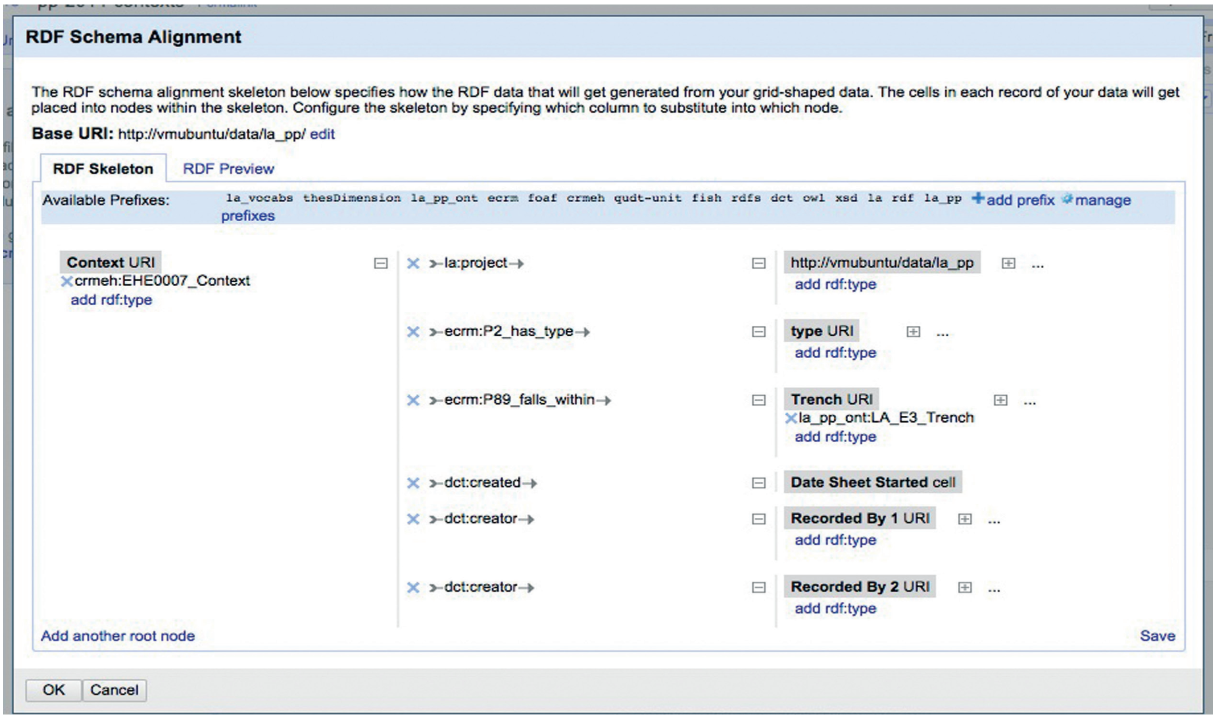
row	value	value.toDate().datePart('years')+'-'+with(value.toDate().datePart('months'), v, if(v.length()==1, '0'+v, v)) +'-'+with(value.toDate().datePart('days'), v, if(v.length()==1, '0'+v, v))
1.	http://lanew.vmubuntu/data/la_pp	Error: Cannot parse to date
2.	http://lanew.vmubuntu/data/la_pp	Error: Cannot parse to date
3.	http://lanew.vmubuntu/data/la_pp	Error: Cannot parse to date
4.	http://lanew.vmubuntu/data/la_pp	Error: Cannot parse to date
5.	http://lanew.vmubuntu/data/la_pp	Error: Cannot parse to date

```
<extrude>1</extrude>
<altitudeMode>relativeToGround</altitudeMode>
<outerBoundaryIs> <LinearRing> <coordinates>
25.73134091005041668,35.12831280813296786,12
25.73134012108765845,35.12831486810637216,12
25.73134599809158374,35.12831740457814789,12
25.73135145992898032,35.12831994643137534,12
25.73135648734948333,35.12832158432474472,12
```

Fig. 9 Data cleaning using Google Refine.

Step 5 – Hosting the RDF data
Fig. 11: The linkedarc.net web app interface displaying a Priniatikos Pyrgos context record.

The data produced by the RDF Refine extension can then be imported directly into an RDF triplestore such as Apache Jena. Apache Jena provides the baseline RDF triplestore functionality for the linkedarc.net project. It is a lightweight and easily configurable RDF hosting solution, which provides both programmatic (using a Java API) and SPARQL 1.1 (via the Fuseki engine) access to the triple data. linkedarc.net is composed of a Python backend and a JavaScript + HTML web app[14] front end **(Figure 11)**.



RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://vmubuntu/data/la_pp/ edit

Available Prefixes: la_vocabs thesDimension la_pp_ont ecrm foaf crmeh qudt-unit fish rdfs dct owl xsd la rdf la_pp

Context URI crmeh:EHE0007_Context add rdf:type	la:project	http://vmubuntu/data/la_pp add rdf:type
	ecrm:P2_has_type	type URI add rdf:type
	ecrm:P89_falls_within	Trench URI la_pp_ont:LA_E3_Trench add rdf:type
	dct:created	Date Sheet Started cell
	dct:creator	Recorded By 1 URI add rdf:type
	dct:creator	Recorded By 2 URI add rdf:type

Fig. 10 Mapping the CSV Priniatikos Pyrgos data to an RDF ontology.

Outputs and conclusions drawn

This paper has considered the process of mapping the data of the Priniatikos Pyrgos archaeological project to the English Heritage CIDOC CRM extension so that it can be published as Linked Open Data to the Semantic Web. Linked Open Data and Semantic Web technologies offer many advantages to archaeological adopters. They promote data visibility, data reuse and the linking of data across different domains. The use of public ontologies increases this value as one dataset can be interrogated alongside other compatible datasets. Ultimately, these technologies allow for the building of a global mesh of archaeological Big Data, in which no one dataset exists as an island. As we move towards that endpoint, working archaeologists will progressively gain from a wider access to bigger and more contextual information sources.

At the time of writing, linkedarc.net holds 5,975,587 triples relating to the Priniatikos Pyrgos material. This RDF data is available via a number of different linkedarc.net data interfaces such as web services and DOI HTTP requests. The linkedarc.net web app also provides an interface to the data, which largely follows the user interfaces of applications such as FileMaker Pro and of data-driven archaeological sites such as Open Context (S. W. Kansa et al. 2012). And for more advanced access to the data, a SPARQL endpoint is also available. It has proved extremely challenging to transform the FileMaker relational data into a Linked Open Data representation that conforms to the CRM-EH model. The data cleaning in particular proved time-consuming and laborious.

These tasks are, however, for the most part once-off costs. Using the history feature of Google Refine, it is possible to reconstruct the many hundreds or even thousands of data cleaning

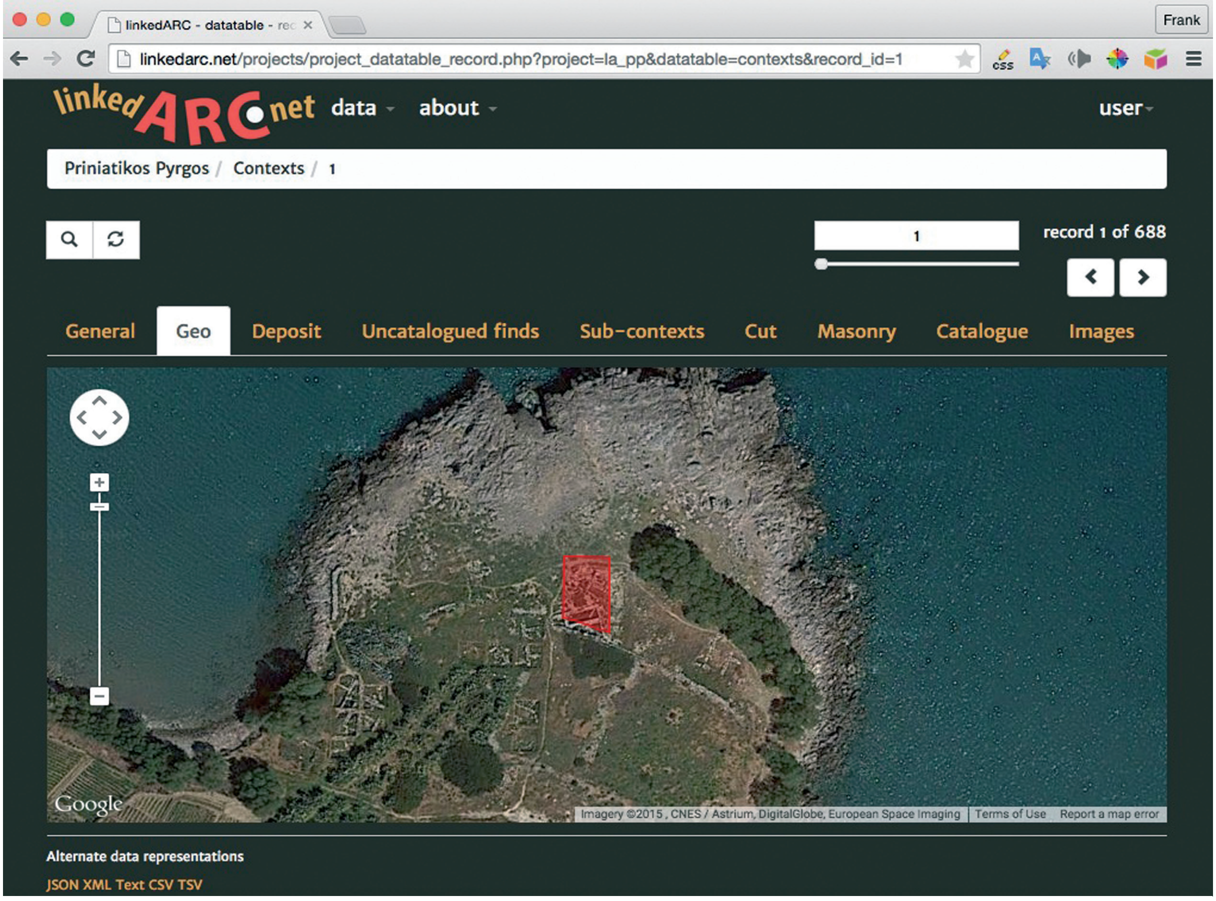

linkedARC - datatable - rec
Frank
linkedarc.net/projects/project_datatable_record.php?project=la_pp&datatable=contexts&record_id=1
linkedARC.net
data
about
user
Priniatikos Pyrgos / Contexts / 1
1
record 1 of 688
General
Geo
Deposit
Uncatalogued finds
Sub-contexts
Cut
Masonry
Catalogue
Images
Google
Imagery ©2015, CNES / Astrium, DigitalGlobe, European Space Imaging
Terms of Use
Report a map error
Alternate data representations
JSON XML Text CSV TSV

Fig. 11 HTML web app front end.

and mapping steps taken in carrying out this transformation. Also, by its nature, data mapping tends to be more time-involved at the beginning as a model is figured out. As the project proceeds and the level of knowledge about the ontologies increases, the amount of time it takes to carry out tasks tends to fall. Therefore, archaeological projects, which choose to adopt Linked Open Data and open ontology-based data strategies, face an initial struggle but these demands reduce over time as experience is gained.

Challenges also present themselves for the consumer of CRM-modelled data. The CRM and its CRM-EH extensions create data structures, which are multi-tiered and complex. It is difficult to work out the ideal way of mapping data onto these models (Carver 2012; E. C. Kansa 2014) but similarly it can be just as challenging for consumers to work out how to get at the data that they are really interested in. SPARQL is a powerful querying interface but without a detailed knowledge of the underlying ontology, it can be difficult for users to know exactly what to ask for. The linkedarc.net project has recently employed the use of autocomplete functionality to address this very problem by providing the user with suggestions as they construct SPARQL queries.[15] More research needs to be carried out in this area but early indications suggest that this is a promising field of investigation.

For the most part, the tools used to map archaeological data to RDF ontologies are not designed for this task specifically. Tools such as Google Refine or even Excel can be very useful when mapping RDF data but there are very few examples of tools that have the needs of the archaeologist solely in mind. The STELLAR project (Tudhope et al. 2011) is one notable exception to this rule but one must imagine that there is a significant amount of duplication of effort happening across the archaeological community as projects go about mapping their data to the CRM.

It is only recently[16] that the entire Priniatikos Pyrgos dataset has been published as Linked Open Data adhering to the CRM-EH model on linkedarc.net. Currently, the project members are being introduced to the new model through a series of usability test sessions. It is hoped that the effort put into the mapping process will deliver a return in terms of tangible interpretive gains for the project archaeologists and for other archaeologists looking to use the service but no assumptions have yet been made in this regard. Ultimately, as with most innovative technologies, success will be determined by popular adoption. To date, it is fair to say this Linked Open Data has yet to reach the tipping point of acceptance. The project outlined here is one example of an attempt to apply Linked Open Data and open ontology methods to the resolution of practical archaeological problems. Only time will tell whether the linkedarc.net project and others like it will be judged successful and of worth to the field of archaeological research or not.

Acknowledgements

The author would like to thank his supervisor Dr Christine Morris for her guidance and support throughout his PhD project. He would also like to acknowledge the support of the Digital Art and Humanities PhD programme, which is generously funded by the PRTLI.

Notes

[1] When compared to other Cretan archaeological sites (Driessen 2011, fig. 4).

[2] The exact location of this original trench or series of trenches remains unknown.

[3] Since the excavations have ended, the project has been engaged in a series of study seasons.

[4] By 2010, there were about 700 context records, 1700 subcontext, 350 loci, 1300 pails, over 5000 catalogue entries and almost 10000 photographs.

[5] Having said that, Roussey et al. (2011) shows that a number of other knowledge fields have their own particular take on the meaning of the term.

[6] It is a convention in ontology syntax to capitalise the first letter of classes and to make the first letter of a property lowercase. Classes and properties cannot have spaces, although underscores are commonly used. Words that follow the first word of a class or property will usually also begin with a capitalised letter.

[7] Alternatives exist, such as CHARM (Gonzalez-Perez et al. 2012) and ArchaeoML (Schloen 2001).

[8] Recently, the ARIADNE project launched a second CIDOC CRM archaeological extension (Cripps et al. 2014). While the English Heritage CRM extension was designed principally to model information derived from single context archaeological investigations in the UK, CRMarchaeo is intended to support all European archaeological data.

[9] The specific class used were EHE0007_Context, EHE2001_ContextExcavationEvent, EHE0077_ProjectTeamMember, EHE0098_ContextExcavationEventTimespan, EHE0008_ContextStuff, EHE2016_ContextStuffMeasurementEvent, EHE0054_ContextStuffMeasurement, EHE1001_ContextEvent, EHE2006_ContextSamplingEvent, EHE0018_ContextSample, EHE0012_ContextEventRecord, EHE1004_ContextFindDepositionEvent, EHE0009_ContextFind, EHE0030_ContextFindMaterial, EHE0053_ContextSampleType, EHE0088_SiteSubDivisionDepiction, EHE2020_ContextFindMeasurementEvent, EHE0031_ContextFindMeasurement, EHE1005_ContextFindUseEvent, EHE2010_DepictionEvent, EHE2002_ContextFindClassificationEvent, EHE1002_ContextFindProductionEvent and EHE0004_SiteSubDivision.

[10] This ontology is outlined in greater detail at http://linkedarc.net/ontologies/la.

[11] Now rebranded as OpenRefine http://openrefine.org

[12] The FileMaker data was exported as CSV before being imported into the Google Refine system.

[13] An example use of this would be to request the geocoordinates of a place name from the Google Maps Reverse Geocoding API (Google Developers 2015).

[14] http://linkedarc.net

[15] See examples at http://linkedarc.net/sparql

[16] January 2015.

Bibliography

Antoniou, G. (ed.) (2012). A Semantic Web Primer. 3rd ed. Cooperative Information Systems. Cambridge, Mass.: MIT Press.

Baudrillard, J. (1994). Simulacra and Simulation. The Body, in *Theory: Histories of Cultural Materialism.* Ann Arbor: University of Michigan Press.

Beall, J. (2010). Determinism and the Semantic Web. http://seealso.tumblr.com/post/2084693211/determinism-and-the-semantic-web [3.12.2014].

Berners-Lee, T. (2006). Linked Data - Design Issues. W3C. 27. http://www.w3.org/DesignIssues/LinkedData.html [16.11.2014].

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web. *Scientific American 284 (May)*, 34-43.

Betancourt, Ph. P. (2014). Priniatikos Pyrgos in 1912: The Last Foreign Archaeological Excavation in Independent Crete. In *Moloy, B. & Dockworth, Chl. (eds.). A Cretan Landscape through Time: Priniatikos Pyrgos and Environs*, pp. 8-14. Oxford: British Archaeological Reports.

Binding, C., May, K. & Tudhope, D. (2008a). Semantic Interoperability Issues from a Case Study in Archaeology. In Kollias, S. & Cousins, J. (eds.). *Semantic Interoperability in the European Digital Library*. Proceedings of the First International Workshop (SIEDL) 2008, Associated with the 5th European Semantic Web Conference, pp. 88-99.

Binding, C., May, K. & Tudhope, D. (2008b). Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM. In Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B. &

Lippincott, J. (eds.). *Research and Advanced Technology for Digital Libraries*, pp. 280-90. Lecture Notes in Computer Science 5173. Berlin: Springer. http://link.springer.com/chapter/10.1007/978-3-540-87599-4_30 [16.11.2014].

Bizer, Chr., Heath, T., Idehen, K. & Berners-Lee, T. (2008). Linked Data on the Web (LDOW2008). In Proceedings of the 17th International Conference on World Wide Web, 1265–66. WWW '08. New York, NY, USA: ACM. doi:10.1145/1367497.1367760 [16.11.2014].

Blackburn, S. (2008). *The Oxford Dictionary of Philosophy*. Oxford: Oxford University Press.

Bojārs, U., Breslin, J. G., Finn, A. & Decker, S. (2008). Using the Semantic Web for Linking and Reusing Data across Web 2.0 Communities. In *Web Semantics: Science, Services and Agents on the World Wide Web, Semantic Web and Web 2.0, 6 (1)*: 21–28. doi:10.1016/j.websem.2007.11.010 [16.11.2014].

Brickley, D., Guha, R. V. & McBride, B. (eds.) (2014). RDF Schema 1.1. http://www.w3.org/TR/rdf-schema/ [16.1.2014].

Carver, G. (2012). "ArcheoInf, the CIDOC-CRM and STELLAR: Workflow, Bottlenecks, and Where Do We Go from Here?" In *CAA2012. Proceedings of the 40th Conference in Computer Applications and Quantitative Methods in Archaeology*, Southampton, United Kingdom, 26-30 March 2012, pp. 498-508. Southampton. http://www.ariadne-infrastructure.eu/ita/content/download/3763/21702/file/Geoff%20Carver-STELLAR[2].pdf [16.11.2014].

Ceusters, W., & Smith, B. (2011). Switching Partners: Dancing with the Ontological Engineers. In Bartscherer, Th. & Coover, R. (eds.). *Switching Codes: Thinking Through Digital Technology in the Humanities and the Arts*, pp. 103-24. Chicago: University of Chicago Press.

Charno, M. (2013). SENESCHAL Vocabularies: Value to the ADS. Archaeological Data Service. July 29. http://archaeologydataservice.ac.uk/blog/2013/07/seneschal-value-to-the-ads/ [16.11.2014].

Cripps, P., Doerr, M., Hermon, S., Hiebel, G., Kritsotaki, A., Masur, A., May, K., Schmidle, W., Theodoridou, M. & Tsiafaki, D. (2014). CRMarchaeo: The Excavation Model. FORTH. http://www.ics.forth.gr/isl/CRMext/CRMarchaeo/docs/CRMarchaeo1.2.1.pdf [16.11.2014].

DERI (16.11.2014). RDF Refine. March 6. http://refine.deri.ie/.

Driessen, J. (2011). History and Hierarchy. In Branigan, K. (ed.). *Urbanism in the Aegean Bronze Age*, pp. 51–71. Sheffield: A&C Black. https://www.academia.edu/2015884/History_and_Hierarchy [16.11.2014].

Ellis, St. & Wallrodt, J. (2011). The Paper-Less Project: The Use of iPads in the Excavations at Pompeii. Paper presented at the 39th Conference on Computer Applications and Quantitative Methods in Archaeology, Beijing, China, April 13th.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., et al. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine 31 (3)*: 59-79. doi:10.1609/aimag.v31i3.2303

Gonzalez-Perez, C., Martín-Rodilla, P., Parcero-Oubiña, C., Fábrega-Álvarez, P. & Güimil-Fariña, A. (2012). Extending an Abstract Reference Model for Transdisciplinary Work in Cultural Heritage. In Dodero, J. M., Palomo-Duarte, M. & Karampiperis, P. (eds.). *Metadata and Semantics Research*, pp. 190-201. Communications in Computer and Information Science 343. Berlin: Springer. http://link.springer.com/chapter/10.1007/978-3-642-35233-1_20 [16.11.2014].

Google Developers (2015). The Google Geocoding API. Google Developers. April 28. https://developers.google.com/maps/documentation/geocoding/ [16.11.2014].

Gruber, Th. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition 5 (2)*: 199-220.

Harpring, P. (2014). The Getty Vocabularies and Linked Open Data: Introduction and Editorial Perspective. Revised September 2014. https://www.getty.edu/research/tools/vocabularies/Linked_Data_Getty_Vocabularies.pdf [16.11.2014].

Hayden, B. J. (2014). Priniatikos Pyrgos and Its Territory: Results of Survey and Excavation. In Molloy, B. & Duckworth, Chl. (eds.). *A Cretan Landscape through Time: Priniatikos Pyrgos and Environs*, pp. 15-22. Oxford: British Archaeological Reports.

Heath, T. (2009). Linked Data? Web of Data? Semantic Web? WTF? Tom Heath's Displacement Activities. March 2. http://tomheath.com/blog/2009/03/linked-data-web-of-data-semantic-web-wtf/ [16.11.2014].

Hendler, J. (2001). Agents and the Semantic Web. *IEEE Intelligent Systems*.

Hendler, J. (2011). The Semantic Web from the Bottom up. In Bartscherer, Th. & Coover, R. (eds.). *Switching Codes: Thinking Through Digital Technology in the Humanities and the Arts*, pp. 125-39. Chicago: University Of Chicago Press.

Isaac, A. & Summers, E. (2009). SKOS Simple Knowledge Organization System Primer. W3C Working Group Note. August 18. http://www.w3.org/TR/skos-primer/ [18.11.2014].

Kansa, E. C. (2012). Openness and Archaeology's Information Ecosystem. *World Archaeology 44 (4)*: 498-520.

Kansa, E. (2014). Open Context and Linked Data. *ISAW Papers. Vol. 7.22*. doi:10.1145/1141753.1141782

Kansa, S. W., Kansa, E. C., Whitaker, J. & Ward, J. (2012). Concepts behind Open Context. Open Context. http://opencontext.org/about/concepts [16.11.2014].

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. 1 edition. Thousand Oaks: SAGE.

Mauthner, N. S. & Parry, O. (2013). Open Access Digital Data Sharing: Principles, Policies and Practices. *Social Epistemology 27 (1)*: 47-67. doi:10.1080/02691728.2012.760663

May, K., Binding, C. & Tudhope, D. (2015). Barriers and Opportunities for Linked Open Data Use in Archaeology and Cultural Heritage. Archäologische Informationen 38, online 4. February 2015: http://www.dguf.de/fileadmin/AI/ArchInf-EV_May-etal.pdf

Molloy, B., Day, J., Klontza-Jaklova, V. & Duckworth, C. (2014). Of What Is Past, or Passing, or to Come: Five Thousand Years of Social, Technological and Environmental Transformations at Priniatikos Pyrgos." In Molloy, B. & Duckworth, Chl. (eds). *A Cretan Landscape through Time : Priniatikos Pyrgos and Environs*, pp. 1-7. Oxford: British Archaeological Reports.

Morgan, C. L. (2010). Where Is Single Context Archaeology? Middle Savagery. February 23. http://middlesavagery.wordpress.com/2010/02/23/where-is-single-context-archaeology/ [16.11.2014].

Motz, Chr. F. & Carrier, S. C. (2012). Paperless Recording at the Sangro Valley Project. In *Archaeology in the Digital Era*, pp. 25–30. Southampton: The University of Chicago Press. https://www.academia.edu/1716491/Paperless_Recording_at_the_Sangro_Valley_Project [16.11.2014].

Museum of London (1994). Archaeological Site Manual. 3rd ed. London: MoLAS.

Oldman, D. & Rahtz, S. (2014). Aligning the Academy with the Cultural Heritage Sector through the CIDOC CRM and Semantic Web Technology. *In Papers from the 42nd Annual Conference of Computer Applications and Quantitative Methods in Archaeology (CAA)*. Paris.

Pavel, C. (2012). Archaeological Recording: Form and Content, Theory and Practice (Atek Na/ En La Tierra 2012, Buenos Aires, 2012, 33-74). *Atek Na/ En La Tierra*, pp. 33-74.

Rahm, E. & Hai Do, H. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Data Eng. Bull. 23 (4)*: 3-13.

Renfrew, C. & Bahn, P. (2004). *Archaeology: Theories, Methods and Practice. 4th ed*. London: Thames & Hudson.

Rosenberg, D. (2013). Data before the Fact. In Gitelman, L. (ed.). *"Raw Data" is an Oxymoron*, pp. 15-40. Cambridge: MIT Press.

Ross, Sh., Sobotkova, A., Ballsun-Stanton, B. & Crook, P. (2013). Creating eResearch Tools for Archaeologists: The Federated Archaeological Information Management Systems Project. *Australian Archaeology 77 (December)*: 107-119.

Roussey, C., Pinet, Fr., Ah Kang, M. & Corcho, O. (2011). An Introduction to Ontologies and Ontology Engineering. In Ontologies in Urban Development Projects, pp. 9-38. *Advanced Information and Knowledge Processing 1*. London: Springer. http://link.springer.com/chapter/10.1007/978-0-85729-724-2_2 [16.11.2014].

Schadla-Hall, T. (1999). Editorial: Public Archaeology. *European Journal of Archaeology 2 (2)*: 147-158. doi:10.1179/eja.1999.2.2.147.

Schloen, J. D. (2001). Archaeological Data Models and Web Publication Using XML. *Computers and the Humanities 35 (2)*: 123-152.

Tudhope, D., Binding, C., Jeffrey, St., May, K. & Vlachidis, A. (2011). A STELLAR Role for Knowledge Organization Systems in Digital Archaeology. *Bulletin of the American Society for Information Science and Technology 37 (4)*: 15-18. doi:10.1002/bult.2011.1720370405

W3C (2013). SPARQL Query Language for RDF. SPARQL 1.1 Overview. March 21. http://www.w3.org/TR/sparql11-overview/ [16.11.2014].

Wallrodt, J. (2012). The Database. Paperless Archaeology. http://paperlessarchaeology.com/the-database/ [16.11.2014].

Abbreviations

AI	Artificial Intelligence
CRM	CIDOC Conceptual Reference Model
CS	Computer Science
DC	Dublin Core
CRM-EH	English Heritage's archaeological extension to the CIDOC CRM
HTTP	Hypertext Transfer Protocol
MoLAS	Museum of London Archaeology Service
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
SC	Single Context archaeological excavation and recording method
URI	Uniform Resource Identifier

About the author

Frank Lynam has more than a decade of experience working in the technology sector. He completed his BA at Trinity College Dublin in Ancient History and Archaeology and Italian, and his MPhil in Archaeology at the University of Cambridge. He is currently in the final year of a 4-year Digital Arts and Humanities PhD under the supervision of Dr Christine Morris at Trinity College Dublin. His doctoral research considers how archaeology might benefit from Big Data analysis using Linked Open Data and Semantic Web techniques.

Frank Lynam
Department of Classics
Trinity College Dublin
flynam@tcd.ie