

University of Southampton Research Repository ePrints Soton

Copyright © and Moral Rights for this thesis are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given e.g.

AUTHOR (year of submission) "Full thesis title", University of Southampton, name of the University School or Department, PhD Thesis, pagination

UNIVERSITY OF SOUTHAMPTON
FACULTY OF PHYSICAL AND APPLIED SCIENCES
School of Electronics and Computer Science

Archaeology and the Semantic Web

by
Leif Isaksen

Thesis for the degree of Doctor of Philosophy

December 2011

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL AND APPLIED SCIENCES
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

ARCHAEOLOGY AND THE SEMANTIC WEB

by Leif Isaksen

This thesis explores the application of Semantic Web technologies to the discipline of Archaeology. Part One (Chapters 1–3) offers a discussion of historical developments in this field. It begins with a general comparison of the supposed benefits of semantic technologies and notes that they partially align with the needs of archaeologists. This is followed by a literature review which identifies two different perspectives on the Semantic Web: Mixed-Source Knowledge Representation (MSKR), which focuses on data interoperability between closed systems, and Linked Open Data (LOD), which connects decentralized, open resources. Part One concludes with a survey of 40 Cultural Heritage projects that have used semantic technologies and finds that they are indeed divided between these two visions.

Part Two (Chapters 4–7) uses a case study, Roman Port Networks, to explore ways of facilitating MSKR. Chapter 4 describes a simple ontology and vocabulary framework, by means of which independently produced digital datasets pertaining to amphora finds at Roman harbour sites can be combined. The following chapters describe two entirely different approaches to converting legacy data to an ontology-compliant semantic format. The first, TRANSLATION, uses a ‘Wizard’-style toolkit. The second, *Introducing Semantics*, is a wiki-based cookbook. Both methods are evaluated and found to be technically capable but socially impractical.

The final chapter argues that the reason for this impracticality is the small-to-medium scale typical of MSKR projects. This does not allow for sufficient analytical return on the high level of investment required of project partners to convert and work with data in a new and unfamiliar format. It further argues that the scale at which such investment pays off is only likely to arise in an open and decentralized data landscape. Thus, for Archaeology to benefit from semantic technologies would require a severe sociological shift from current practice towards openness and decentralization. Whether such a shift is either desirable or feasible is raised as a topic for future work.

To Jessica Rose Ogden, who was there from the beginning.

Contents

List of Figures	vii
List of Tables	ix
Declaration of Authorship	xi
Acknowledgements	xii
Nomenclature	xiii
1 Introduction	1
1.1 Preamble	1
1.2 The Semantic Web	2
1.2.1 Potential Benefits	3
1.2.2 Potential Limitations	5
1.3 Archaeology	7
1.4 Semantic Technologies and Archaeology	9
1.5 Research Aims and Contribution	12
1.5.1 Additional Outputs	12
1.6 Chapter Summary	13
2 The Semantic Web and Cultural Heritage	17
2.1 Overview	17
2.2 A Brief History of the Semantic Web	17
2.2.1 The Semantic Web Era (2001-2005)	20
2.2.2 The Linked Data Era (post-2005)	27
2.3 The Semantic Web in Cultural Heritage	33
2.3.1 The CIDOC CRM	34
2.3.2 Galleries, Libraries, Archives and Museums	36
2.3.3 Archaeology	40
2.4 Conclusions	45
3 Survey — Evaluating Semantic Technologies for Data Publication in Cultural Heritage	47
3.1 Overview	47
3.2 Method	48
3.2.1 Participants	49
3.2.2 Procedure	50

3.2.3	Bias and Limits of Scope	50
3.3	Findings	53
3.3.1	The Projects	53
3.3.2	Intentions	56
3.3.3	Semantic Technologies Employed	59
3.3.4	Data Conversion	60
3.3.5	Access and Consumption	62
3.3.6	No Linked Open Data?	64
3.3.7	Respondents' Comments	64
3.4	Conclusions	68
4	MSKR in Archaeology: Building a Framework	71
4.1	Overview	71
4.2	Case Study: Roman Port Networks	72
4.3	Requirements	74
4.4	Infrastructure	76
4.4.1	Domain Ontology	77
4.4.2	Thesauri	80
4.4.3	Hosting and Maintenance	82
4.5	Conclusions	82
5	TRANSLATION — A Toolkit for Generating RDF	85
5.1	Overview	85
5.2	Stage 1: Mapping	86
5.2.1	Configuration File	86
5.2.2	Architecture	87
5.2.3	Configuration details	88
5.2.4	Schema Mapping	89
5.2.5	URI Minting	91
5.2.6	URI Search	92
5.2.7	Literal Standardization	95
5.3	Stage 2: Export	97
5.4	Stage 3: Representation	98
5.5	Evaluation	98
5.6	Conclusions	101
6	Introducing Semantics — A Semantic Cookbook	105
6.1	Overview	105
6.1.1	<i>Introducing Semantics</i>	106
6.1.2	Infrastructure	108
6.1.3	Layout	109
6.2	Semantic Level 1: Literal Standardization	110
6.2.1	Overview and Aims	110
6.2.2	Recipes	111
6.2.3	Visualization and Analysis	121
6.3	Semantic Level 2: Introducing URIs	123
6.3.1	Overview and Aims	123

6.3.2	Recipes	124
6.3.3	Visualisation and Analysis	133
6.4	Semantic Level 3: Introducing RDF	137
6.4.1	Overview and Aims	137
6.4.2	Recipes	138
6.4.3	Visualisation and Analysis	144
6.5	Evaluation	147
6.6	Conclusions	148
7	Discussion, Conclusions and Future Work	151
7.1	Review	151
7.2	Discussion	153
7.3	Conclusions	155
7.3.1	Openness	155
7.3.2	Decentralization	158
7.3.3	Linked Open Data	159
7.4	Contributions and Future Work	160
A	Semantic Web Conference Papers	165
B	STCH Survey: Participant Information Sheet	171
C	STCH Survey: Consent Form	175
D	STCH Survey: Questionnaire	177
E	STCH Survey: Participants	187
F	STCH Survey: Data	191
G	Roman Port Networks: Partners	211
H	<i>Introducing Semantics</i> Questionnaire	215
	Glossary	217
	Bibliography	217

List of Figures

2.1	The Semantic Web Layer Cake in 2000	23
2.2	The Semantic Web Layer Cake in 2007	23
2.3	The LOD cloud diagram in 2007 (Cyganiak and Jentzsch, 2007).	30
2.4	The LOD cloud diagram in 2011 (Cyganiak and Jentzsch, 2011).	30
2.5	Google Trends: Search volume index by year for the terms ‘semantic web’, ‘linked data’, ‘web of data’ and ‘web 3.0’ (2004-2011)	32
2.6	Google Trends: Search volume index by year for the terms ‘semantic web’ and ‘web 2.0’ (2004-2011)	32
4.1	A <i>Dressel 20</i> Roman amphora (Keay and Williams, 2005)	73
4.2	ArchVocab Excavation Ontology diagram	79
5.1	UML Package Diagram of the Data Inspector Wizard	88
5.2	Data Inspector Wizard: Basic configuration information	89
5.3	Data Inspector Wizard: Ontology-to-column schema mapping	90
5.4	Data Inspector Wizard: URIs for instance data	91
5.5	Data Inspector Wizard: Mapping excavations to GeoNames URIs	93
5.6	Data Inspector Wizard: Amphora typology system mapping	94
5.7	Data Inspector Wizard: Amphora type mapping	95
5.8	Data Inspector Wizard: Entering <i>terminus post quem</i> Literals	96
5.9	Results of RDF generation displayed in SIMILE Exhibit	99
5.10	Formalization benefit variance across different users	102
5.11	Wave-type formalization process	103
6.1	An example of a recipe in PBWorks	109
6.2	Example Accumulation Graph	121
6.3	Amphora data from Carthage visualized in Many Eyes	122
6.4	A Roman excavation site in Freebase	126
6.5	A dataset in Google Refine	129
6.6	A spreadsheet exported from Google Refine showing hyperlinked contents	132
6.7	A dynamic network diagram showing the probable origins of ceramic finds from a Roman harbour	136
6.8	The RDF Explorer	146

List of Tables

1.1	Potential benefits of semantic technologies for Archaeology	11
2.1	Key developments in the Semantic Web (2001-2011)	21
2.2	Number of blog posts aggregated per year by Planet RDF (2006-2011). . .	33
2.3	STCH Survey: New Cultural Heritage projects utilising semantic technologies per year (2002-2009)	34
2.4	Number of papers presented at Museums and the Web conference on semantic technologies (2002-2010)	37
2.5	Number of papers presented at CAA on semantic technologies (2001-2011)	41
3.1	STCH Survey: Projects by region	50
3.2	STCH Survey: Projects by domain	51
3.3	STCH Survey: Respondents by specialism	52
3.4	STCH Survey: Respondents by role	52
3.5	STCH Survey: Projects by duration	53
3.6	STCH Survey: Open-ended vs. Fixed-term	53
3.7	STCH Survey: Projects by people involved	54
3.8	STCH Survey: Projects by institutions involved	54
3.9	STCH Survey: Projects by datasets integrated	55
3.10	STCH Survey: Projects by datasets with different schemas integrated . .	55
3.11	STCH Survey: Projects by intended user group(s)	56
3.12	STCH Survey: Project data availability	56
3.13	STCH Survey: Projects' priorities — data complexity vs. data quantity .	57
3.14	STCH Survey: Projects' priorities — data utility vs. data integrity . . .	57
3.15	STCH Survey: Project data sources	58
3.16	STCH Survey: Project aims	59
3.17	STCH Survey: Semantic Technologies used	59
3.18	STCH Survey: URI usage	60
3.19	STCH Survey: URI hosting	60
3.20	STCH Survey: External URI mapping	61
3.21	STCH Survey: RDF generation — process	61
3.22	STCH Survey: RDF generation — technologies	62
3.23	STCH Survey: RDF formats used	62
3.24	STCH Survey: Data consumption	63
3.25	STCH Survey: Respondent advocacy	63
3.26	STCH Survey: Respondent satisfaction	64
3.27	STCH Survey: Conformance with 'Linked Open Data Rules'	65

3.28	STCH Survey: Variables associated with Fixed-term and Open-ended projects	68
4.1	RDF generation support tasks	83
5.1	Principal Elements of XML configuration file	87
6.1	Semantic Level 2 Column URI mappings	133
6.2	Archvocab RDF Object Columns divided by Subject	141
6.3	Semantic Level 3 <code>rdf:Predicate</code> mappings	143
F.1	STCH Survey Data 1A: Project Information	192
F.2	STCH Survey Data 1B: Project Goals	193
F.3	STCH Survey Data 2 & 3: Methods and Controlled Vocabularies	194
F.4	STCH Survey Data 4: URIs	195
F.5	STCH Survey Data 5 & 6: RDF & Ontologies	196
F.6	STCH Survey Data 7: Consumption	197
F.7	STCH Survey Data 8: Satisfaction	198
F.8	STCH Questions 3.2 & 6.6: Other vocabularies and ontologies used	199
F.9	STCH Questions 2.2 & 5.2: Other technologies used	200
F.10	STCH Question 7.4: Other APIs provided	200
F.11	STCH Question 7.6: Other human interfaces provided	200

Declaration of Authorship

I, LEIF ISAKSEN

declare that the thesis entitled

ARCHAEOLOGY AND THE SEMANTIC WEB

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission.

Signed:

Date:

Acknowledgements

This research has benefitted from the direct and indirect contributions of a great many people. It would be impossible to mention them all but the following people deserve explicit and grateful recognition.

My supervisory team, Kirk Martinez, Graeme Earl, Nick Gibbins, Simon Keay and Mark Weal have expertly guided me through the perils and pitfalls of interdisciplinary research. This work could never have been achieved without their support.

I am indebted to the many members of Roman Port Networks who offered constructive feedback after suffering through bugs, eccentricities, and typos in TRANSLATION and *Introducing Semantics*. I am especially grateful to those who offered me such genuine hospitality and friendship during my visits to their institutions: Giulia Boetto, Michel Bonifay, Daniela Giampaola, Anna Gutiérrez, Cèsar Carreras Monfort, Marinella Pasquinucci, Roberta Tomber and Enrique García Vargas.

The STCH survey forms a central plank of this thesis and the 67 people who completed it are owed a major debt of gratitude. I hope they will find the results as interesting and useful as I have.

Many people have been important sources of assistance or helped shape my thinking in various ways. Again, they are too numerous to list, but special thanks go to: Lucy Blue, Lance Draper, Paul Lewis, John McNabb, David Peacock, Tim Sly, David Wheatley and David Williams (Southampton), and Sean Elan-Gaston, Graham Klyne, Joel Phillips, John Sheridan, Monika Solanki, Stephen Stead, Jeni Tennison, the STAR/STELLAR team, and my colleagues on the Pelagios project and CAA Semantic SIG.

Help and support has been given by many friends over the past four years. Once again, it is impossible to name all of them but specific shout-outs go to: ‘The Stile Council’ — John Bustard, Dirk De Jager, Tom Hebborn, Tobias Kleemann, Rebecca Newland and Rikki Prince; My co-supervisees — Joshua Ellul and Salma Noor; All my friends in the Archaeology Department. A very special thanks goes to Nicole Beale, Keith May and Jessica Ogden for commenting on the manuscript.

Finally, I would most like to thank my family, and especially my partner, Jessica, for their enduring love and support throughout.

Thank you all for making this work as enjoyable as it was stimulating.

Nomenclature

AAT	Getty Art and Architecture Thesaurus
API	Application Programming Interface
BCE	Before Common Era
CAA	Computer Applications and Quantitative Methods in Archaeology
CE	Common Era
CIDOC	<i>Comité International pour la Documentation</i>
CRM	Conceptual Reference Model
CRM-EH	Conceptual Reference Model — English Heritage
CSV	Comma-Separated Values
ECS	School of Electronics and Computer Science
ETL	Extract-Transform-Load
EVE	Estimated Vessel Equivalent
GIS	Geographic Information System
GLAM	Galleries, Libraries, Archives and Museums
GUI	Graphical User Interface
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IDE	Interactive Development Environment
ISWC	International Semantic Web Conference
JDOM	Java Document Object Model
JSON	JavaScript Object Notation
JUNG	Java Universal Network/Graph Framework
KML	Keyhole Markup Language
KR	Knowledge Representation
LOD	Linked Open Data
MNI / NMI	Minimum Number of Individuals
MSKR	Mixed-Source Knowledge Representation
N3	Notation 3
NLP	Natural Language Processing
OWL	Web Ontology Language
R2RML	Relational Database-to-RDF Mapping Language
RCAHMS	Royal Commission on the Historic and Ancient Monuments of Scotland

RDB2RDF	Relational Database-to-Resource Description Framework
RDF	Resource Description Framework
RDFa	Resource Description Framework in Attributes
RDFS	Resource Description Framework Schema
SKOS	Simple Knowledge Organisation System
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
STCH	Semantic Technologies in Cultural Heritage Survey
SWRL	Semantic Web Rule Language
TGN	Getty Thesaurus of Geographic Names
UML	Universal Modeling Language
URI	Uniform Resource Identifier
URL	Uniform Resource Location
VGI	Volunteered Geographic Information
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language — Transformations

Namespaces

For ease of reading, URIs in the text have been shortened using the namespaces listed below. They are printed in monospace.

amphora:	<code>http://archvocab.net/amphora/</code>
archvocab:	<code>http://archvocab.net/excavation/ontology/</code>
batlas:	<code>http://atlantides.org/batlas/</code>
freebaseAmphora:	<code>http://rdf.freebase.com/ns/base.romanamphorae/</code>
freebaseRDF:	<code>http://www.freebase.com/view/</code>
freebaseRomanEmpire:	<code>http://rdf.freebase.com/ns/</code> <code>user.robert.roman_empire.roman_region/</code>
freebaseVisualArt:	<code>http://rdf.freebase.com/ns/visual_art/</code>
geonames:	<code>http://www.geonames.org/</code>
heml:	<code>http://www.heml.org/rdf/2003-09-17/heml#</code>
rdf:	<code>http://www.w3.org/1999/02/22-rdf-syntax-ns#</code>
rdfs:	<code>http://www.w3.org/2000/01/rdf-schema#</code>
skos:	<code>http://www.w3.org/2004/02/skos/core#</code>

Chapter 1

Introduction

1.1 Preamble

This research originally set out to address the question:

1. How can Semantic Web technologies usefully be applied to the discipline of Archaeology?

In the course of its development it has become clear that the issue might better be framed as three separate but connected questions:

1. What are the benefits to archaeologists of using semantic technologies to express their data?
2. What are the social and technical requirements for expressing archaeological data with semantic technologies?
3. To what extent is contemporary archaeological practice able and willing to meet the social and technical requirements of expressing data with semantic technologies?

It should be made clear from the outset that this research is not purely an exercise in Computer Science insofar as it does not attempt to solve an engineering problem. The entanglement of Web technology theory and practice within the specific domain of Archaeology makes the study of it an inherently interdisciplinary practice with few obvious templates to work from. It might be deemed an exercise in Archaeological Computing but that field is more usually concerned with the application of digital technologies (such as GIS, database management and virtual representation) by expert practitioners. In contrast, the Web typically draws its strength from participation by large numbers of

non-technical specialists. The growing field of Digital Humanities has begun to encompass New Media as a topic of research but its emphasis is frequently on the analysis of language and text in a way that sits somewhat at odds with the more fragmentary and statistical nature of archaeological data. As this research deals with the ‘why’, as well as the ‘how’, of the Semantic Web in Archaeology, the closest intellectual lens through which to explore it seems to be the nascent discipline of **Web Science**. As an academic subject born even since the inception of this PhD, adopting it as a framework may raise difficulties as well as mitigate them, but its guiding principle is very much in line with that which has come to underpin this inquiry — that the Web should be understood as both a social and technical structure (Hendler et al., 2008).

The particular difficulty faced by the Web Scientist, and indeed any interdisciplinarian, is how to address such issues in a way that does not leave one half of their intended audience in limbo while attending to the theoretical considerations of the other half. This introductory chapter is therefore an attempt to provide a *prima facie* case that semantic technologies *may* have benefits to offer the discipline of Archaeology and therefore that a critical appraisal of their requirements is worthwhile. It begins with a brief overview of the Semantic Web and some of the potential virtues which have been ascribed to it, compares them with some of the traditional goals of Archaeology and concludes that there is enough overlap to justify further investigation. The chapter finishes by laying out the structure of this thesis and its principal arguments.

1.2 The Semantic Web

The term ‘**Semantic Web**’ was originally associated with a *vision* rather than a technology (such as the Internet), a protocol (such as HTTP), an established practice (such as Web 2.0) or a resource (such as Wikipedia). It can be credited directly to Tim Berners-Lee (Shadbolt et al., 2006) but its original formulation is of a way in which the Web might be used, and not a set of implementations, specifications, observations, or data. Over time however, these other elements have also come about and the term can no longer easily be distinguished from them, or even the community and conceptual ‘brand’ that make use of them. It is futile to insist on a single definition, for language is embedded in communities rather than dictionaries. Yet with no little irony for an idea so concerned with the nature of meaning, such ambiguity has resulted in the phrase ‘Semantic Web’ (along with its near-synonyms, ‘Linked (Open) Data’ (Berners-Lee, 2006), ‘Web 3.0’ (MacManus, 2009b), ‘Web of Data’ (Korth, 2009) and ‘Giant Global Graph’ (Scott, 2009)) meaning many things to many people. Its perception in the minds of computer scientists, academics, politicians and opinion-formers alike varies tremendously and their views about its current state and future direction are equally diverse. It is therefore necessary to try and tease apart this multiplicity of motivations and views

in order to determine which, if any, have a relevance to the theory, method and culture of Archaeology.

Such an analysis is rendered more difficult by a technical, political and cultural landscape which is continually and rapidly changing. Some of the technologies referred to in this thesis were unthought of, let alone developed, when it commenced in 2007 and others were yet to become well-established. The Linked Data initiative,¹ in particular, has caused a major shift in the way that the Semantic Web is perceived by many outside of the academic community. The Open Data movement² has provided an important fillip in terms of accessible content and surfaced important social issues rarely encountered in university research laboratories. The commercial sector has also begun to play an important role but not always in the way that earlier proponents had envisaged.³ All these developments have both moulded and fragmented its popular image and in turn the research work that has gone into it. Despite this diversity, a ‘semantic killer app’ is no more apparent in 2011 than in it was when this research began, and thus there are few obvious pioneers who we might turn to for direct emulation. While it is not uncommon to hear people say that ‘the Semantic Web is now where the **Web of Documents** was in 1999’, the analogy is not terribly illuminating. Following the dot-com bust of 2000, the Web of Documents only seriously re-established itself after a major shift in the way it was used and understood (to the so-called **Web 2.0** paradigm). We may yet see equally dramatic changes in direction for the Web of Data.

1.2.1 Potential Benefits

Our task in this chapter is to ask, or perhaps even predict, what semantic technologies can offer Archaeology and we begin by examining what claims have been made in their favour. Motivations for the Semantic Web are numerous and not all of them are subscribed to, or even deemed desirable, by any given proponent (Marshall and Shipman III, 2003). The following non-exhaustive list is an attempt to summarize some of the principle benefits that have been attributed to it at various times.⁴

Machine-readable data

One of the earliest cited benefits of the Semantic Web is that it provides Web-based data that a machine can automatically parse and process. HTML, in particular, is a technologically mixed blessing, providing a simplicity that has helped the Web to grow, while enforcing upon it a Web of Documents that is human-readable but opaque to computation. Semantic technologies are envisaged as a means to make raw data equally accessible to machines (Berners-Lee and Fischetti, 1999, p. 191).

¹<http://linkeddata.org>

²<http://okfn.org>

³<http://schema.org/>

⁴References are by way of example only and not necessarily the publication in which the benefit was first cited.

Interactive data

A potential advantage for human consumers is the ability to ‘follow your nose’ in order to drill down for additional information about a concept. Sometimes referred to as ‘Data you can click on’ (Roberts, 2011).

Integrated data

Virtually all datasets suffer from some form of siloing, be it through access restrictions, schema differences, or offline storage. While substantial theoretical obstacles may remain, semantic technologies have frequently been vaunted as a technical solution for merging distributed, heterogeneous data (Berners-Lee et al., 2001).

Inferencing

One of the ambitions of the **Knowledge Representation (KR)** movement was the ability to reason over data in order to deduce second-order facts. The inclusion of well-defined properties and ontologies appears to make this a possibility on the Semantic Web (Allemang and Hendler, 2008).

Data integrity

The atomistic nature of data makes it highly susceptible to corruption. While it is impossible to rule out all unintended change, ontologies can help automatically identify internal inconsistencies which may be symptomatic of corrupt or problematic data (Tao, 2010).

Self-describing data

The importance of metadata has long been acknowledged but almost as frequently ignored by knowledge-workers due to the (fairly arbitrary) distinction between content and description. Self-describing resources could ideally generate metadata with little or no additional effort and at the level of the content itself, rather than high level summaries of a document (Bergman, 2009).

Persistent global data

Most datasets do not have a means of access which is globally accessible or use identifiers that cannot be conflated with those of other datasets. Semantic technologies potentially offer both (MacManus, 2009a).

Provenance

Knowing the source of information is often as important as the information itself. Semantic technologies can provide such information at a variety of levels (Omitola et al., 2010).

Personalization

‘Semantic agents’ are a special case of semantic integration and inference that apply user defined preferences and functions to generic data so as to facilitate interaction with complex or banal data processing situations (Berners-Lee et al., 2001).

Data contextualization

By automatically associating data with relevant online content elsewhere it may be possible to assist users in their comprehension of a topic and even lead to serendipitous discovery (Bourg, 2010).

Document decomposition

Semantically-encoded documents contain elements which can be more easily recombined or visualized in new ways (Bizer et al., 2009).

Repurposing

Perhaps a guiding philosophy of the Semantic Web, rather than a benefit, it proposes that data should be remixed and utilized in ways unforeseen by the original author. Most famously expressed in the intuitively correct but unsettling maxim that *“the coolest thing that can be done with your data will be thought of by someone else”* (Miller, 2007, quoting Rufus Pollock).

We need to remember that these are not necessarily interdependent goals. Different semantic applications emphasise them in different ways (or indeed, not at all), which makes it misleading to talk of the ‘average’ semantic website, service or dataset. This has led to widely differing expectations as to what advantages might be gained by investing in them. The difficulty is exacerbated by the association of semantic technologies with a range of other topics, including Web 2.0, Web Science, The Internet of Things and software agents. In such cases it may (sometimes) be a matter of more carefully disambiguating terms (Martinez and Isaksen, 2010), but the effect on public perception, which is ultimately the driving force behind adoption, should not be underestimated (Parry et al., 2008). Stakeholders of any stripe tend to evaluate the performance of a new technology based on their prior beliefs about what it can achieve, rather than any retrospective correction of them. Nonetheless, as our list makes clear, semantic technologies evidently offer potential solutions to a wide array of longstanding issues in knowledge management.

1.2.2 Potential Limitations

As a formal **W3C** Activity,⁵ there has been considerable evangelizing about the utility of semantic technologies from a number of key figures in the wider Web community. It is therefore a subject of considerable concern that such technologies have emphatically *not* seen adoption on a scale that might be expected given the long list of benefits. Möller et al. (2010) note that high profile semantic datasets such as that provided by DBpedia⁶ have received almost no growth in traffic and what traffic there is comes principally from

⁵<http://www.w3.org/2001/sw/>

⁶<http://dbpedia.org/sparql>

university research projects. Google Trends⁷ shows a continuous decline in searches for the term ‘Semantic Web’, and while the term ‘Linked Data’ has gained traction the numbers are paltry in contrast to other Web technologies (see Section 2.2.2, below). An email⁸ raising this issue on the Semantic Web mailing list in 2010 provoked over one hundred responses. Views on both diagnosis and cure were highly divergent and occasionally even contradictory. Among the opinions cited were the beliefs that:

- There is not strong enough integration with social networks;
- Better user interfaces need to be developed;
- Semantic data should be better integrated with extant user interfaces;
- Semantic services need to be commoditized;
- The Semantic Web requires a broader narrative;
- Semantic technologies need use cases beyond academia;
- Better cost-benefit analyses are required;
- The Semantic Web does not provide straightforward answers to users’ questions.

Similar public conversations elsewhere have included both the necessity for more careful brand management and the contrasting need to be a broad church (Davis, 2009; Wilde, 2009). It has more recently been claimed that semantic technologies are becoming fragmented between the major Web corporations (Murphy and Spivack, 2010). In short, while reports of the Semantic Web’s death (Newcomb, 2007; Beall, 2010),⁹ may be somewhat exaggerated, any recommendation as a paradigm for a resource-poor discipline such as Archaeology must at least be evaluated with caution.

The Semantic Web is not simply a solution looking for a problem. After all, there is a lengthy list of well-documented challenges in data management that semantic technologies may be capable of providing at least partial solutions to. Likewise we must acknowledge that the Semantic Web has never been offered as a miracle cure (despite occasional hyperbole). It is unreasonable to criticize new technologies on grounds which are equally applicable to more mainstream approaches. Instead, and as with all such tantalizing but uncertain opportunities, we must turn from the generic considerations which are the inevitable focus of the technical community, and take the time to consider them in the light of the specific needs of the domain of application, in this case Archaeology.

⁷<http://www.google.com/trends>

⁸‘Call to Arms’ <http://lists.w3.org/Archives/Public/semantic-web/2010Mar/0160.html>

⁹Let alone that of the Web itself (Anderson and Wolff, 2010).

1.3 Archaeology

On the face of it, semantic technologies would appear to have much to offer archaeologists. Archaeological discourse is fundamentally composed of fragmentary heterogeneous data. A discipline framed by a topic — the human experience understood through its material culture — rather than a method, it is extraordinarily diverse in its range. In addition to recording a range of material covering the full range of human industry and beyond, its theoretical analysis and practice draw upon disciplines that include Art History, Biology, Chemistry, Critical Theory, Economics, History, Geography, Law, Linguistics, Philosophy, Physics, Psychology, Sociology and of course Computer Science. Much as it is difficult to think of a discipline which has not been impacted by informatics, it is extremely hard to name one which does not contribute to archaeological thought in some manner. Furthermore, its combination of empirical observation and humanistic theorizing make it both a bridge and something of an outsider amongst the Sciences and Humanities. This incredible diversity makes Archaeology an exceptionally rich and engaging subject, but presents formidable obstacles to schematization.

Such obstacles are not merely conceptual. The breadth of the archaeological field, along with an Archaeology-for-Archaeology’s-sake philosophy typical of many Humanities subjects, leads to much diversity in opinion amongst archaeologists themselves about the methods of their discipline (Richards, 2009). We need not agree with Sayre’s Law¹⁰ when we note that any attempt to manage archaeological information that requires homogenization or consensus as a prerequisite is almost certainly doomed to failure. This paradox lies at the heart of the archaeological endeavour: almost all archaeological information is the surfacing of phenomena through the sum of multiple observations, yet the observations themselves are typically made by practitioners pursuing an extremely wide variety of theoretical agendas.

So much is self-evident to anyone working within the domain of Archaeology. Less apparent, and perhaps less agreed upon, however, is the issue of purpose. To what end is archaeological data created, managed, manipulated and disseminated? In scientific disciplines the answer to such a question is ostensibly more clear cut. While every theory is susceptible to refutation, as the probability of an observation tends to 1.0 it becomes ever more useful in terms of application. Regardless of the complexity of a medical support algorithm, for example, the output must be relatively simple — whether to operate on a patient or not, for example.¹¹ In turn, the efficacy of such decision-making processes can be determined readily by indicators such as a reduced mortality rate. In contrast, the Humanities rarely have such metrics by which to establish the ‘success’ or otherwise of their enterprise. It is certainly true that groundbreaking theories are

¹⁰ “*Academic politics is the most vicious and bitter form of politics, because the stakes are so low.*” (Shapiro, 2006, ‘Wallace S. Sayre’)

¹¹ I am grateful to Sir Michael Brady for this observation, made in response to a question at a workshop on ‘Understanding Image-based Evidence’, Oxford, 17 Nov. 2009. (Giacometti, 2009)

expounded and adopted, and discoveries made through careful detective work. Yet overall, the value of Archaeology, and thus its constituent contributions to the Academy and beyond, are essentially a Social Good. We invest in Archaeology as societies because we believe it enriches our cultural lives and plays an important role in creating our sense of identity and community. The operative word here is ‘create’. If we invested less in Archaeology we would not feel any the less Human or socially-minded — we would simply hold different beliefs about our place in the world, albeit some what more at odds with the scientific paradigm our society has come to hold dear.

While such reflections may appear abstract, they are fundamental to understanding one of the key challenges of archaeological data integration. Archaeology is not merely a means to end, but undertaken for its own sake. No system is therefore likely to find substantial adoption within the discipline which does not provide immediate benefit to the practitioner and is furthermore flexible enough to account for the inevitable idiosyncrasies of their approach. In contrast to scientists, who are to some degree both beneficiaries and dependents of complex data management systems and standards, humanists can continue to function, *qua* humanists, without recourse to such technology. Occupying the scientific-humanistic middle ground, it is not yet clear which path Archaeology will ultimately tread, but for the time being there is little or no pressure to agree on standardised methodologies, let alone those involving a high level of technical complexity.¹²

We must also bear in mind that university research represents only a relatively small proportion of archaeological activity. The academic environment is complemented by other archaeological communities with equally diverse motives. Commercial units undertake assessment and excavation as part of the planning and development process. Here the discoveries occur essentially at random and motivation is a combination of financial profit (rarely very much) and passion for the job. Local and national governments play an important role in determining archaeological policy and providing some digital infrastructure. They may in turn be supported by other public and commercial organisations who provide specific services and software. The Galleries, Libraries, Archives and Museums (GLAM) sector is responsible for curating much of the physical and archival material that is produced, and beyond these ‘official’ organisations there is an enormous global antiquities trade that ranges from legitimate antiques shops and private collections on the one hand, to industrial-scale looting and organised crime on the other. Despite their common interests, collaboration and even mutual understanding between them is rare (Bradley, 2006).

It is worth mentioning that despite an emerging movement promoting openness, and a wide acceptance among professional archaeologists that their duty is ultimately to the

¹²Note that even the fully mature standards proposed by the W3C remain ‘Recommendations’. Likewise, archaeological institutions typically choose to adopt recording approaches through forums such as FISH (2008) and even professional bodies such as the Institute for Archaeologists tend to emphasise guidance over standards in terms of actual practice (Hinton, 2008).

wider public, Archaeology remains a predominantly closed discipline and the realm of the ‘expert’. The reasons for this are complex and not infrequently emotive or even personal. They range from legitimate concerns about looting and vandalism, through greyer issues of academic authority and self-interest, to personal aggrandisement and even vendetta. They are also widely divergent from country to country. In many jurisdictions, such as those within the United Kingdom, the majority of archaeological material is simply recorded and archived, and, while accessible in principle, largely forgotten in practice. In other countries archaeological research may be considered the property of the State, to be presented in ways that reinforce certain aspects of a national narrative. Yet other societies, particularly those with colonial histories, may have complex arrangements with indigenous populations who might wish to see remains repatriated and put beyond the reach of researchers.

In summary, the archaeological record is often contested, and even more commonly restricted or altogether inaccessible to all but a few. That this remains the case in the twenty-first century is clearly regrettable, but there is little evidence from either archaeologists or legislators to suggest that this will change in the foreseeable future. This is not to say that no effort is made to disseminate results. Journals and books continue to be published and the wider public are increasingly invited to participate in Open Days, visit large excavations, or follow site blogs. Projects such as OASIS,¹³ are also dramatically improving access to the catalogue of so-called ‘grey literature’, the reports which summarize excavation findings (Hardman, 2006). There is however no *culture* of sharing direct access to archaeological results either among peers or beyond (Mantegari et al., 2007).

This state of affairs seems especially unfortunate given that Archaeology is a very poorly resourced sector. While most evident in the depressed wages of the private sector — archaeological wages range between 13% and 53% lower than those in comparable industries (Price and Geary, 2008) — this is merely symptomatic of a field which usually has highly restricted budgets with which to obtain, create, manage and disseminate archaeological information. Complex technical solutions appropriate for large or well-funded institutions and consortia are likely to be beyond the reach of most archaeologists. Solutions which benefit from wide uptake and economies of scale must therefore have very low requirements thresholds to be viable within the archaeological domain as a whole.

1.4 Semantic Technologies and Archaeology

Given these characteristics, we must ask how well semantic technologies seem to fit with the practice of Archaeology. Gauging ‘suitability’ is a difficult task however, given that adoption is a piecemeal process undertaken by numerous individuals with differing

¹³<http://oasis.ac.uk/>

perspectives. There is no objective measure of how much a technology might benefit the discipline as a whole (much though one might have an opinion on the matter). Instead, two dimensions, ‘desirability’ and ‘capability’, might be a better way of considering its appeal. In other words, we should take into account both the degree to which practitioners might see its advantages in their day-to-day work and the extent to which it is practicable given the resources they are likely to have available. If either of these seem unduly problematic then further evaluation would seem unnecessary.

Let us begin with desirability. Table 1.1 lists the potential benefits articulated above, along with reflections on the level of significance mainstream archaeologists are likely to assign to them. This is not intended to be an authoritative declaration of archaeological motives, but rather a first pass at detecting areas of commonality. It is clear that many, if not all, of the advantages ascribed to the Semantic Web are likely to be of interest to archaeologists. Interactive data, data integration, contextualization, provenance and document decomposition all seem like functionality with obvious appeal, while logical inferencing, personalization and Web persistence seem less likely to attract demand. Machine readability, data integrity and content repurposing may produce important secondary benefits but are perhaps not of such immediate significance in themselves.

If it seems there is a case to be made that at least some of the outcomes are worthwhile, how capable is the discipline of adopting such an approach? Factors to consider include:

1. Ease of data production;
2. Ease of data consumption;
3. Cost;
4. Requirements;
5. Alternatives.

With regard to the first two criteria, the simple fact is that there are as yet no clearly laid out generic methodologies for either the production or consumption of archaeological semantic data that do not require specialist intervention. Thus, not only is it difficult to estimate the third criterion, cost — either as a whole or to individual practitioners — but the requirements themselves from which that cost can be calculated are equally hard to determine. Likewise, the option of cheaper alternatives that are nevertheless sufficient to meet the immediate needs of archaeologists must be taken into consideration. In an industry where earnings are low and budgets comparatively small, dependency on highly paid developers and expensive computer equipment can jeopardise other items on a project’s budget. This is an issue faced across the Humanities but, for the time being at least, an obvious equilibrium between digital and non-digital costs has yet to be established.

TABLE 1.1: Potential benefits of semantic technologies for Archaeology

Criterion	Desirability	Notes
Machine-readable	**	Few archaeologists can program software, but allowing computers to automate simple tasks would clearly be beneficial.
Interactive data	***	A large amount of archaeological research time is spent looking up information about concepts. Instant access to such data would be a significant boon.
Integrated data	***	Archaeology is fundamentally about piecing together fragmentary data.
Inferencing	*	Archaeologists may be sceptical of fully-automated processes of logical deduction.
Data integrity	**	It is accepted that logical inconsistency is inevitable in such a broad discipline. Identifying such inconsistencies may suggest interesting areas of research however.
Self-describing	**	Archaeologists are aware of metadata's importance but not as good at recording it. Automated processes may help.
Persistent global data	*	Few archaeologists are yet willing to make data freely available on the Web, preferring traditional publishing paradigms.
Provenance	***	Knowing the source of archaeological data is very important in judging its utility.
Personalization	*	Archaeologists usually (if not always) intend to be scientifically objective.
Contextualization	***	Providing additional information of potential value is highly helpful to the archaeological process.
Document decomposition	***	The ability to extract and utilize data from grey literature would be extremely useful.
Repurposing	**	Archaeology depends heavily upon prior work but there is a tendency to release 'finished products', rather than raw data, which inhibits repurposing.

1.5 Research Aims and Contribution

It seems that there is indeed a *prima facie* case for evaluating semantic approaches in Archaeology but this stems as much from uncertainty about their actual demands on the discipline as from their supposed benefits. The aim of this research is thus to evaluate the potential of semantic technologies for archaeological practice. Every effort has been made to make it accessible to both an archaeological and a Computer Science audience, hopefully without significant detriment to either.

The first part of this thesis (Chapters 1–3) reviews both past and current applications of semantic technologies in Archaeology and the wider Cultural Heritage sector by means of a literature review and targeted survey. The results of this review suggest that two separate visions of the Semantic Web appear to have evolved: Mixed-Source Knowledge Representation (MSKR), and Linked Open Data. The former is centred on formal approaches to data integration and inferencing in a closed environment, the latter attempts to enhance the interconnectivity of documents across the Open Web.

Given the comparatively limited possibilities for working with Open Data in Archaeology, the second part of this thesis (Chapters 4–7) evaluates the possibilities of applying an MSKR approach through a specific case study, Roman Port Networks. It begins by outlining an overall infrastructural framework required to achieve the goals of the project. The following chapters then evaluate two alternative technical methodologies intended to facilitate the semantic publication process in a closed collaborative environment. Chapter 7 concludes that MSKR is likely to be of limited application within Archaeology. In particular, it notes that the principal value of the Semantic Web is its scalability. The relatively high investment required in time across multiple publishers is unlikely to be sufficiently rewarded by the comparatively limited quantity of data that can be aggregated in the closed environments typical of today's discipline. Some final reflections are then made as to whether an Linked Open Data-based approach may be more viable within the field.

1.5.1 Additional Outputs

In addition to this thesis a number of additional resources have been developed:

- Archvocab.net: An ontology for archaeological excavation,
- TRANSLATION: Several bespoke java software applications for mapping and visualizing archaeological data,
- *Introducing Semantics*: A wiki containing both guides to conversion of archaeological data.

- Data from the ‘Semantic Technologies in Cultural Heritage’ survey.

These resources are made freely available, and will be deposited along with the digital version of this thesis, under Creative Commons Zero¹⁴ and GNU Public License (GPL)¹⁵ licenses. They were developed solely for the purposes of research however, and no claim is made for their future maintenance or utility in other contexts.

1.6 Chapter Summary

The thesis is divided into seven chapters.

This introduction forms Chapter 1. It begins by outlining the principal questions to be addressed. It then offers a brief overview of some proposed benefits of the Semantic Web, as well as some of its potential drawbacks, including its apparently sluggish adoption rate. It goes on to explore a number of current issues in archaeological data management, concluding that semantic technologies may have much to offer them. The path ahead is by no means clear, however, due to the lack of clear precursors, the reluctance of archaeologists to make their data openly available, and the high costs associated with digital practices. It then describes the research aims and contribution of the thesis: to evaluate the potential benefits of semantic technologies for MSKR in Archaeology through a combination of wide ranging review and practical experiment with a case study, Roman Port Networks. It concludes with this Chapter Summary.

Chapter 2 begins with a literature review of developments in semantic technologies since its initial conception by Berners-Lee in the mid-1990s. It highlights a number of debates within in the community as well as some of the more general critiques and responses to them. In particular, it notes a significant split in 2006 between two visions which are here characterized as ‘Mixed-Source Knowledge Representation’ and ‘Linked Open Data’. It continues with a more in-depth study of semantic projects related to Cultural Heritage and Archaeology, citing specific examples. The study finds that there is a very large degree of variation between projects and few clearly defined methodologies or even terminology. Instead, most Cultural Heritage projects that could be categorised as ‘semantic’ draw from a pool of technologies that rarely permit interoperability between each other in practice. The chapter concludes by stating the need for a comprehensive survey of Cultural Heritage projects that have used semantic technologies in order to establish whether they are chiefly influenced by the MSKR or Linked Open Data paradigm, or a mixture of both.

Chapter 3 discusses the results of a survey carried out by the author across 40 Cultural Heritage projects using semantic technologies. Topics include the actual technologies

¹⁴<http://creativecommons.org/publicdomain/zero/1.0/>

¹⁵<http://www.gnu.org/copyleft/gpl.html>

used, the projects' social structures, the intended goals, the current status of the data and the success of the semantic approach taken in the view of the respondents. The results indicate wide variation and mixed, although predominantly positive, feelings about semantic approaches. An important conclusion is that projects do indeed seem to be divided between MSKR and Linked Open Data, and that this is frequently associated with whether the project is fixed-term or open-ended.

Chapter 4 begins a practical examination of the possible benefits to Archaeology of MSKR for closed, collaborative projects, based on a central case study, Roman Port Networks. It provides some detail of the historic and archaeological issues the project addresses as well as the methods employed. It then discusses the need to lower the difficulty threshold of publishing semantic data for the so-called data micro-providers who make up the majority of archaeologists. The chapter goes on to describe the principles and implementation of the infrastructural components of a possible semantic framework. This includes the creation of the ArchVocab.net ontologies and the use of additional external resources, such as the GeoNames¹⁶ gazetteer.

Chapter 5 describes TRANSLATION, a Wizard-style approach to publishing semantic data that makes use of a software tool suite to map and export tabular excavation data to the ArchVocab ontology. The tool suite inspects the input dataset and leads the user through a number of steps that utilise Natural Language Processing (NLP)-based suggestion mechanisms in order to simplify the mapping process. The chapter then evaluates this approach, concluding that it is simultaneously too complex for a task which will be carried out infrequently, and conceals too much of the underlying process for the archaeologists to fully understand (and thus feel fully in control of) the nature of the transformation.

Chapter 6 discusses *Introducing Semantics*, an entirely different 'cookbook'-based approach to semantic data publication that avoids using new technologies wherever possible and requires the archaeologist to transform the data by hand. This increases the overall workload but also provides a number of useful 'break points' where the archaeologist can work with the data in a form that is familiar to them while providing previously unavailable functionality. It is also more open to adaptation without requiring programming skills. The process is directed by a series of consecutive guides, each of which equates to one of three 'Semantic Levels': *Literal Standardization*, *Introducing URIs*, and *Introducing RDF*. Following evaluation of this approach, the chapter concludes that while publication is achievable without the high level of support required by TRANSLATION, it seems that the level of time investment for perceived individual benefit is still too low to make it truly viable.

Chapter 7 summarizes the findings of the previous chapters before returning to the central question: whether semantic technologies are suitable for archaeological practice. It

¹⁶<http://www.geonames.org/>

finds that there are few, if any, technical impediments to MSKR but that its dependence on scale to offer significant return on users' time limits its attractiveness for those working with closed datasets. There follows a more theoretical discussion as to whether Linked Open Data approaches might offer a more promising alternative. This cautions that a significant shift towards openness and decentralization in the culture of archaeological practice would be required — as well as significant investment in infrastructural capacity — before a positive feedback loop of mutual publication and consumption can be reached. The question as to whether this is feasible, or even desirable, is raised as an important topic for future work.

Chapter 2

The Semantic Web and Cultural Heritage

2.1 Overview

In Chapter 1 we made the case that semantic technologies may have something to offer the archaeological profession(s) based on the ostensible benefits they provide. Just how close are they to delivering on those promises, however? Are there any significant consequences or limitations to their use? How have they been developed and deployed to date? And what might that tell us about their value in the archaeological domain? This chapter will explore those questions through a literature review, starting at “*a level of 20,000 feet*” (to paraphrase Berners-Lee (1998b)). This will lead us into a discussion of the historic development of the Semantic Web, which can be divided into two distinguishable eras of conceptualization and development. We will then look at its application in the Cultural Heritage sector, and particularly Galleries, Libraries, Archives and Museums, before finally considering its adoption and impact in Archaeology.

2.2 A Brief History of the Semantic Web

The concept of a Web of Data, in contrast to a Web of Documents, has been around as long as the Web itself (Berners-Lee, 1989, ‘Data Analysis’), and its roots go back yet further. The potential of universal identifiers and repositories of discrete but interconnected pieces of information was foreseen at least as far back as 1918 with Paul Otlet’s description of the ‘monographic principle’ (Rayward, 1994). An explanation of semantics as the accumulation of predicates around an empty variable, is the basis of Bertrand Russell’s groundbreaking philosophical paper ‘On Denoting’ (Russell, 1905). Tripartite links that specify their own meaning are a feature of the Memex machine

described in Vannevar Bush's landmark article 'As We May Think' (Bush, 1945). Even the notion of a 'document' as the basic unit of informational exchange was undermined by the concept of 'intertwining' introduced by Ted Nelson (1974). Nevertheless, the combination of these ideas is novel.

The earliest ideas about a Semantic Web were articulated by Tim Berners-Lee at the first World Wide Web conference¹ in 1994 (Shadbolt et al., 2006), but its present formulation as an official W3C Activity² can more or less be dated to three W3C *Design Issues* notes written by him for consultation in 1998–9: 'The Semantic Web Road Map', 'What the Semantic Web is Not' and 'Web Architecture at 50,000 feet' (Berners-Lee, 1998b,c, 1999). These documents form an interesting archive because they specify some of the earliest intentions behind the Semantic Web quite explicitly. Particularly marked is a tension, even an oscillation, between apparently mundane, if pragmatic, technical concerns on the one hand, and a vision of harmonious semantic integration on the other. At a basic level the Semantic Web is described as "*a web of data, in some ways like a global database*" and the method of implementation is to develop "*languages for expressing information in a machine processable form*" (Berners-Lee, 1998b). There are clear limits to this vision however:

A Semantic Web is not Artificial Intelligence. The concept of machine-understandable documents does not imply some magical artificial intelligence which allows machines to comprehend human mumblings. It only indicates a machine's ability to solve a well-defined problem by performing well-defined operations on existing well-defined data. Instead of asking machines to understand people's language, it involves asking people to make the extra effort. (Berners-Lee, 1998c)

Two important caveats are mentioned here. First there is a reminder that machines will never be able to correct for human error, ambiguity or misapprehension, a problem that has been known since the very dawn of computing technology.³ The second, and perhaps the more striking, is that it involves "*asking people to make the extra effort*". This may be to save labour in the long term, but it is only at the cost of significant human investment in the short term. Despite these warnings as to its limitations — the first common sense, the second less intuitive — Berners-Lee remains sanguine about its potential:

Though there will still not be a machine which can guarantee to answer arbitrary questions, the power to answer real questions which are the stuff of our

¹<http://www94.web.cern.ch/WWW94/>

²<http://www.w3.org/2001/sw>

³ "On two occasions I have been asked,—'Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?'... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question." (Babbage, 1864, p. 67)

daily lives and especially of commerce may be quite remarkable. (Berners-Lee, 1998b)

Other passages in the same text refer to a “*consistent logical web of data*” or even “*a Web in which machine reasoning will be ubiquitous and devastatingly powerful.*”. Yet even these passages are not without notes of caution. ‘Commerce’ and ‘questions which are the stuff of our daily lives’ are foreseen as the most likely beneficiaries. Is this because Berners-Lee believes that there is a crucial difference between the goals of these domains and others? Perhaps, but failing a direct critique it seems unduly pessimistic to dismiss the Semantic Web’s utility in other fields out of hand. A more revealing passage is that discussing its relationship to the field of **Knowledge Representation (KR)**, quoted here in full:

Knowledge Representation goes Global

Knowledge representation is a field which currently seems to have the reputation of being initially interesting, but which did not seem to shake the world to the extent that some of its proponents hoped. It made sense but was of limited use on a small scale, but never made it to the large scale. This is exactly the state which the hypertext field was in before the Web. Each field had made certain centralist assumptions – if not in the philosophy, then in the implementations, which prevented them from spreading globally. But each field was based on fundamentally sound ideas about the representation of knowledge. The Semantic Web is what we will get if we perform the same globalization process to Knowledge Representation that the Web initially did to Hypertext. We remove the centralized concepts of absolute truth, total knowledge, and total provability, and see what we can do with limited knowledge. (Berners-Lee, 1998c)

A number of things are interesting about this statement. The first is the confession that this is not so much a theory as an experiment. Berners-Lee openly acknowledges that KR approaches have not performed well in the past and offers no *a priori* proof that the Semantic Web will fare any better. The goal is simply to “*see what we can do*”. The second is that it declares explicitly that the Semantic Web has the same relationship to KR that the Web of Documents has to Hypertext — it is the result of performing a ‘globalization process’. What is this process? The next sentence defines it in terms of decentralization: “*We remove the centralized concepts of absolute truth, total knowledge, and total provability*”. A quote from his book *Weaving the Web*, published in 1999, provides further evidence that this decentralizing tendency is fundamental to the nature of the Web.

What was often difficult for people to understand about the design [of the World Wide Web] was that there was nothing else beyond URIs, HTTP and HTML. There was no central computer ‘controlling’ the Web, no single network on which these protocols worked, not even an organisation anywhere that ‘ran’ the Web. The Web was not a physical ‘thing’ that existed in a certain ‘place’. It was a ‘space’ in which information could exist. (Berners-Lee and Fischetti, 1999, p. 39)

We will return repeatedly to the theme of decentralization throughout the course of this thesis, but must now continue with an overview of the Semantic Web’s history following these earliest writings. We can begin by looking at developments within the community of informatics specialists working to develop it, orienting ourselves with a list of the most significant developments in a timeline (Table 2.1). Naturally it is hard to provide a hard and fast rule as to what constitutes a ‘significant development’ (and opinions will vary) but here the focus has been on the foundation of major conferences, key articles, the W3C Recommendation status of the principal technologies, the release of a handful of notable services, and its adoption by multinational corporations. Following this scheme, the history of the Semantic Web’s development can be divided into two broad periods. Somewhat helpfully, both of them broadly commence around the time of a key publication by Berners-Lee, reflecting not only on past work, but further providing an influential template for the development work which was to follow. We shall therefore refer to them as the **The Semantic Web Era** (2001-2005) and **The Linked Data Era** (2006-present).

2.2.1 The Semantic Web Era (2001-2005)

At the start of the seminal *Scientific American* article entitled ‘The Semantic Web’ (Berners-Lee et al., 2001) the Beatles song ‘We Can Work it Out’ is playing on a hypothetical hi-fi. It seems a fitting anthem, not only for the fictionalized software agents which support the protagonist in a manner slightly akin to the animals in a Walt Disney film, but also for the optimistic spirit of the piece itself and the ensuing direction of Semantic Web research. The general tone is similar to the earlier *Design Issues* notes, but two themes are given particular emphasis. The first is a restatement of the decentralized KR architecture that was discussed above:

Knowledge representation...is currently in a state comparable to that of hypertext before the advent of the Web: it is clearly a good idea, and some very nice demonstrations exist, but it has not yet changed the world [...]
To realize its full potential it must be linked into a single global system.
Traditional knowledge-representation systems typically have been centralized

TABLE 2.1: Key developments in the Semantic Web (2001-2011)

Year	Developments
2001	Semantic Web Working Symposium (later ISWC) 'The Semantic Web' (Berners-Lee et al., 2001)
2002	3store
2003	<i>Journal of Web Semantics</i>
2004	OWL (W3C Recommendation) RDFS (W3C Recommendation)
2005	SemTech conference
2006	'Design Issues: Linked Data' (Berners-Lee, 2006) 'The Semantic Web Revisited' (Shadbolt et al., 2006) GeoNames
2007	Linkeddata.org DBpedia Freebase
2008	SPARQL official W3C recommendation <i>Semantic Web for the Working Ontologist</i> (Allemang and Hendler, 2008) RDFa (W3C Recommendation) Microsoft and Yahoo adopt semantic search
2009	Talis Connected Commons OWL2 (W3C Recommendation) SKOS (W3C Recommendation) Data.gov and data.gov.uk
2010	<i>Semantic Web Journal</i> Freebase acquired by Google Facebook adopts RDFa
2011	Schema.org

[...] But central control is stifling, and increasing the size and scope of such a system rapidly becomes unmanageable.

Semantic Web researchers, in contrast, accept that paradoxes and unanswerable questions are a price that must be paid to achieve versatility [...] The challenge of the Semantic Web, therefore, is to provide a language that expresses both data and rules for reasoning about the data and that allows rules from any existing knowledge-representation system to be exported onto the Web.

Adding logic to the Web—the means to use rules to make inferences, choose courses of action and answer questions—is the task before the Semantic Web community at the moment. (Berners-Lee et al., 2001)

In short, the Semantic Web will create a framework in which different Knowledge Representation models can exist, and thus rules and inferencing are required to work out the differences between them. Decentralization is fundamental to this vision and equally so is the acceptance that the output will at times be messy. No guarantees are made about the quality or consistency of the results. While the initial input will be undertaken by humans, in time this process can be taken on by semi-autonomous software agents:

The real power of the Semantic Web will be realized when people create many programs that collect Web content from diverse sources, process the information and exchange the results with other programs. The effectiveness of such software agents will increase exponentially as more machine-readable Web content and automated services (including other agents) become available. The Semantic Web promotes this synergy: even agents that were not expressly designed to work together can transfer data among themselves when the data come with semantics. (Berners-Lee et al., 2001)

Machines talking to machines — is this vision of an entirely technical utopia? Not quite. There are explicit references to the human component of the Semantic Web. Not only does the fictional scenario's human protagonist interact with the software agents in order to correct their results, but the data itself is created by a human being:

[The] semantics were encoded into the Web page when the clinic's office manager (who never took Comp Sci 101) massaged it into shape using off-the-shelf software for writing Semantic Web pages along with resources listed on the Physical Therapy Association's site. (Berners-Lee et al., 2001)

In other words, human involvement is required at both the start and end of the process. Yet despite these human elements, the research which was to follow concentrated predominantly on the Semantic Web's KR aspects, in particular the standards and software required for reasoning over semi-structured networks of information. Of perhaps greatest significance was the development and formal Recommendation by the W3C of several key technologies in the so-called Semantic Web 'Layer Cake'. First produced by Berners-Lee in 2000, the Layer Cake represents the Semantic Web as a stack of technologies (Figures 2.1⁴ and 2.2⁵). The diagram has evolved considerably over time as layers slowly transformed from generic concepts (such as 'ontology vocabulary') to concrete specifications (such as OWL).

The Layer Cake not only provides an overview of the key technologies used by the Semantic Web, it also it makes clear their dependencies. While an overview of all the

⁴<http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html>

⁵<http://www.w3.org/2007/03/layerCake.png>

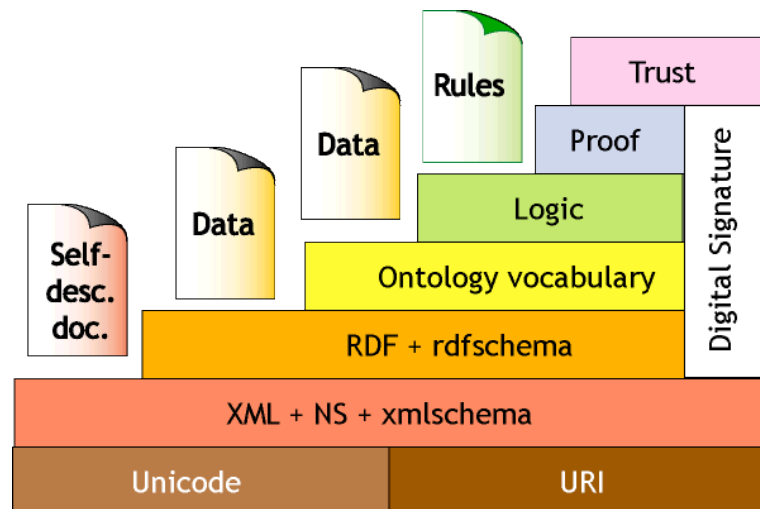


FIGURE 2.1: The Semantic Web Layer Cake in 2000

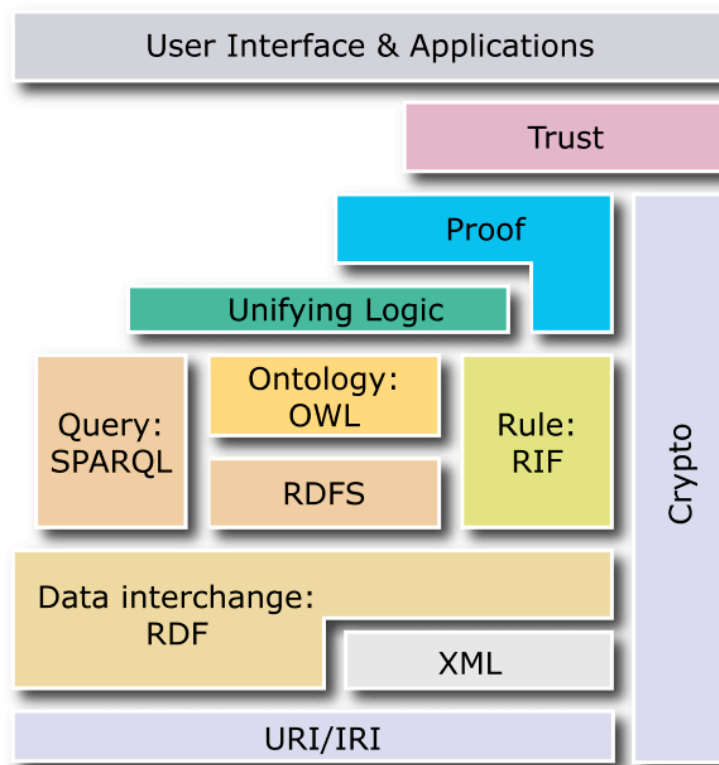


FIGURE 2.2: The Semantic Web Layer Cake in 2007

semantic technologies developed during this period would beyond the scope of the thesis, some knowledge of the three most fundamental layers — URIs, RDF and ontologies — is vital for understanding the ensuing discussion and they are worth describing briefly.

URIs

Uniform Resource Identifiers (URI) are globally identifying strings of text for a **Resource** which can be thought of as any atomic concept, whether a tangible

object, a class of things, an electronic document or something else. Although their specification (Berners-Lee et al., 1998) predates any formal Semantic Web activity they are fundamental to understanding how it works. A URI is composed of two sections — the ‘URI Scheme Name’ (such as ‘http’, ‘mailto’ or ‘urn’) and the ‘scheme-specific’ part (the syntax of which is scheme dependent) — separated by a colon. Examples of URIs might be:

- `http://en.wikipedia.org/wiki/Flinders_Petrie`
- `ftp://example.org/resource.txt`
- `urn:issn:1535-3613`

URIs are identifiers that remain independent of the data system in which they are stored, thereby creating the possibility for different data systems to refer unambiguously to the same concept. Although different schemata can be used for URIs it is considered best practice to use the HTTP URI scheme as this allows them to be **dereferenced** over the Web. In other words, a Web browser request to this identifier should redirect the user to a document containing information about it. This information may contain different words (or **labels**) that denote the concept in different languages or contexts, thereby providing a partial solution to the **Vocabulary Problem** (Furnas et al., 1987).

As full-length URIs are not easy for humans to read or remember, it is common to use abbreviated **namespaces** when referring to them. For example, one might use the prefix `wikipedia:` to refer to the Web subdomain `http://en.wikipedia.org/wiki/`. This would allow us to shorten the first URI example as `wikipedia:Flinders_Petrie`. Such abbreviations are not globally defined however, and must be provided to the parsing machine.

Although two instances of the same URI must be assumed to be references to the same concept, the Semantic Web is also based on the **Nonunique Naming Assumption** which postulates that two distinct URIs may in fact refer to the same concept (Allemang and Hendler, 2008, p. 11). As different users may well use different URIs when talking about the same thing, establishing this synonymy across the open Web is a difficult challenge known as the **Co-reference Problem** (Glaser et al., 2009).

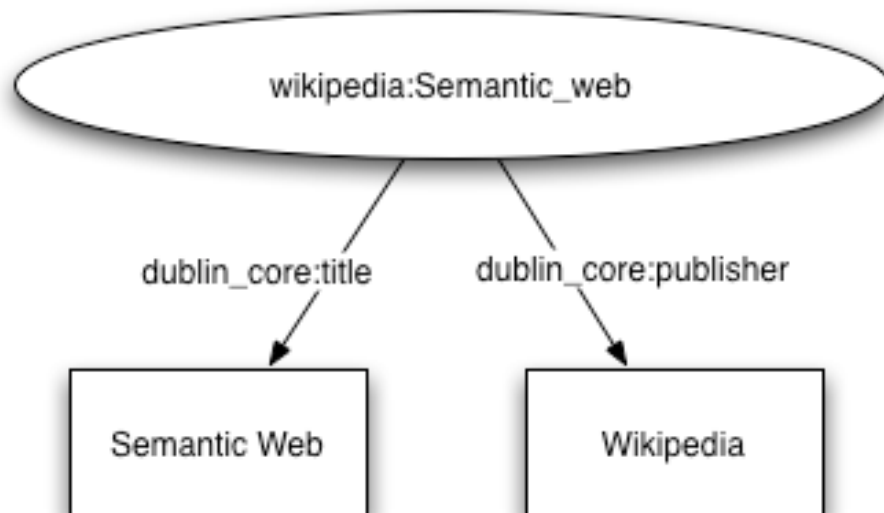
RDF

If URIs provide the fundamental building blocks, or ‘words’, of the Semantic Web, then **Resource Description Framework (RDF)** (Lassila and Swick, 1999) can be thought of as its ‘sentences’. RDF is a way of modeling knowledge by means of simple tripartite statements called **triples** formed of a Subject, a Predicate and an Object. The first two must be URIs identifying conceptual Resources, whereas the last may either be a URI or a **Literal** — typically a string of text or a number that is a simple value rather than an identifier. The great advantage of using RDF is

that any number of datasets can be merged simply by adding the triples together. When triples share the same URIs, combining them together forms a graph which can be analysed and queried.

URIs themselves are generally considered to be opaque and should not be **decomposed**, i.e. their meaning should not be deduced from their textual form. Instead, it should be inferred from the RDF triples associated with them. RDF is notation-independent and there are a number of different ways of expressing it including **Notation3 (N3)** (Berners-Lee, 1998a), **RDF/XML** (Beckett, 2004), **Turtle** (Beckett and Berners-Lee, 2011). It can be embedded in HTML in a format known as **RDFa** (Adida et al., 2008) or even represented diagrammatically. For instance, the following descriptions of the ‘Semantic Web’ entry in Wikipedia are all equivalent RDF representations:

Diagram



N3

```

<http://en.wikipedia.org/wiki/Semantic_web>
<http://purl.org/dc/elements/1.1/title>
"Semantic Web" .
  
```

```

<http://en.wikipedia.org/wiki/Semantic_web>
<http://purl.org/dc/elements/1.1/publisher>
"Wikipedia" .
  
```

RDF/XML

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://en.wikipedia.org/wiki/Semantic_web">
    <dc:title>Semantic Web</dc:title>
    <dc:publisher>Wikipedia</dc:publisher>
  </rdf:Description>
</rdf:RDF>
```

Ontologies

By itself, RDF puts very few constraints on what kinds of statements can be made. This could result in the same kind of information Babel that plagues database integration. To militate against this, digital **ontologies** — popularly defined as “*an explicit specification of a conceptualization*” (Gruber, 1993) — are used to restrict what can and cannot be said computationally about a given domain in a given model. In order to define the key concepts for describing such ontologies, several special RDF meta-ontologies have been developed including **RDF Schema (RDFS)** (Brickley and Guha, 2004), the **Web Ontology Language (OWL)** (McGuinness and van Harmelen, 2004) and **Simple Knowledge Organisation System (SKOS)** (Miles and Bechhofer, 2009). The URIs defined by them can be used to create statements that define the overall knowledge schema: what classes of concept exist and how they interrelate. Such ontologies make it a great deal easier for those describing their data, as well as those consuming it, to situate the meaning of individual triples within a broader semantic context.

These are perhaps the most central technologies to the Semantic Web but they are by no means the only ones. Other important technical innovations in this period include triple-stores for managing RDF (Harris and Gibbins, 2003), development frameworks (such as Jena⁶ and Sesame⁷), and both network and facet-based browsing interfaces, developed in order to aid visualisation of rich datasets.⁸ At a community level the Semantic Web Working Symposium, later to become the **International Semantic Web Conference (ISWC)**, was launched in 2001 with an additional European regional conference, **ESWC**, begun in 2004. 2005 saw the first **Semantic Technologies (SemTech)** conference, focussed on Enterprise applications. The *Journal of Web Semantics* published its first issue in 2003 and books offering high-level overviews of the topic also emerged in print (Daconta et al., 2003; Antoniou and Harmelen, 2004).

⁶<http://jena.sourceforge.net/>

⁷<http://www.openrdf.org/>

⁸See, for example, mSpace (Schraefel et al., 2005) and Tabulator (Berners-Lee et al., 2006).

Despite continual activity at a number of research institutions and the W3C however, Semantic Web content was much slower to materialise (Auer et al., 2009). Even from an academic perspective this was not without consequences, for the Semantic Web — like the Web of Documents — is ultimately dependent on continual use and contribution. At the same time, a number of critiques of the Semantic Web project began to emerge, concerned that it either under-represented the richness of human semantics (Gärdenfors, 2004) or was too confused to make progress (Marshall and Shipman III, 2003). Arguably a greater problem was brewing however. Even as the technical challenges faced by the Semantic Web were solved, it came at the expense of introducing ever greater complexity. This was not a problem for research laboratories looking to find new topics for research papers — the belief that ‘we can work it out’ seemed, if anything, more justified then ever. Unfortunately, it was also rapidly building an insurmountable knowledge threshold for newcomers and those outside of academia. If the Semantic Web was ever to appeal beyond the ivory tower and enter mainstream adoption it needed to find a new approach.

2.2.2 The Linked Data Era (post-2005)

It is important to recall that the period we have been discussing was notable for seismic shifts in the mainstream Web as well. The bursting of the ‘Dot-com Bubble’ in March 2000 — and the ensuing depression in Web investment that it led to — both bred cynicism about the social and economic models upon which early Web ventures had been based and sparked new innovation in order to overcome their limitations. The nature of that innovation, later to be branded Web 2.0, encouraged greater activity by making it easier for the general public to write to the Web, as well as read from it. While Berners-Lee can legitimately claim to have argued for such a ‘Read-Write Web’ since its earliest days (Berners-Lee and Fischetti, 1999, p. 77), it is not clear that this vision was as central to the work of early Semantic Web pioneers.

In 2006 Berners-Lee wrote a new *Design Issues* note entitled ‘Linked Data’. It outlined four so-called ‘rules’ (which he more explicitly defines as expectations) intended to “*make the web grow*”:

Like the web of hypertext, the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they link between arbitrary things described by RDF. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the web grow:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.

3. When someone looks up a URI, provide useful information[, using the standards (RDF*, SPARQL)].⁹
4. Include links to other URIs so that they can discover more things.

Simple. In fact, though, a surprising amount of data isn't linked in 2006, because of problems with one or more of the steps.

[...]

I'll refer to the steps above as rules, but they are expectations of behavior. Breaking them does not destroy anything, but misses an opportunity to make data interconnected. This in turn limits the ways it can later be reused in unexpected ways. It is the unexpected re-use of information which is the value added by the web. (Berners-Lee, 2006)

Was this note prompted by an apparent stall in progress? The frank acknowledgement that “*a surprising amount of data isn't linked in 2006*” certainly suggests so, but it is the simultaneous publication of an article (with Nigel Shadbolt and Wendy Hall) entitled ‘The Semantic Web Revisited’ (Shadbolt et al., 2006) which makes it clear that, in the view of its originator, the Semantic Web vision was heading off-track. It noted that strong progress *had* been made in terms of technologies but was nonetheless forced to conclude that the Semantic Web “*remains largely unrealized*”. In particular, this was deemed to be due to a lack of openness and decentralization from early adopters. While the technologies had seen some limited uptake, it was clear that the spirit of the Semantic Web — as described in the writings of 1998–9 — had not.

It is also instructive to consider typical Semantic Web projects of the past five years. They demonstrate a distinctive set of characteristics. Typically, they generate new ontologies for the application domain —whether its information management in breast diseases or computer science research. They either import legacy data or else harvest and redeposit it into a single, large repository. Then they carry out inference on the RDF graphs held within the repositories and represent the information using a custom-developed interface.

These projects have been important proving grounds for a number of techniques and methods. They show how to facilitate harvesting and semantic integration by using ontologies as mediators. They have served as a development context for RDF stores and a whole range of important Semantic Web middleware. In general, however, they lack real viral uptake. Moreover, in most cases, we aren't able to look up a URI and have the data returned. The data exposure revolution has not yet happened. (Shadbolt et al., 2006)

⁹The phrase in square parentheses was not included in the original post and subsequently inserted in June 2009. Source: WayBack Machine (<http://web.archive.org/web/20090526040403/http://www.w3.org/DesignIssues/LinkedData.html>).

Whether or not these publications provided the direct impetus, a number of important developments followed that appeared to take note of the issues they raised. Within less than a year several major ‘open’ repositories were either set up or chose to incorporate Linked Data Rules. **GeoNames**¹⁰ combines both Web-scraped and publicly volunteered spatial and placename data to create persistent URIs for places and human- and machine-readable data (including RDF) about their geographic location (Wick, 2006). **DBpedia**¹¹ generates URIs and extracts complex data automatically from the ‘infoboxes’ embedded in many Wikipedia pages. (Auer et al., 2007). **Freebase**,¹² a semantic encyclopaedia site developed by Metaweb, allows users to create their own content and properties but automatically prompts re-use of those previously created by others (Markoff, 2007). These projects and others often retain links back to their original source material and the heavy reliance on Wikipedia by all of them is indicative of their shared belief in the value of crowd-sourced and openly accessible public data.

Linkeddata.org was established in 2007 to “*provide a home for, or pointers to, resources from across the Linked Data community*”.¹³ In a deliberate shift away from many of the theoretical and technical discussions of the previous era, the site provides links to a variety of resources, including tutorials for those interested in putting their data online, but its chief aim is to encourage interconnectivity between them and not directly to itself. To this end it has published regular updates to the **Linking Open Data cloud diagram** since May 2007 (Figures 2.3 and 2.4). The diagram represents the interconnections between an increasing number of openly available datasets which follow the Linked Data Rules and has arguably come to replace the Layer Cake as the visual symbol of the Semantic Web.¹⁴ The data itself comes from a loose-knit community of contributors including public sector bodies (Scott and Smethurst, 2009), media companies (Rogers, 2009), academic mashups (Auer et al., 2007) and both for- and non-profit Web organisations (Dodds, 2009; Wick, 2006). The content may originate from archival material, be community generated, or created specifically with the Semantic Web in mind. It also ranges from instance data to vocabulary data (including taxonomies) to ontologies of various size and complexity (Idehen, 2008). Although the diagram’s authors explicitly state that it is not an exhaustive list of semantic resources (its data is drawn selectively from the CKAN¹⁵ directory), its rate of growth in both size and complexity has often been cited as a barometer of Semantic Web adoption.

¹⁰<http://www.geonames.org>

¹¹<http://dbpedia.org/>

¹²<http://www.freebase.com/>

¹³<http://linkeddata.org/>

¹⁴Although calls for the retirement of the Layer Cake had begun even earlier (Zacharias, 2007).

¹⁵<http://ckan.net/>

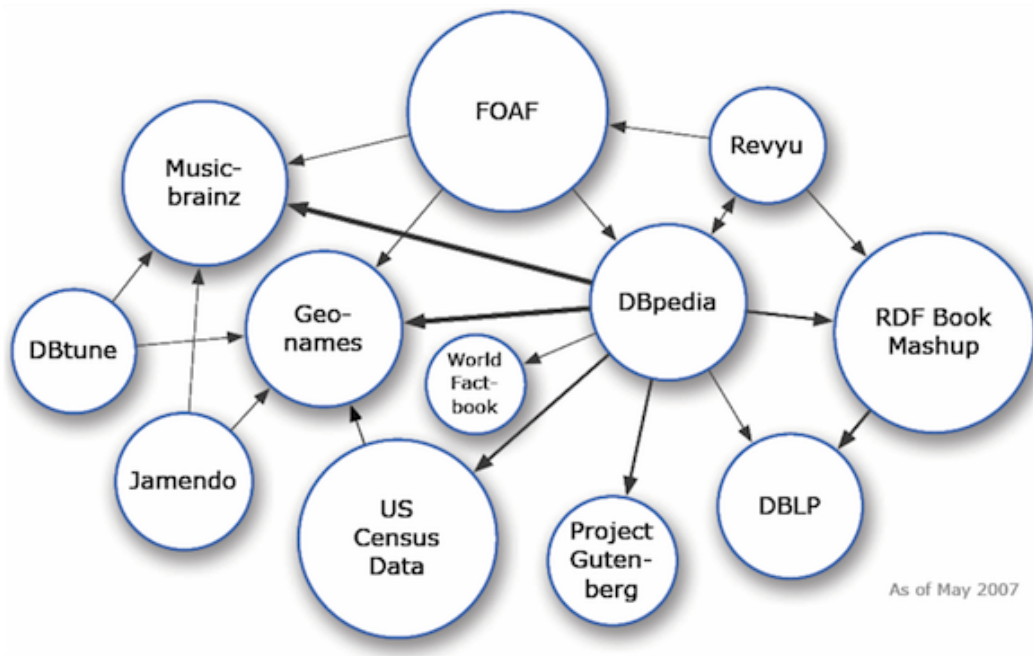


FIGURE 2.3: The LOD cloud diagram in 2007 (Cyganiak and Jentzsch, 2007).

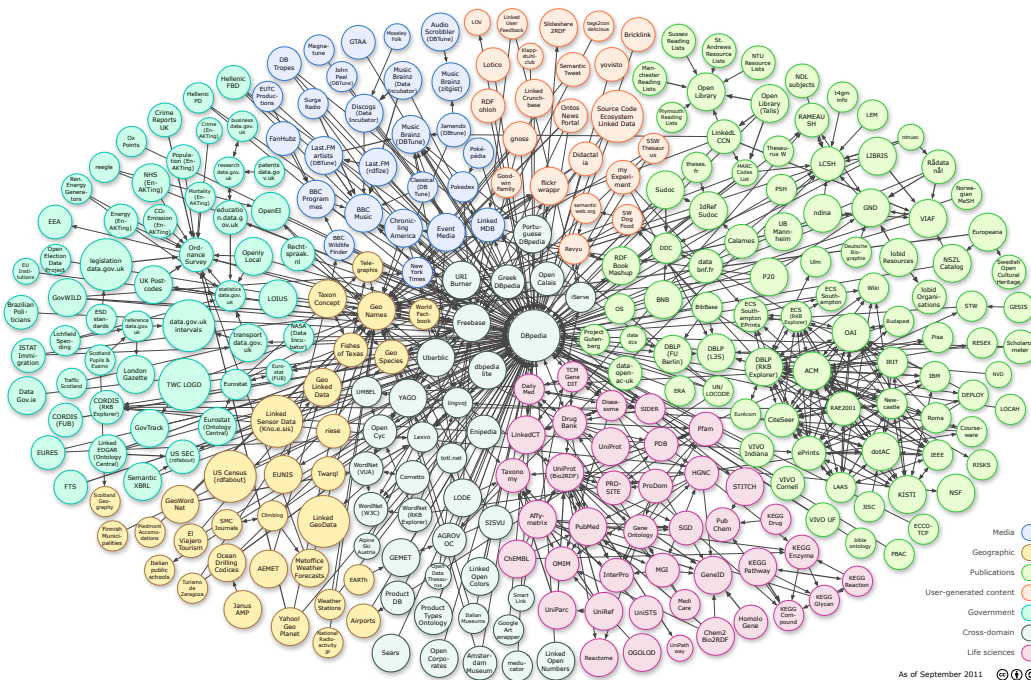


FIGURE 2.4: The LOD cloud diagram in 2011 (Cyganiak and Jentzsch, 2011).

2008 saw W3C Recommendation of RDFa (Adida et al., 2008) and **SPARQL**, a standardized query language for RDF analogous to the SQL language for querying relational databases (Prud'hommeaux and Seaborne, 2008). It also saw the high profile adoption of RDF by Yahoo! for its search technologies (Kumar, 2008) and the publication of *Semantic Web for the Working Ontologist* (Allemang and Hendler, 2008), a book providing

practical advice in implementing semantic technologies, rather than merely the theory behind them.

In 2009, two more technologies reached W3C Recommendation status: OWL2 (an extended version of OWL) (W3C OWL Working Group, 2009), and the SKOS ontology for thesauri (Miles and Bechhofer, 2009). Semantic content was given a huge fillip by the adoption of semantic technologies by the UK government data store¹⁶ (Tennison, 2010). 2010 was marked by two more high profile technology corporations adopting Semantic Technologies: Facebook's **Open Graph**¹⁷ protocol based on RDFa (Bell, 2010), and Google's acquisition of Metaweb, with Freebase data quickly being incorporated in its search results (Menzel, 2010).

The growing importance of the SemTech conference (Miller, 2010), the launch of infozines such as *Nodalities* by the Talis corporation (Miller, 2008), and the Open Access (and open review) journal *Semantic Web* (Hitzler and Janowicz, 2010) also indicate that the Semantic Web's demographic has begun to reach beyond the Academy since 2006. The addition of a new '5-Star Data' categorisation (with branded mugs and T-shirts) to the original 'Linked Data' *Design Issues* post also re-confirms Berners-Lee's opinion that openness is fundamental to the Linked Data vision. The very first criterion is that data be "*available on the web (whatever format), but with an open licence*" (Berners-Lee, 2006). So has the Semantic Web turned a corner? The picture remains complex at several levels.

One concern is the apparent decline in its public perception. According to Google Trends,¹⁸ searches for the term 'Semantic Web' and its near-synonyms have fallen as a proportion of Web searches year-on-year since records began in 2004 (Figure 2.5). The term 'Linked Data' may have increased slightly, but nothing like in the manner one might expect of Internet revolutions. 'Web of Data' saw a significant rise in late 2006 but even this has started to decline in recent years. All these are in marked contrast to the enormous growth in search traffic for the term 'Web 2.0' between 2005 and 2007, and it remains an overwhelmingly more frequent search term (Figure 2.6).

In the blogosphere the decline appears to be even worse. A chart of the number of posts per year on Planet RDF¹⁹ — a blog aggregator for Semantic Web content operational since late 2005 — shows a significant amount of material in 2006–8 before rapidly collapsing to one third of its previous volume by 2010 (Table 2.2). Print publications may be following suit with *Nodalities* publishing its final issue in August 2011 (Beauvais, 2011).

Meanwhile, questions as to what is impeding adoption continue to generate impassioned debate. While arguments continue to rage as to whether the problem is best solved

¹⁶<http://data.gov.uk/>

¹⁷<http://ogp.me/>

¹⁸<http://www.google.com/trends>

¹⁹<http://planetrdf.com/>

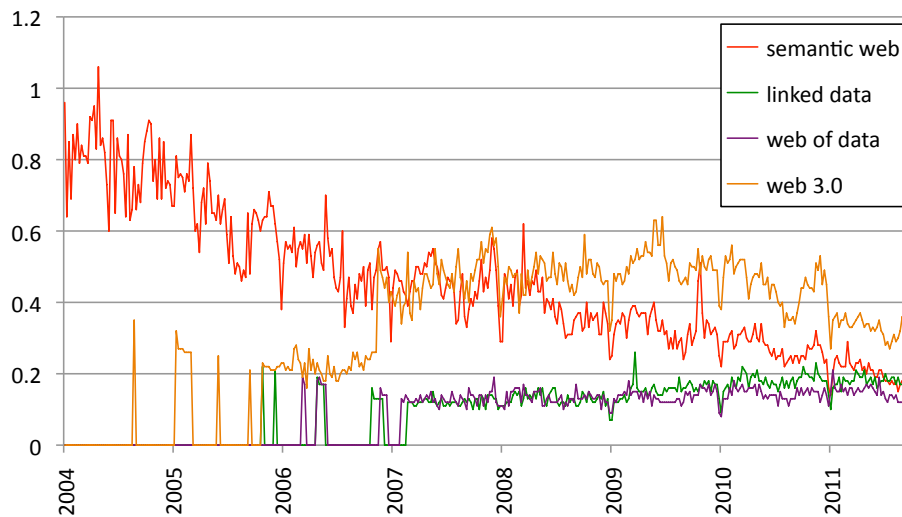


FIGURE 2.5: Google Trends: Search volume index by year for the terms ‘semantic web’, ‘linked data’, ‘web of data’ and ‘web 3.0’ (2004-2011)

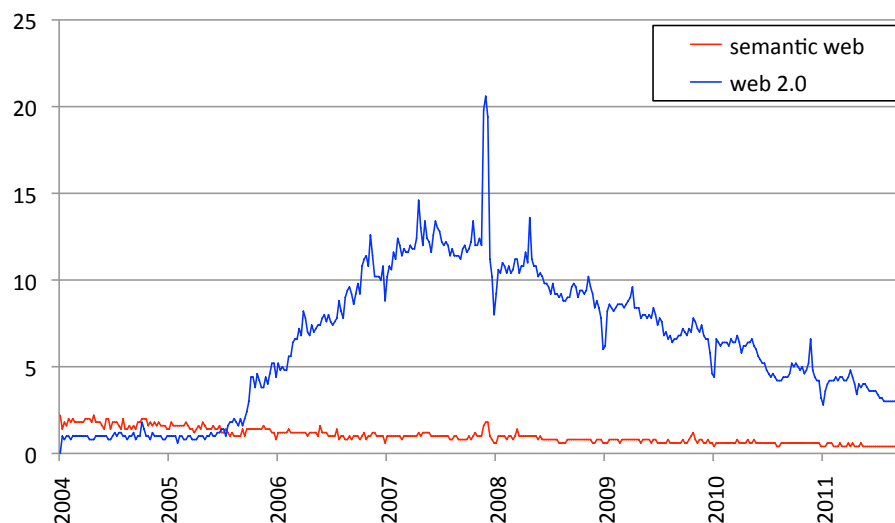
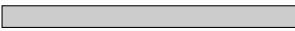





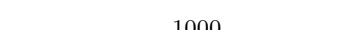


FIGURE 2.6: Google Trends: Search volume index by year for the terms ‘semantic web’ and ‘web 2.0’ (2004-2011)

through improved branding (Boutin, 2009), reduced complexity (Alani et al., 2008), or just getting our terms straight (MacManus, 2009b), there is certainly a consensus that the Semantic Web must vastly increase the level of engagement if it is not to become a footnote in the history of the Web. Even the investment by major corporations, such as Google, Facebook, Microsoft and Yahoo!, has proved to be a mixed blessing

TABLE 2.2: Number of blog posts aggregated per year by Planet RDF (2006-2011).

Year	Number of Blog Posts	
2006	1725	
2007	1943	
2008	1678	
2009	973	
2010	590	
2011 (projected)	560	
		

to Semantic Web purists. The launch of **Schema.org**²⁰ to encourage the uptake of **Microdata** (Hickson, 2011) — arguably at the expense of support for RDFa — has been seen by some as a direct challenge to the Semantic Web vision while others consider it to be a pragmatic intermediary solution (Sporny, 2011; Bergman, 2011). Even prior to this development there have been those who question whether RDF is a necessary requirement for data semantics (Miller, 2009; Davis, 2009; Boutin, 2009).

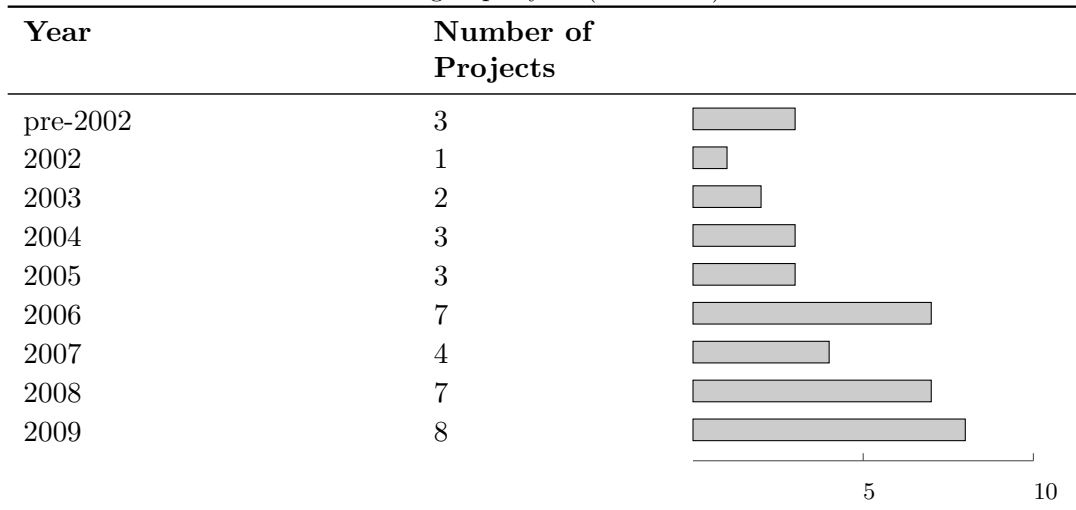
In conclusion, the Linked Data Era has seen the Semantic Web mature technologically in a number of ways but even more importantly it has seen some significant shifts in its perception. This change has not been universal. Debates about its fundamentals continue and they are far from being resolved. At the same time, much academic research has continued in a similar vein to that of the preceding Semantic Web Era. This mix of perceptions and motivations, and the speed with which they have occurred, has made it difficult for those outside the Computer Science community to identify Best Practices, or even a single underlying philosophy behind the vision, which in turn has contributed to a great diversity of approach in their own domains. We will therefore now turn to those applications of semantic technologies that are specifically concerned with the Cultural Heritage sector so that we can properly evaluate its levels of success to date.

2.3 The Semantic Web in Cultural Heritage

The Cultural Heritage sector — which also embraces **Galleries, Libraries, Archives and Museums (GLAMs)** — adds even greater diversity to that which we have already noted as being characteristic of Archaeology. Despite a common interest in the analysis, curation and dissemination of material culture, it is in some ways a series of interlocking communities, with different motivations, organisational structures and funding models.

²⁰<http://schema.org/>

TABLE 2.3: STCH Survey: New Cultural Heritage projects utilising semantic technologies per year (2002-2009)



Nevertheless, we will be unable to fully grasp the way that archaeologists have understood and applied semantic technologies without first situating their activities within this wider context. Archaeologists are by no means alone amongst Cultural Heritage practitioners in having experimented with semantic technologies, and the results of a cross-sector survey conducted by the author suggest that it has grown significantly since 2006 (Table 2.3). The question thus arises, was this interest stimulated by the new Linked Open Data philosophy or are other factors at play?

At risk of simplifying the overall complexity of the sector, this literary review will consider these questions first with regard to the GLAM community, followed by a specific discussion of archaeological adoption. This will bring to light some interesting contrasts with the wider picture discussed earlier in Section 2.2. It will be difficult to discuss either topic, however, without a brief digression on the CIDOC CRM to which we now turn.

2.3.1 The CIDOC CRM

First published in 2000 by the ICOM²¹ Documentation Committee (CIDOC),²² the **CIDOC Conceptual Reference Model (CIDOC CRM)** (Crofts et al., 2010) was accepted as an ISO Standard²³ in 2006 and is now the best known domain ontology in Cultural Heritage. However, the CIDOC CRM is not, and was never intended to be, a ‘Semantic Web’ project despite its close association with the field. Rather, its aim is to provide an explicit conceptualization of key concepts required for recording the

²¹International Council of Museums.

²²<http://cidoc.mediahost.org/>

²³ISO 21127:2006

interactions between human agents and cultural entities independently of a technical implementation. The CIDOC CRM's domain is no less than the full spectrum of Cultural Heritage knowledge, including Archaeology. Version 5 — the most recent — contains 90 entity types and 148 property types but is relatively compact and efficient given its extremely ambitious scope. In order to deal with the innate uncertainty of information about the past, it uses an 'event-based' model which describes interactions rather than states (Doerr, 2003). Yet two aspects separate it from the Semantic Web and set it closer to more traditional modes of KR.

The first is that CIDOC have historically been reluctant to use either persistent HTTP URIs or RDF for its definition, leaving implementation details to individual projects. The result has been a wide variety of formats — from the CIDOC natural language definition (Crofts et al., 2010) to relational databases (Hiebel and Hanke, 2008) to full-blown RDF ontologies (Oischinger et al., 2008). Consequently, integrating separate datasets has not been technologically straightforward, irrespective of the use of a common ontology. Fortunately, there have been more recent moves towards convergence including an official RDFS representation of the CIDOC CRM²⁴ which may go a considerable way towards improving this situation, although the natural language publication is still officially considered the authoritative definition.²⁵ In an even stronger advocacy of persistent HTTP URI use, a draft Recommendation has been published by CIDOC strongly encouraging museums to generate URIs for their holdings. The principal motivation for this is that “*LoD and the Semantic Web have no declared policy about who will create a URI for well-known things like Shakespeare or the Acropolis*” leading to “*a quagmire of competing authorized and unauthorized content about museum holdings*” (Crofts et al., 2011). This is not unqualified good news, for the belief that by creating their own URIs museums can prevent others from doing so is not only misconceived but also directly contradicts the Nonunique Naming Assumption. It is, nevertheless, an indication that CIDOC, and by extension ICOM, are now increasingly keen for museums to give their holdings a stable Web presence.

The second factor is that the CIDOC CRM is intended to be a *universal* domain ontology, albeit with the encouragement to extend it as appropriate, and its utility is largely considered to derive from its universal (or at least widespread) adoption. This is in strong contrast to the decentralized view espoused by the Semantic Web vision. Despite these structural differences, the CIDOC CRM is often seen as more or less interchangeable with 'digital semantics' by Cultural Heritage practitioners and continues to exercise considerable influence in the field. A more in-depth discussion of its relationship to the Semantic Web from the perspective of one its chief proponents is given by Doerr and Iorizzo (2008).

²⁴<http://www.cidoc-crm.org/rdfs/5.0.2/cidoc-crm>

²⁵http://www.cidoc-crm.org/definition_cidoc.html

The CIDOC CRM is by no means the only significant semantic infrastructure project in Cultural Heritage. Others include the **Pleiades**²⁶ gazetteer of ancient places (Gillies, 2011) and the **Heml**²⁷ and **LODE**²⁸ ontologies for describing historic events (Robertson, 2009; Shaw et al., 2009). These projects, perhaps better described as services, have begun to create a common URI framework by which cultural information can be expressed as RDF. They may additionally provide useful data such as geographic coordinates, but fundamentally they remain a backbone to which additional information can be attached, rather than datasets *per se*. All of these projects make their infrastructure openly available on the Web as their very utility is predicated on uptake by others.

2.3.2 Galleries, Libraries, Archives and Museums

GLAMs form a significant proportion of the Cultural Heritage sector and were particularly quick to begin exploring the potential benefits of the Semantic Web. Although chiefly concerned with the curation and dissemination of cultural materials, much of that content is derived from the archaeological process and the two cannot be considered entirely distinct. There are substantial differences nevertheless. GLAMs, whether publicly or privately owned, tend to have long-term goals in mind and funding is often in the form of capital investment in training and equipment, or increasing and improving their physical and information assets. Although public and scholarly engagement is usually a central part of their remit, this is rarely in an unmediated format. Most GLAMs consider it their duty to contextualise information about materials in their possession before releasing it and the raw data itself is rarely directly available. Likewise, many are either fully or partly funded by the sale of content derived from their holdings — such as post cards or digital images — and many are unwilling to risk making it freely available online for fear of undercutting an important revenue stream. Others may be restricted by licensing conditions imposed by the original contributor or even national legislation in the case of national institutions. The principles of openness and decentralization may therefore pose significant challenges for such organizations.

It is difficult to establish a clear, independent picture of how popular semantic technologies have been in any domain, but one method is to see how many related papers are published at relevant conferences. Fortunately it is possible to search the proceedings of the annual **Museums and the Web** conference by keyword.²⁹ Table 2.4 shows the number of all published papers that have keywords that refer either to a ‘semantic’ technology or Linked Data (titles are given in Appendix A). While at first glance it seems that there are spikes in 2002 and 2008, we must remember that conferences — and Humanities conferences in particular — have considerable latency. Abstracts are

²⁶<http://pleiades.stoa.org/>









²⁷<http://www.heml.org/rdf/2003-09-17/heml>

²⁸<http://linkedevents.org/ontology/>

²⁹<http://conference.archimuse.com/researchForum>

typically submitted six months (or longer) in advance, and frequently report on work already undertaken. If we assume a lag of 1 year then peaks in interest in 2001 and 2006–7 seems very much in line with picture which emerged in our discussion of the wider Semantic Web. Of particular concern, however, is that these peaks seems to be exceptional and interest has not been substantially maintained afterwards. Let us then examine the work that has been undertaken in this area and try to establish how well the chart reflects reality on the ground.

TABLE 2.4: Number of papers presented at Museums and the Web conference on semantic technologies (2002-2010)

Year	Number of Papers	
2002	3	
2003	0	
2004	1	
2005	1	
2006	1	
2007	3	
2008	6	
2009	2	
2010	2	

5 10

2001-2005

Prior to 2006 a small number of pioneering projects attempted to apply semantic technologies to GLAMs. Building on the **ARTISTE** project which provided cross-archival search capabilities for high-profile galleries using RDF (Addis et al., 2002), **SCULPTEUR** used an ontology-driven approach to provide adaptive search and visualisation mechanisms for 2D and 3D objects including digital images, 3D models, associated metadata, free text documents and numerical tables. Museum databases were mapped to the CIDOC CRM (with several extensions) and machine interoperability was achieved using a Z39.50³⁰ Search and Retrieve Webservice protocol (Addis et al., 2005). The ensuing **eCHASE** focussed on the creation of a toolset and framework by which third parties could contribute and draw multimedia entities from a semantically integrated network of repositories across Europe in order to increase exploitation of otherwise moribund resources (Sinclair et al., 2005).

MuseumFinland³¹ provided tools and services so that Finnish museums can present their collections online through a common interface backed by semantic technologies.

³⁰<http://www.loc.gov/z3950/>

³¹<http://www.museosuumi.fi/>

A bespoke domain ontology was developed for the project and contributing institutions made their data available as RDF/XML. The website allows users to search and browse cultural artefacts in a ‘follow-your-nose’ fashion via their properties (Hyvönen et al., 2005). The project came second in the 2004 Semantic Web Challenge,³² and formed part of a wider scheme by the Finnish government to semantically enable public web services. The scope of this research has continued to expand ever since (Mäkelä and Hyvönen, 2012).

The **Virtual Lightbox for Museums and Archives (VLMA)** project created a federated search tool that could be used across independent museum collections. The VLMA framework used the Jabber/XMPP³³ protocol to discover relevant material which was then returned as RDF via a Web query. The data, including images and metadata, was hosted independently and aggregated locally to be visualized in the user’s Web browser or in a Java WebApp.³⁴ As well as providing content, different repositories were able to recommend additional resources which could be browsed in turn (Fuchs et al., 2005).

These projects demonstrated the potential of semantic technologies for interlinking datasets but were in some regards still heavily influenced by traditional models of data aggregation. The use of Z39.50 and Jabber/XMPP, for instance, are strongly suggestive that RDF was seen as a data management format rather than a means of connectivity between separate resource bases. Furthermore, the data itself was not directly accessible in its semantic form over the Web, and URIs were generally not resolvable. Instead, RDF was manipulated behind the scenes and served either through webservices or human-readable interfaces. This is not to suggest that the projects themselves were ‘closed’. On the contrary, all three explicitly aimed to improve public and scholarly access to cultural materials that were otherwise difficult to access. Nonetheless, it seems that the idea of semantic technologies (including both URIs and RDF) as a dissemination mechanism had not been seriously considered.

A certain amount of theoretical discussion took place over the same period, generally through workshops and symposia which generated a combination of both excitement and scepticism. The DigiCULT project in particular brought together a panel of thirteen European Cultural Heritage experts to discuss the Semantic Web’s potential for the field Geser et al. (2003). Ross’s Position Paper for the event predicted, rather pessimistically, that *“the heritage sector is likely to be left behind because the financial rewards for creating the mark-up necessary to make the Semantic Web a reality are only evident to the commercial sector”* (Ross, 2003). Nonetheless, a panel discussion among the thirteen seemed to conclude that they *“would put their money on the Semantic Web”* (Steemson,

³²<http://challenge.semanticweb.org/>

³³<http://www.jabber.org/>

³⁴http://java.sun.com/developer/technicalArticles/tools/webapps_1/

2003). A particularly interesting aspect of this conversation is its emphasis on ‘interoperability’ and the need for a single common ontology (with the CIDOC CRM a strong candidate), rather than the need to increase openness or decentralization within the sector.

Post-2005

As Table 2.4 illustrated clearly, 2006–7 saw a significant surge in interest about the Semantic Web in the Cultural Heritage sector, leading to a raft of new projects. These may have been ‘born semantic’, derived from a single data resource, or aggregated in some way from a set of resources. Currently, very few projects are generated from scratch and when this does take place it is typically through heads-up digitization of an unstructured digital resource. A good example is the **Henry III Fine Rolls Project** which created a semantic database of complex relations between medieval manuscripts (Spence and Ciula, 2009). Legacy resources may or may not be held by the digitizing organization however, and even when they are, issues of ownership and copyright can have a substantial impact on a project’s ability or desire to provide public access to them. Large memory institutions, such as the **Rijksmuseum** in Amsterdam, the **British Museum** in London and the **Metropolitan Museum of Art** in New York, have largely focused on integrating their internal holdings using a predefined ontology (often the CIDOC CRM). This is typically used as the basis for a publicly accessible website while maintaining private access to the RDF and locally defined URIs (Aroyo et al., 2007; Smith and Miller, 2009; Oldman, 2009). Occasionally access to the data is restricted due to the use of a licensed thesaurus, rather than the content itself (van Ossenbruggen et al., 2007). Far fewer organisations have started to make their data semantically available over the Web although the **UK National Gallery**³⁵ is a noteworthy example.

A second set of services developing over this period act as semantic metadata aggregators, combining information from multiple heterogeneous sources. While the original content may itself be available under a variety of licensing conditions, the aggregator will usually point the user directly to the original resource. This greatly assists in the process of discovery while maintaining a decentralized approach to hosting and deferring issues of access to the original data provider. Examples of this approach include the **Cuneiform Digital Library Initiative (CDLI)**³⁶, **Contexta/SR** (Astudillo et al., 2008), **CLAROS**³⁷ (Kurtz et al., 2009), **Europeana** (Purday, 2010) and **Pelagios** (Barker, 2011).

Alongside these concrete implementations, the theoretical discussions have continued. Many of the published project reports indicate that, while semantic approaches can

³⁵<http://rdf.ng-london.org.uk/>

³⁶<http://cdli.ucla.edu/>

³⁷<http://explore.clarosnet.org/>

prove highly beneficial, mapping legacy datasets to ontologies — especially those with a high degree of complexity — remains a difficult task (Addis et al., 2005; Nussbaumer and Haslhofer, 2007; Mäkelä and Hyvönen, 2012). This has not been the only critique — the wider uncertainties expressed by DigiCULT do not seem to have been resolved either. The **Semantic Web Think Tank**,³⁸ a series of AHRC-funded workshops held in 2008, concluded that “*There is no coherent answer to the question ‘How do I do the Semantic Web?’ and almost no information with which to make an informed decision about technologies, platforms, models and methodologies.*” This was deemed to have created a gap between the vision and the reality of the Semantic Web “*which critically undermines the ability of the sector to move forward in a clear and constructive way*” (Parry et al., 2008).

If the Linked Data initiative was intended to simplify the Semantic Web then it seems that no-one had informed the GLAM community. Certainly, many of them continue to either ignore, or defer, the issues of decentralization and openness that seem so fundamental to it. But if that is the case, then what caused the significant rise in new projects in 2006–7? And just as importantly, is this pattern of peak followed by decline — that we have now observed amongst computer scientists, heritage specialists and the general public — just part of the **Gartner Hype Cycle** (Fenn, 1995), or must we conclude that the Semantic Web is a failed experiment and turn our attention elsewhere? Unfortunately, identifying the extent to which individual projects were driven by divergent technical philosophies is a difficult attribute to measure from published reports alone. In order to establish the foundations for a more in-depth analysis, we need to conduct a survey among those employing semantic technologies across the Cultural Heritage sector. The results of such a survey will be considered in detail in Chapter 3, but first we will complete this literature review by looking at the application of semantic technologies in Archaeology itself.

2.3.3 Archaeology

The international **Computer Applications and Quantitative Methods in Archaeology (CAA)** conference has also seen a good number of papers related to semantic technologies since 2006, often related to the CIDOC CRM (Table 2.5; Titles are listed in Appendix A). Given the apparent regularity of all our previous results, it is therefore surprising that they tell a different story. Archaeological interest in the Semantic Web appears to have grown since 2001 — with a tell-tale spike in 2006–7 — but has yet to show any significant sign of abating. What is frustratingly difficult to establish is whether this is caused by an underlying trend towards growth, or whether we are seeing the same effect as in other domains until 2009 when, instead of continuing to decline as

³⁸<http://culturalsemanticweb.wordpress.com/>

TABLE 2.5: Number of papers presented at CAA on semantic technologies (2001-2011)

Year	Number of Papers	
2001 (proceedings)	0	
2002 (proceedings)	1	■
2003 (proceedings)	4	■■■■
2004 (proceedings)	5	■■■■■
2005 (conference)	4	■■■■
2006 (conference)	9	■■■■■■■■■
2007 (conference)	12	■■■■■■■■■■■
2008 (conference)	9	■■■■■■■■■
2009 (conference)	15	■■■■■■■■■■■■■
2010 (conference)	12	■■■■■■■■■■■
2011 (conference)	12	■■■■■■■■■■■

in other domains, it starts to pick up again for some reason. It is possibly not coincidental that the first CAA session dedicated to semantic technologies was held in that year, followed by a **CAA Semantic Special Interest Group**³⁹ in 2010. So have archaeologists found a successful semantic recipe, or are they blindly pursuing a technology which others seem, after a short period of appraisal, more willing to abandon?

As noted previously, the field of Archaeology contrasts strongly with GLAMs in several ways. In particular, it is frequently dependent on fixed, often short-term, funding with which to produce a pre-specified output that rarely includes more than rudimentary dissemination objectives. As a result, archaeologists tend to have less time and fewer resources to invest in data management technologies. This picture is highly diverse however, with many long-term excavations, commercial units and international missions having both the resources and the desire to make strategic digital investments. Let us then briefly review the work carried out in this area to date so that we can begin to see whether archaeological approaches have been substantively different from those of others working in Cultural Heritage.

2001-2005

One of the earliest pioneers in this area was the doctoral research of Alani (2001), undertaken in collaboration with **Royal Commission on the Historic and Ancient Monuments of Scotland (RCAHMS)**.⁴⁰ While not strictly ‘semantic’ in the sense of using URIs, it employed both spatial and semantic distance metrics in order assist

³⁹<http://groups.google.com/group/caa-semantic-sig>

⁴⁰<http://www.rcahms.gov.uk/>

search and retrieval over data from RCAHMS tagged with concepts from the Getty Research Institute's Art and Architecture Thesaurus (AAT)⁴¹ and Thesaurus of Geographic Names (TGN).⁴² More recently the RCAHMS has also supported the doctoral work of Byrne (2009) who developed Natural Language Processing (NLP) techniques to extract RDF descriptions of archaeological events from free text in the RCHAMS database, as well as documents from the National Library of Scotland and the National Museum of Scotland.

Much of the research in this period, both theoretical and technical, is influenced by, or based upon, the CIDOC CRM. Crofts (2004) describes a project that used the the CRM to successfully integrate a range of heterogeneous archaeological resources inherited by Geneva's newly-established *Direction du Patrimoine et des Sites*. The process is entirely based on XML transformation (using XSLT) and an Oracle database however, and neither URIs nor RDF are mentioned. Perhaps the first explicitly archaeological project to do so was **VBI-ERAT-LUPA**⁴³ which created a centralized RDF repository of CRM-compliant RDF describing the iconography and epigraphy of Roman stone monuments from multiple data sources (Doerr et al., 2004). It is unclear whether this forms the back-end of the current website, as the RDF itself is not directly available, but all data is available through a human-readable search interface.

Arguably the most significant development in this period was the specification of a CIDOC-CRM extension for archaeological information by English Heritage, entitled **CRM-EH** (Cripps et al., 2004). A large amount of work has since been done by the **STAR** project to explore the possibilities this ontology affords and develop a variety of experimental interfaces both to convert data into this format and to query it (Tudhope, 2008; May et al., 2009). The same team has gone on to work with the Archaeological Data Service (ADS) on a further project (**STELLAR**) that has created more mature tools that enable non-expert users to produce datasets that conform to CRM-EH (Tudhope et al., 2011a).

Post-2005

In the years following 2005, semantic technologies have attracted a good deal more attention, prompted at least in part by regular workshops explaining the the relevance and structure of the CIDOC-CRM. The projects themselves display a surprising degree of variation, and many of them are multifaceted. For the purpose of this review however, they can be loosely categorised as technical platforms, archaeological research projects and aggregators, although most combine features from two or more of these categories.

⁴¹<http://www.getty.edu/research/tools/vocabularies/aat/>

⁴²<http://www.getty.edu/research/tools/vocabularies/tgn/>

⁴³<http://www.ubi-erat-lupa.org/>

Several projects have developed a generic technical platform intended to ingest, manage and query arbitrary archaeological information. The **NavEditOW** system is a server and suite of editing tools principally developed for hosting an e-library on pre- and proto-history in Italy (Bonomi et al., 2007). **ArcheoInf**⁴⁴ is building a ‘mediator’ application that displays information from multiple heterogeneous excavation datasets within a common visualization framework using SKOS and the CIDOC CRM (ArcheoInf, 2008). Alongside these explicitly semantic platforms it is interesting to note that several other platforms for archaeological and historical data, including the **Archaeological Recording Kit (ARK)**⁴⁵ (Eve and Hunt, 2008), **Intrasis**⁴⁶ and **Heurist**⁴⁷ (Johnson, 2009) all use a Subject-Predicate-Object data format in order to maximize structural flexibility across widely divergent use-cases. While they do not make use of URIs, RDF or formal ontologies, the leap towards incorporating them seems small if the intention were present. Many of these platforms are Web-based, permitting remote access to the data in principle, even though there are only a few examples of them serving data openly in practice. A radically different solution to the problem of data management in an archaeological context is that proposed by Isto Huvila. He argues for a participatory archive, based on the **Semantic MediaWiki**⁴⁸ platform, that could provide “*decentralised curation, radical user orientation, and contextualisation of both records and the entire archival process*” (Huvila, 2008).

A second category of project focusses on a specific archaeological excavation or research agenda. The reasons for employing semantic technologies vary considerably. For example, the **Athenian Agora Excavations**⁴⁹ have begun to record excavation data as RDF in order to have better associative indexing between finds, features and other archaeological outputs. The pottery at **Ilion (Troy)**⁵⁰ is being experimentally published as RDFa in order to make them easier for digital archaeologists to query from directly within the Webpage (Heath and Tekk  k, 2009). While these projects have a strongly disseminatory angle, others are more focussed on the inferencing capacities of RDF. The **Tracing Networks**⁵¹ project has used semantic technologies to combine disparate data on trade and exchange networks across the ancient Mediterranean. The **ArcheoKM** system, developed for an industrial archaeology site at the former Krupp steel works, uses the SWRL⁵² Rule Language in order to automatically categorise spatial entities (Karmacharya et al., 2010).

Finally there are a series of projects principally concerned with the aggregation and further dissemination of archaeological material. Some of these have a history that long

⁴⁴<http://www.archeoinf.de/>

⁴⁵<http://ark.lparcnaeology.com/>

⁴⁶<http://www.intrasis.com/>

⁴⁷<http://heuristscholar.org/heurist/>

⁴⁸<http://semantic-mediawiki.org/>

⁴⁹<http://www.agathe.gr/>

⁵⁰<http://classics.uc.edu/troy/grbpottery/>

⁵¹<http://www.tracingnetworks.ac.uk/>

⁵²<http://www.w3.org/Submission/SWRL/>

predates their interest in digital semantics but RDF and stable URIs have become part of a suite of technologies they use to achieve their goals. The **Nomisma**⁵³ project uses RDF to combine data on Greek coin hoards so that their origins and distribution can be visualized online and readily integrated with other ancient world resources (Heath, 2011). The **SPQR**⁵⁴ project also makes epigraphic and papyrological data available in RDF format (Jackson, 2011). A number of major national archiving initiatives — including the UK **Archaeology Data Service (ADS)**⁵⁵ and **Portable Antiquity Scheme (PAS)**,⁵⁶ the US **Digital Archaeological Record (tDAR)**⁵⁷ and the German Archaeological Institute’s **Arachne**⁵⁸ database — have also experimented with either exposing or storing data in semantic formats although none of them have made it central to their operations (Tudhope et al., 2011a; Krempel, 2011).

In a further contrast to the GLAM community, there have been few collaborative attempts by archaeologists to evaluate the value of semantic technologies to the domain. The CAA Semantic Special Interest Group mentioned earlier has functioned predominantly as a forum for the exchange of information and practical tips rather than deeper philosophical reflection. Such reflection has certainly taken place however. Within the context of archaeological journal publication, Richards (2006) has argued for the importance of ontologies in assisting with the search and retrieval problems faced by self-archivers and e-print journals.

Within archaeology agreed common data structures do not exist and their development is the greatest obstacle on the road to the development of an archaeological semantic web. Bell and Eiteljorg (2006) agree that XML is just a technology and that it is not the issue of prime importance. Rather it is an agreed ontology that is required, defining terminologies, data structures and discipline-specific schema. (Richards, 2006)

Richards notes the emergence of the CIDOC CRM as an emerging standard core ontology, and in particular the work of Cripps et al. cited above. Yet he highlights the fact that the CRM, in and of itself, “*does not specify those terminologies and structures which must be used if data are to be meaningfully agglomerated*”, and appears to advocate a far more lightweight process of semi-automated XML tagging with unique identifiers, especially in those cases “*where there is agreement on classifications and terms*”. A follow-up article in 2009 suggests that he has not changed his mind, emphasising the power of a ‘What’, ‘When’ and ‘Where’ approach employed by the **Archaeotools**⁵⁹

⁵³<http://nomisma.org/>

⁵⁴<http://spqr.cerch.kcl.ac.uk/>

⁵⁵<http://archaeologydataservice.ac.uk/>

⁵⁶<http://finds.org.uk/>

⁵⁷<http://www.tdar.org/>

⁵⁸<http://www.arachne.uni-koeln.de/>

⁵⁹<http://archaeologydataservice.ac.uk/research/archaeotools>

project (Richards, 2009). A further passage in the 2006 article is also pertinent to the current discussion.

The argument that we should...structure both publications and archives to be machine-readable [is separate from] the debate about open access and e-repositories. The Semantic Web does not require that information is freely available. Indeed its development will undoubtedly be driven by the commercial advantages to advertisers and sellers. Nonetheless there is an area of overlap in that whilst the Semantic Web does not require open access, it would arguably be greatly facilitated by it. If all archaeologists placed their articles in an online repository they would undoubtedly be easier to find. Therefore the Semantic Web also needs to be considered alongside current debates about scholarly publication. (Richards, 2006)

Two important points stand out. The first is that Richards, like both Berners-Lee (1998b) and Ross (2003) before him, suggests that the commercial sector is likely to be the main beneficiary of the Semantic Web. It is not entirely clear why this need be the case other than a general assumption that financial reward is the most common engine of innovation. In contrast, a requirement for openness and decentralization may well cause industry to steer clear. Either way, we must not forget that the overwhelming quantity of archaeological output is produced by commercial archaeological organizations. Nevertheless, the repeated view that commerce may be the natural domain of digital semantics is striking. The second significant statement is his opinion that, "*whilst the Semantic Web does not require open access, it would arguably be greatly facilitated by it.*" The degree to which this is in fact the case will become a central theme of this thesis.

2.4 Conclusions

What can we now conclude about archaeologists' experiences with semantic technologies and how do they compare with those of Computer Scientists and the GLAM community? The first — apparent from the Google search statistics, the Planet RDF aggregation count and the number of papers published in the proceedings of *Museums and the Web* — is that general interest in the Semantic Web peaked shortly after 2008 and has been on the wane ever since. This is certainly not to say that it has become irrelevant, but neither can it realistically be considered a strong growth area, at least in its present incarnation. Curiously, Archaeology appears to have followed this trend only until 2009 when, rather than falling off, interest in semantic technologies seems to have picked up again.

The second is that two visions have been driving Semantic Web development, although the terminology is distinct for neither of them. One, which finds notable expression in Berners-Lee's earlier writings, and again after 2006, is often described as **Linked Open Data**. It emphasises openness, decentralization and the interconnection of independent resources. For want of a pre-established term, we shall refer to the other vision as **Mixed-Source Knowledge Representation (MSKR)**. This perspective was particularly prevalent between 2001–5 and continues to have a strong following. It tends to focus on formal description and reasoning over multiple heterogeneous datasets, typically within a closed digital environment. Although these two visions are very different, it is important to stress that they have become strongly intertwined in both literature and practice, and any given Semantic Web project is likely to have been influenced to some degree by both of them. What is less certain is whether Cultural Heritage projects have been particularly influenced by one or other of them, and hence whether their success (or otherwise) is coloured by a specific set of beliefs about what semantic technologies are intended, and able, to achieve. In the following chapter we will discuss the results of a comprehensive survey intended to answer that question.

Chapter 3

Survey — Evaluating Semantic Technologies for Data Publication in Cultural Heritage

3.1 Overview

We have seen that a wide range of Cultural Heritage projects have made use of semantic technologies. Their variety however, along with the diversity of formats in which they have been reported, makes it very difficult to compare them easily and thus identify common trends. This chapter discusses the **Evaluating Semantic Technologies for Data Publication in Cultural Heritage (STCH) Survey** undertaken to establish a broad overview of how semantic technologies have been applied across a range of Cultural Heritage activities. The aim of the survey was to evaluate the costs and benefits of publishing data with semantic technologies in Cultural Heritage projects and institutions, and in particular whether they appear to follow the distinct philosophies of MSKR and Linked Open Data identified in Chapter 2. As no clear terminology exists to identify these different perspectives the survey focussed on a number of specific aspects pertinent to them, including:

- Which semantic technologies were adopted.
- Motivations for using such technologies.
- The relative cost of using them.
- Assessment of their utility.

3.2 Method

The study was carried out as a questionnaire among an invited set of participants who had worked, or continue to work, on Cultural Heritage projects that utilise semantic technologies to publish (rather than simply consume) data. The term ‘project’ is used loosely here, and includes PhD theses and long-term institutional initiatives. Where possible, two respondents were invited from each project¹ — a computing specialist and a humanities specialist. Although this turned out to be practical for only a fraction of the projects, the differences in responses raised interesting issues about perceptions of the technology, as well as the robustness of surveys such as this one. These issues will be discussed in due course. In all cases, respondents were asked to self-identify which field(s) they consider themselves to be in.

When identifying prospective participants the definition of ‘semantic’ was deliberately left vague, partly due to the aforementioned unclarity in the literature and largely because very few projects implement all four of the ‘Linked Data Rules’ (Berners-Lee, 2006). A project was initially considered eligible to participate in the survey if it exhibited any of the criteria from List A and any of the criteria in List B below.

A: Cultural Heritage

- Project content pertains to Archaeology
- Project content pertains to Galleries, Libraries, Museums or Archives
- Project content pertains to Classics
- The project is explicitly designated a ‘Cultural Heritage Project’

B: Semantic Technologies

- The project makes use of URI identifiers
- The project makes use of RDF
- The project makes use of RDF-based ontologies
- The project makes use of technologies specifically designated as ‘semantic’.

The electronic questionnaire asked the respondent to describe the motivations of the project, along with the technologies used, and to provide a brief evaluation of them. The full text is provided in Appendix D. The great majority of questions were either

¹In one case three respondents replied.

single- or multiple-choice (radio buttons or check boxes), with a small number of numeric and text fields. The survey was intended to take approximately 20-30 minutes to complete in order to provide a reasonable amount of detail without deterring potential respondents. A small number of well-known ‘successful’ non-Cultural Heritage projects were invited to participate as a control group. Two agreed to take part and completed the survey² but unfortunately the small sample made it impossible to derive meaningful statistical comparisons. A list of participating projects is provided in Appendix E. For purposes of data protection, the names of respondents themselves is not given and the free-text responses listed in Appendix F have been redacted where necessary to maintain anonymity.

3.2.1 Participants

79 individuals from approximately 70 different projects were identified and invited to participate altogether. The projects cover the great majority — if not all — known and relevant projects undertaken globally in this time period. Targets of the survey were individuals who had worked on projects drawn principally from papers published in journals such as *The Semantic Web Journal*, *The Journal of Web Semantics* and *Internet Archaeology*, or the conference proceedings of *ISWC*, *ESWC*, *CAA* and *Museums and the Web*. These were supplemented to a lesser a degree by projects identified through mailing lists or other branches of the research network, and occasionally through recommendation by other survey participants. As a result, academic and institutional projects are more likely to have been included than initiatives by commercial organisations who rarely publish in academic circles. It is notable however that the author is aware of virtually no such initiatives. This is not to say that they have no interest in the subject — and conference sessions on the topic often see attendance by representatives from the private sector — but given the common assumption by Berners-Lee (1998b), Ross (2003) and Richards (2006) that semantic technologies would likely be pioneered by commercial interests, it seems a significant observation.

67 individuals agreed to participate — with some respondents reporting on multiple projects — representing 57 individual projects altogether (Appendix E). 12 projects had multiple respondents. Once all the responses had been returned, the projects were additionally filtered to exclude those which did not use URIs, in line with Berners-Lee’s requirement that “*if it doesn’t use the universal URI set of symbols, we don’t call it Semantic Web*” (Berners-Lee, 2006).³ This criterion was used in order to ensure that the results of the survey bore maximum relevance to questions outlined at the beginning of Chapter 1 and establish a formal scope to the survey. This reduced the final results to a set of responses from 40 projects, upon which the following discussion is based.

²GeoNames, BBC.

³Persistence, dereferencing and HTTP were not a requirement.

3.2.2 Procedure

In compliance with the University of Southampton’s Ethics Policy,⁴ all survey documentation and procedures were vetted prior to use by the University of Southampton’s Electronics and Computer Science Ethics Committee. Individuals were contacted by email and asked if they were willing to take part. A copy of the Participant Information Sheet (Appendix B) and Consent Form (Appendix C) was provided. If only one person was known from the project they were asked if they could suggest a colleague who represented the other specialism (Humanities or Computing). If the individual consented to participate (by returning the signed Consent Form by mail or from a personally identifiable email address), they were directed to a website hosted by the Survey Monkey webservice⁵ to fill out the questionnaire. Project and respondent details were sent by the respondent, along with the Consent Form, and stored independently of the survey on an access-controlled university server. The respondent was given a codename by which they could be identified in the survey. An anonymized summary of the responses is provided in Appendix F.

3.2.3 Bias and Limits of Scope

Any survey is inevitably both a snapshot in time and subject to the interests, bias and limitations of its compiler and this one is no exception. The author is a UK-based researcher and, despite efforts to be as global in perspective as possible, the high proportion of projects included that are based in Great Britain, and to a lesser extent Europe, are almost certainly partly (if not entirely) due to the limitations of his own research network (Table 3.1). Smaller projects based outside Europe are especially likely to have been missed.

Despite this likely bias toward UK-based research however, the author suspects that the figures do to some degree reflect the prevailing research interests of these regions. Europe (and especially the UK) has a particularly strong history of research and investment in

⁴http://www.soton.ac.uk/inf/ethics_policy.html

⁵<http://www.surveymonkey.com/>

TABLE 3.1: STCH Survey: Projects by region




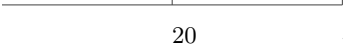




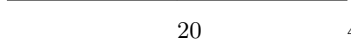
Project base	Number of Projects	
UK	13	
Europe (non-UK)	15	
Americas	12	
		

TABLE 3.2: STCH Survey: Projects by domain

Project domain	Number of Projects	
Archaeology	17	
Museums and Archives	15	
‘Culture’	5	
Other	2	
		

Heritage Computing, both in terms of national and EU-level funding. Likewise, many of the major centres of Semantic Web Research, such as DERI, Southampton and Berlin, are within European countries. There is comparatively less engagement with either field in Asia, South America and Africa, although their almost complete absence from the survey (there is a single South American project) is still surprising.

Time is another dimension upon which we must orient the survey, and accept both that older projects may have disappeared into obscurity and that more recent ones may not have established enough of a public profile to be identified. This survey took place in July and August 2010, and thus the practices reported by the respondents probably best reflect a reality lasting up until the preceding six months, although, once again, it is the author’s suspicion that projects that post-date that time have not significantly diverged from the practices recorded.

A third area of bias to consider are the individual fields of Cultural Heritage activity reported on. When asked ‘which term best describes the nature of the project data?’, responses were fairly evenly split between Archaeology and GLAMs (with a small remainder of ‘Cultural’ and ‘Other’ data) (Table 3.2). These communities, though intertwined in many ways, tend to operate independently with different strategic priorities. From the purely pragmatic perspective of this research, it was helpful to see that Archaeology was not under-represented and that there is a similar body of research within the GLAM sector for purposes of comparison.

Finally, comparing responses from different people working on the same project indicates that the data cannot be assumed to be trustworthy in all particulars. In particular there was a tendency for those who were not directly engaged in the technical elements of the project to be unaware (or have forgotten) about the use of specific technologies. For instance, they might claim that RDF was used but not URIs (without being aware that the former requires the latter). Thus, the use of specific technologies is more likely to be under-represented than over-represented and responses in the negative to questions of this nature cannot be regarded as a definitive ‘no’. A presumption towards the affirmative has been made where two such responses were conflict. Where multiple

TABLE 3.3: STCH Survey: Respondents by specialism





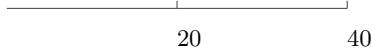



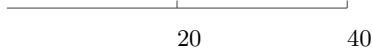
‘Do you consider yourself to be a specialist in:’ (single choice)	Number of Respondents	
Computing	16	
Humanities	6	
Both	17	
Neither	1	
		

TABLE 3.4: STCH Survey: Respondents by role

‘Your role in the project’ (multiple choice)	Number of Respondents	
Developer	25	
Project Director	24	
Content Curator	8	
		

responses were given to questions that required evaluation of a technology or similarly scalar response, an average value has been taken for the purposes of this analysis. A summary of all responses is provided in Appendix F. A digital version of the original data, edited only to maintain anonymity, will be deposited along with the electronic version of thesis in the University of Southampton ePrints repository.⁶

As regards the biases of the participants, the respondents themselves largely self-identified as Computing specialists, with or without additional humanistic skills (82.5%). In contrast, just 58% claimed to have a Humanities skillset (Table 3.3). This division is mirrored in the respondents’ roles in their projects, which were predominantly Developers and Project Directors (sometimes the sole member of the project), with only a handful of Content Curators (Table 3.4). This distribution seems to be largely a byproduct of the fact that i) technically-oriented staff were more likely to publicize their results through the technical journals and conferences through which they were identified, and ii) the occasional tendency for Humanities specialists to refer the author on to a Computing specialist colleague ‘who knows more about these things than I do.’ This may suggest that answers to technical questions are more likely to be correct when answered by Computing specialists (i.e. in the majority of cases), although perhaps we should be equally wary of their evaluation of benefits to the heritage sector.

⁶<http://eprints.soton.ac.uk/>

3.3 Findings

3.3.1 The Projects

Before discussing which semantic technologies have been used and how, it will be worthwhile looking at a few basic statistics about the projects themselves. Starting with the duration of the projects, we can see that many of them are comparatively long by Humanities standards. The great majority of the projects (82.5%) are over one year in duration and, as half of the projects are open-ended, the figures somewhat under-represent the actual figures (Tables 3.5 and 3.6).

Project staff size seems to be more evenly spread, with approximately one third having five staff or fewer, a third having 6–10, and a third having more than 10 staff (Table 3.7). The reasons for this emphasis on scale are suggested by three further sets of figures. When indicating the number of institutions involved (Table 3.8), the number of datasets merged (Table 3.9) and the number of dataset schemas merged (Table 3.10), over two-thirds of the projects have three or more. It is difficult to be certain of the significance of this across such a diverse range of projects, but intuitively it suggests that semantic technologies become desirable precisely when data integration cannot be solved with a bilateral mapping between two systems. In other words, they are predominantly seen as

TABLE 3.5: STCH Survey: Projects by duration







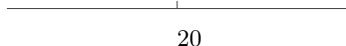
Project Duration (months)	Number of Projects	
0–12	8	
13–24	11	
25–36	7	
37–48	8	
49–60	2	
61+	4	
		

TABLE 3.6: STCH Survey: Open-ended vs. Fixed-term




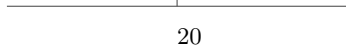
Project is/was:	Number of Projects	
Open-ended	20	
Fixed-term (ongoing)	10	
Fixed-term (finished)	10	
		

TABLE 3.7: STCH Survey: Projects by people involved

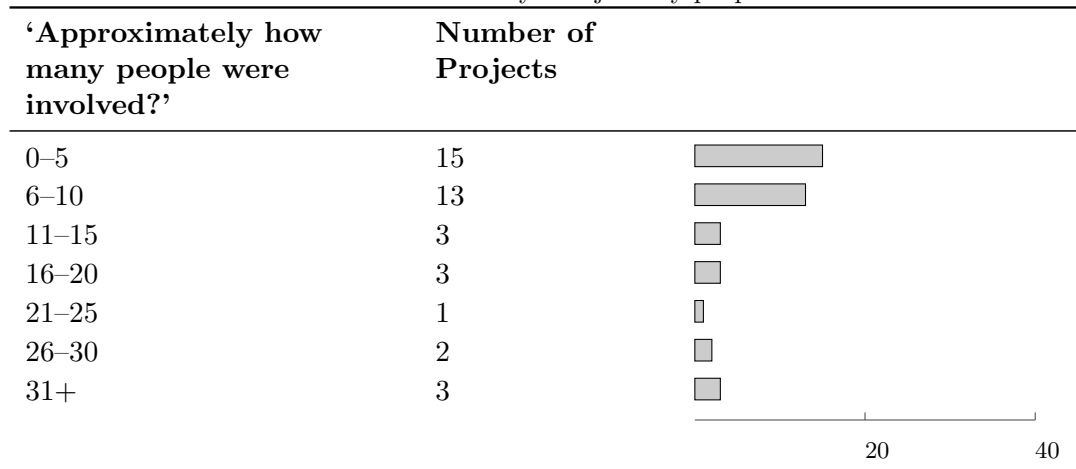
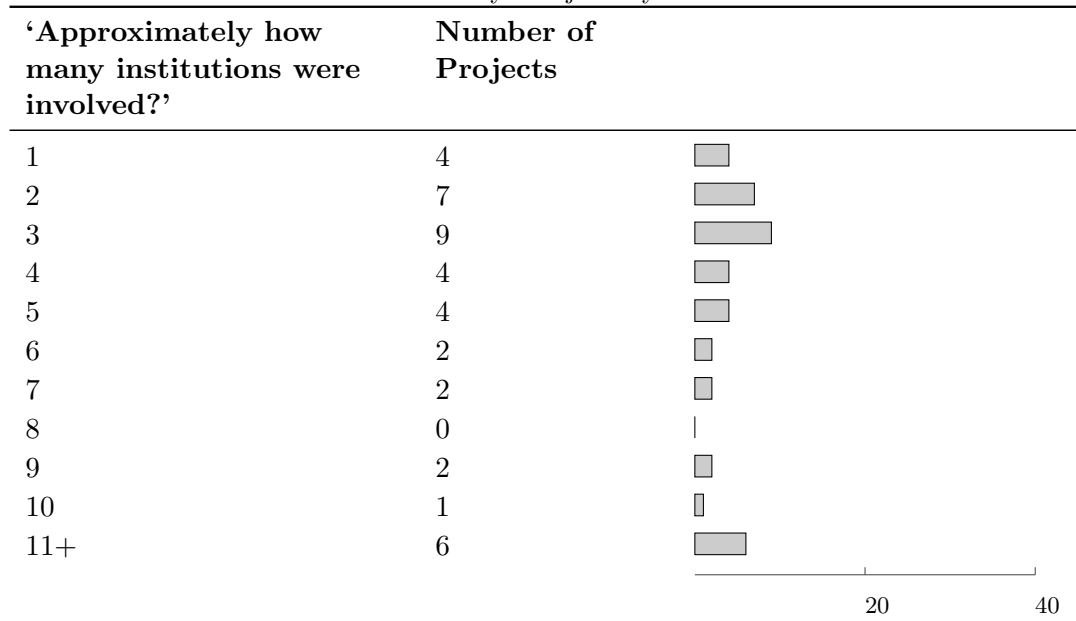


TABLE 3.8: STCH Survey: Projects by institutions involved



a data integration strategy for complex data environments. The implications of this are not only that they are adopted by medium-to-large-scale projects as we have just seen. It also suggests that they are less commonly being used to *provoke* incoming connections from others (even if this is seen as a desirable by-product) so much as to solve specific data integration problems between pre-identified data sets.

TABLE 3.9: STCH Survey: Projects by datasets integrated

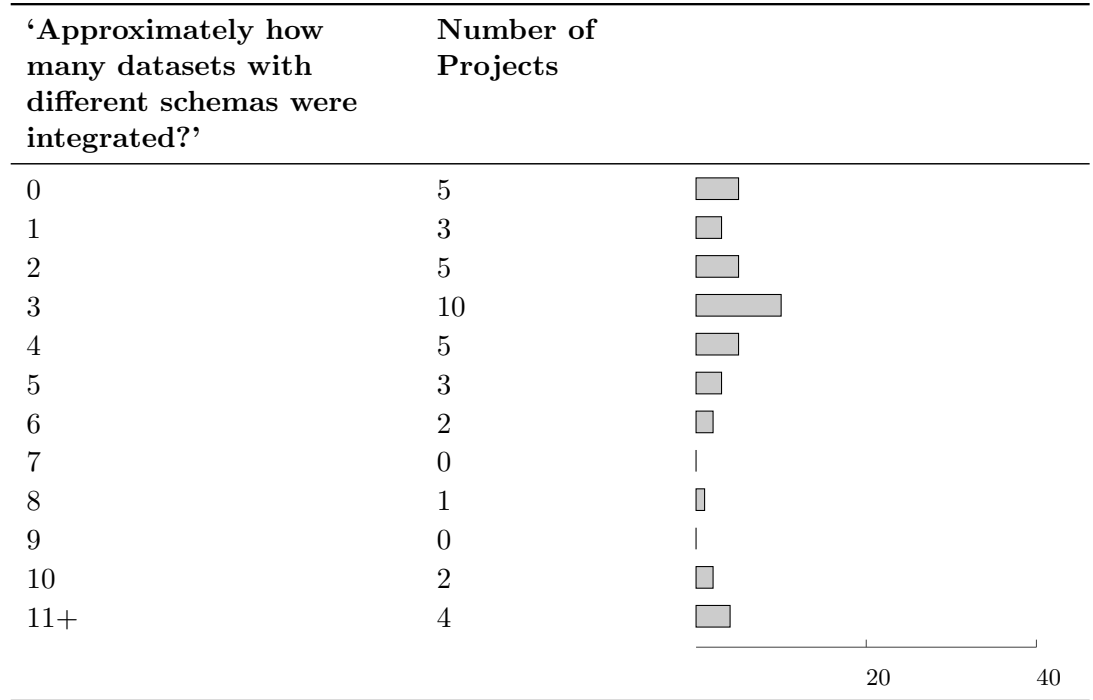
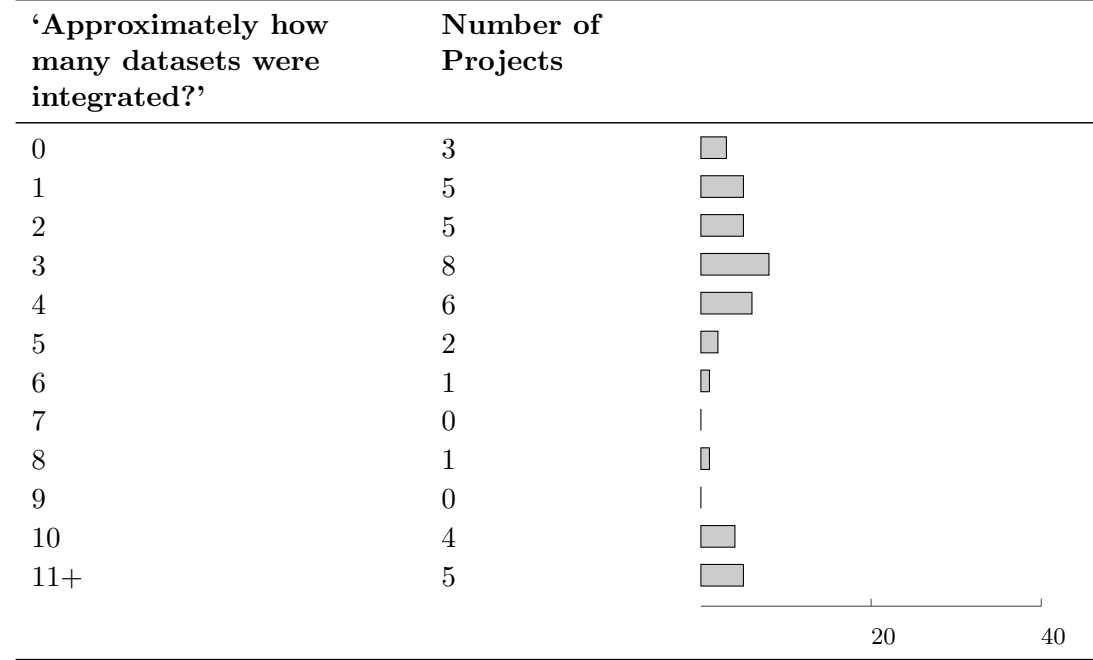


TABLE 3.10: STCH Survey: Projects by datasets with different schemas integrated

3.3.2 Intentions

Are these then predominantly a set of closed, centralizing initiatives, indicative perhaps of an MSKR philosophy? Not necessarily. There are at least some indications that a wider user-base is being sought. In response to the question ‘who are your intended consumers?’ only 10% responded that they were restricted target groups such as project partners, and two thirds said they had no target group (Table 3.11). Of course, this may not always have been a full-throated endorsement of openness — it may just as well reflect a lack of any thinking in this direction at all. In fact openness *per se* often seems not to have been a consideration or to have fallen foul of sustainability issues. Only 60% of the projects said that their data is publicly available (Table 3.12).

Questions about the relative priorities of the projects also deliver interesting results. When asked, in the opinion of the respondent, whether data complexity or data quantity took precedence for the project, the results were evenly balanced, with a considerable proportion (22.5%) giving no answer at all (Table 3.13). This is perhaps to be expected as both are important goals for archaeologists and heritage curators alike. In contrast, when asked whether data utility⁷ or data integrity⁸ took precedence, the answer was overwhelmingly (70%) in favour of utility with a further 7.5% giving no answer (Table 3.14).

⁷ “That the output could be put to better or different use than with other technologies.”

⁸ “That the output was an accurate representation of the semantics of the input.”

TABLE 3.11: STCH Survey: Projects by intended user group(s)




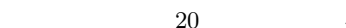
‘Who are your intended consumers?’ (single choice)	Number of Projects	
No restriction, no target group	27	
Unrestricted target group(s)	13	
Restricted target group(s)	4	
		

TABLE 3.12: STCH Survey: Project data availability



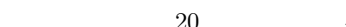
‘Is the data publicly available?’	Number of Projects	
Yes	24	
No	16	
		

TABLE 3.13: STCH Survey: Projects' priorities — data complexity vs. data quantity




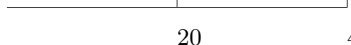
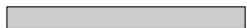


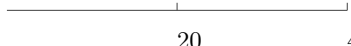
'In your opinion, which of the following two criteria took precedence?'	Number of Projects	
Complexity over quantity	16	
Quantity over complexity	15	
Neither	9	
		

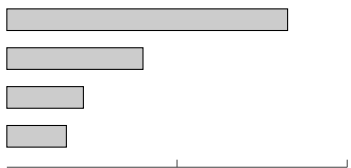
TABLE 3.14: STCH Survey: Projects' priorities — data utility vs. data integrity

'In your opinion, which of the following two criteria took precedence?'	Number of Projects	
Data utility	28	
Data integrity	9	
Neither	3	
		

Does this indicate a trend towards a Linked Open Data paradigm? It seems an extremely curious result for two reasons. First of all the heritage sector is typically thought of as one in which the creation and curation of accurate data are paramount. Archaeologists frequently cite the need to publish 'good quality data' as a reason for not publishing it in its raw form and 'utility' is rarely used as a justification for Humanities scholarship. Secondly, semantic data is explicitly supposed to be an improvement in quality, helping to avoid the difficulties of ambiguity that plague other formats. Of course it could be argued that this is precisely so that the data becomes more useful (and we should remember that the respondents are not indicating that data integrity is unimportant) but nevertheless, such a strong emphasis on utility does seem to require an explanation. Significantly, it does not seem to be influenced by the respondents' background or role although Computer Scientists were slightly *less* likely to prioritize utility. There was also no correlation with either archaeological or GLAM projects, nor with those that began before or after 2006.

One possibility is that it is a consequence of the source data. Table 3.15 shows that it is overwhelmingly converted legacy content that is being used rather than data that is 'born semantic'. There is a kind of Catch-22 involved here. The need to semantically format data arises precisely in those cases where it is not already machine-readable. But therefore no generic algorithm can convert it perfectly, by definition, and so perfect data integrity is unattainable by automated means. Manual conversion can naturally achieve far better results (issues of source data ambiguity, variable precision and human

TABLE 3.15: STCH Survey: Project data sources

Project data source:	Number of Projects	
Legacy digital data	33	
Legacy paper-based data	16	
Born semantic	9	
Controlled vocabulary only	7	

error notwithstanding) but is often unworkable in practice, especially with large data volumes. Could it be that those projects working with legacy data are simply forced to accept ‘good enough’? In fact the answer again seems to be no. There is no discernible correlation between the use of legacy data (whether digital or paper-based) and a prioritization of utility.

There are in fact just two variables which appear to reflect a similar distribution. The first is the employment (or not) of text-mining algorithms, which seems entirely consistent with the ‘good enough’ hypothesis as this also an inherently imprecise process. 86.7% of projects which used text-mining also prioritized utility, but as they make up just 37.5% of the entire survey they cannot entirely explain the trend. The second factor, which was not expected, is whether the project is fixed-term or open-ended. The projects surveyed were almost exactly divided by those which were continually ongoing and those which were fixed-term (whether ongoing or completed). Of the open-ended projects, 80% prioritized data utility (with one giving no answer). In contrast, only 65% of fixed-term projects were, and of these, about 50% used Text-mining as a method. The disparities do not end there. Open-ended projects were much more likely to have no restriction on their intended consumer base (fixed-term 50%, open-ended 85%). In contrast, fixed-term projects were much more likely to have a Humanities expert as part of the data conversion team (fixed-term 60%; open-ended 25%), use the CIDOC CRM (fixed-term 65%, open-ended 30%) and least likely to have publicly available data (fixed-term 45%, open-ended 75%). This last statistic is not an effect of data ‘going dark’ after the completion of the project — the same result was observed for both completed and uncompleted projects. In short, it seems that fixed-term projects are more likely to be concerned with data modelling and integration within a closed environment, whereas open-ended projects seem more liable to prioritize access and utility. This in turn appears to have had some important effects on their choice of technologies, as we shall see.

3.3.3 Semantic Technologies Employed

Having established that there may indeed be differing intentions behind these projects, we now turn to the technologies they employed. Initially it would seem that data-linking of all types — internal, incoming and outgoing — was a common goal, although by no means in every project (Table 3.16). Fixed-term projects were only slightly more likely to have internal or outgoing links and neither was more or less likely to encourage incoming links.

Of the mainstream semantic technologies there is an unsurprising fall-off in uptake dependent on its place in the Semantic Web Layer Cake (Table 3.17).⁹ Some flavour of RDF was used by about 75% of the projects whereas SPARQL, OWL and SKOS were adopted by only about half of the projects. Similarly, although HTTP URIs were relatively common (90%), resolvable and persistent URIs were considerably less so (60% and 55% respectively) (Table 3.18). When projects used resolvable URIs they were almost always hosted by a partner institution (91.7%) rather than an independent hosting service (just a single project took this approach) (Table 3.19¹⁰).

⁹The fact that all projects used URIs is merely an artefact of the method (See Section 3.2.1).

¹⁰Some projects stated that their URIs were hosted but not resolvable. These have been excluded from the table.

TABLE 3.16: STCH Survey: Project aims




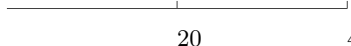
‘The project aimed to:’ (multiple choice)	Number of Projects	
Link data to external data	30	
Link data internally	27	
Allow external data to link to project data	26	
		

TABLE 3.17: STCH Survey: Semantic Technologies used







‘Semantic technologies used:’ (multiple choice)	Number of Projects	
URIs	40	
RDF/RDFS	32	
SPARQL	24	
OWL	20	
SKOS	18	
		

TABLE 3.18: STCH Survey: URI usage








‘If you created (minted) URIs, were they:’ (multiple choice)	Number of Projects	
HTTP based	36	
Resolvable	24	
Persistent	22	
Used content negotiation	12	
		<div style="text-align: right;">20 40</div>

TABLE 3.19: STCH Survey: URI hosting

‘If resolvable, are [Project URIs] hosted by:’	Number of Projects	
An organization associated with the project	22	
An external organization	1	
NO ANSWER	1	
		<div style="text-align: right;">20 40</div>

It is notable that ontologies were much more commonly used by the fixed-term projects (OWL: fixed-term 70%, open-ended 30%; SKOS: fixed-term 60%, open-ended 30%; CIDOC CRM: fixed-term 65%, open-ended 30%). In marked contrast however, the open-ended projects were much more likely to use URIs that are persistent (fixed-term 35%, open-ended 80%) and resolvable (fixed-term 45%, open-ended 75%). Once again, it seems that the fixed-term projects are principally concerned with modelling — and presumably querying and reasoning over — a controlled and integrated dataset, whereas the open-ended projects were more intent on establishing an open and stable Web presence.

3.3.4 Data Conversion

As noted above, the majority of projects converted legacy data into a semantic format, rather than generate it from scratch. For both fixed-term and open-ended projects about half mapped terms to URIs defined elsewhere (fixed-term 40%, open-ended 50%), overwhelmingly by a semi-automated process (Table 3.20). A very small number of projects of both kind did mapping solely by hand (2 projects in each category) and there were two fixed-term projects that used a fully automated process.

TABLE 3.20: STCH Survey: External URI mapping







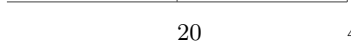
External URIs identified by:	Number of Projects	
Automated process	2	
Semi-automated process	18	
Manual process	4	
		

TABLE 3.21: STCH Survey: RDF generation — process

‘RDF generation is/was:’	Number of Projects	
Dynamic	18	
One-off Export	10	
		

A more surprising result was that RDF, when generated, was more frequently produced dynamically than as a one-off export (Table 3.21). It should be noted that this was especially common for fixed-term projects with 11 projects (55%) claiming to use this method. There is an interesting twist however: of the 11 RDF-producing projects that do not make their data publicly available, 8 produced it as a one-off export and 3 produced it dynamically. In complete contrast, of the 17 projects which claim both to have publicly available data and produce RDF, 14 unambiguously used a dynamic process, while only 2 produced it as one-off export. This would seem to suggest that ‘open’ projects frequently have a non-semantic back-end which is continuously updated, whereas ‘closed’ projects commonly export the data once and then perform analyses. On the other hand, there was no direct correlation with projects being open-ended or fixed-term. The conversion process itself always involved a Computer specialist with considerably less involvement from Humanities specialists, although this happened much more frequently in fixed-term projects. The mapping itself was mainly done with either bespoke software or XSLT (Table 3.22). Protégé¹¹ and D2R MAP¹² each received a single mention.¹³

75% of the projects claimed that data was still available in a semantic form irrespective of whether it was open access. Table 3.23 shows that this is was largely by means of a file-based format (typically RDF/XML). As the majority of these projects stated that they used a dynamic RDF generation process it is presumed that this was a regular

¹¹<http://protege.stanford.edu/>

¹²<http://www4.wiwiiss.fu-berlin.de/bizer/d2rmap/d2rmap.htm>

¹³These were originally classified as ‘Other’, rather than ‘Generic software’, by the respondents, which may call into question whether the categorisation used by the survey is satisfactory.

TABLE 3.22: STCH Survey: RDF generation — technologies





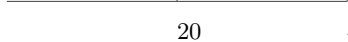






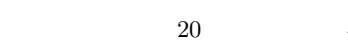
‘RDF generation was done by:’ (multiple choice)	Number of Projects	
Bespoke software	17	
XSLT	12	
Generic software	2	
Other	1	
		

TABLE 3.23: STCH Survey: RDF formats used

RDF formats available	Number of Projects	
RDF/XML	21	
SPARQL	10	
RDFa	6	
RDBMS with URI values	5	
Turtle	3	
N3	2	
		

dump rather than a live connection to a relation backend as might be achieved with the D2RQ Server. Data was also made available (if considerably less frequently) as a SPARQL endpoint, RDFa, a relation database with URI values, in Turtle and N3.

3.3.5 Access and Consumption

We have already noted that just 60% of the projects made their data publicly available — principally those of an open-ended nature (Table 3.12). 50% of all the projects made their data available via an alternative API or computer readable interface such as JSON or Atom, irrespective of whether it was also available in a semantic format. Two thirds (67.5%) of the projects also provided tools for visualization or an alternative human readable interface and results were broadly similar in both cases between fixed-term and open-ended projects. Yet despite this fact, levels of consumption seem to be low. Only half the projects report that their data is being consumed by partners involved with the project, and just one quarter report consumption by target groups or anyone else (Table 3.24). Open-ended projects dominate this last category with 50% reporting consumption by external parties as compared to just 20% of fixed-term projects.

TABLE 3.24: STCH Survey: Data consumption















‘To the best of your knowledge is your data being consumed by:’ (multiple choice)	Number of Projects	
Project partners	21	
Target groups	10	
Anyone else	11	
		

TABLE 3.25: STCH Survey: Respondent advocacy

‘Would you advocate the use of semantic technologies for similar tasks in future?’	Number of Projects	
Definitely	25	
Probably/Possibly	14	
Probably not	1	
Definitely not	0	
		

Once again, people’s perceptions of the technology in this regard seems to be coloured by their views about its purpose. When asked ‘would you advocate the use of semantic technologies for similar tasks in the future?’, 65% of the projects had at least one respondent reply that they ‘definitely’ would (Table 3.25) and expectations, regardless of what they may have been initially, largely appear to have been met (82.5%) and in sometimes even exceeded (25%) (Table 3.26). Of those that were content, just 33% reported consumption by external parties and 30% reported no consumption by anyone at all. However the distribution is heavily skewed. 78% of fixed-term projects that claimed they would ‘definitely’ advocate its use had no external consumption, compared to just 36% of open-ended projects. Is this simply due to the fact that many more open-ended projects are externally consumed? Apparently not — of the 6 fixed-term projects that were externally consumed, only 4 (67%) would definitely recommend them, whereas of the 10 open-ended projects that were externally consumed, 9 (90%) would do so. It therefore seems clear that different sets of goals are being aimed at here.

TABLE 3.26: STCH Survey: Respondent satisfaction

‘Did semantic technologies live up to your expectations?’	Number of Projects	
Much better than expected	5	
Better than expected	5	
As expected	23	
Less than expected	5	
Not at all	0	

20 40

3.3.6 No Linked Open Data?

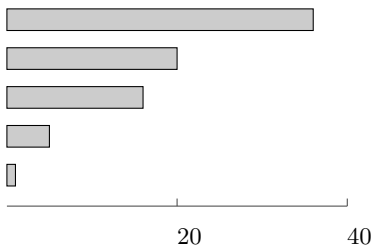
One final finding came as a considerable surprise and may tell us a lot about current perspectives on semantic technologies in Cultural Heritage: although many of the ‘Linked Data Rules’ are followed independently, comparatively few projects follow them all, and those that do are almost entirely ‘closed data’. In other words there is virtually no entirely Linked Open Data in Cultural Heritage (Table 3.27). While the majority use URIs to express data (Linked Data Rule #1),¹⁴ just half also have dereferencable HTTP URIs (Linked Data Rule #2). Of these, 16 projects express their data as RDF (Linked Data Rule #3), but just 5 projects link to external URIs as well (Linked Data Rule #4). Of the entire set of projects, only a single one meets all these criteria and is publicly available. This last project — while an excellent exemplar — is highly atypical, comprising a small dataset of manually curated data expressed as RDFa in an otherwise conventional website. Of course, the Linked Data Rules are not required to be followed in that specific order, and it is generally agreed that following some of them is better than following none of them, yet it seems hard to conclude that any part of the Cultural Heritage sector has fully embraced the Linked Open Data paradigm when so few examples exist in practice.

3.3.7 Respondents’ Comments

While the great majority of questions were either multiple-choice or required short responses, the end of the questionnaire asked participants four questions that gave them an opportunity to express themselves more fully. The full text of these comments can be found in Appendix F but a number of themes emerged which are collated below.

¹⁴Four projects are excluded by virtue of providing only an ontology, rather than data.

TABLE 3.27: STCH Survey: Conformance with ‘Linked Open Data Rules’

Conformance with Linked Data Rules:	Number of Projects	
Rule 1	36	
Rules 1 and 2	20	
Rules 1, 2 and 3	16	
Rules 1, 2, 3 and 4	5	
Linked Open Data	1	

‘What was/were the greatest advantage(s) of using semantic technologies?’

The Web / Sharing Data

Unsurprisingly, the benefits associated with Web access were commented on by many of the respondents. “*Web-friendliness*” and “*data exchange*” were both mentioned, as was “*the ability to automatically, repeatedly, and quickly convert from a museum professional point of view to a public point of view*”. One participant particularly appreciated the “*radical transparency of previously opaque, specialized data [and] ability to reuse/remix the now ‘atomic’ data in unforeseeable ways.*”

Data Integration

This was also a major focus of respondents’ comments, often in a localised context (“*sharing data across partners*”, “*internal re-use of data*”). The possibilities for “*pulling together and cross searching disparate data*” and “*faceted searches*” were also brought up. Data integration does not always appear to have been seen as a one-off process. One comment highlights the “*open-ended nature of data structuring used, allowing an incremental approach to data integration.*”

Miscellaneous

While exposure and integration were by far the most common themes, other interesting reasons for adoption included “*support for multilingual access*”, “*flexibility of modeling*” and the fact that “*semantic technologies enabled us to greatly enlarge the amount of usable data from an otherwise often sparsely populated, text-block-heavy data set.*” One participant noted potential cost savings: “*The adaptation costs are likely to be low in the long term: I’m front loading costs to the early, set-up, stage so that they become part of the method and process*”, although they also made clear that it was “*too early*” to be sure.

‘What was/were the greatest disadvantage(s) of using semantic technologies?’

Four of the participants stated explicitly that they saw no disadvantages to using semantic technologies but this was certainly a minority view.

Difficulty

Criticism focussed overwhelmingly on the difficulty of implementing semantic technologies, with at least half the respondents mentioning this in some manner. Phrases such as *“incredible amount of work”* and *“bloody hard”* give an indication of the challenges faced by some. This seemed to be at almost every level of the process. Comments include: *“difficulty to explain the complexity and potentials of the approach to the project partners”*, *“effort to build community consensus on ontologies”*, *“considerable effort... to build the required ontologies”*, *“grounding local URIs against authoritative ones is hard”*, *“integrating legacy content production to the system is hard”* and *“not easy to understand for non-experts.”*

Tools and Training

Some of this difficulty certainly stems from what was considered to be a lack of suitable software. An *“immaturity of tools”*, *“complexity of toolset”* and *“buggy software”* meant projects were not *“able to offer [museum staff] the tools to allow them to improve the ontologies and semantic logic”*. Likewise, *“The lack of good clear examples”*, *“trained human resources”*, and *“training”* were all cited as additional problems.

Consumption

Less frequent, but still common, were criticisms related to consumption. Several mentioned *“performance”* and *“slow response times for complex queries”* as specific issues. Two participants stated that *“There are very few end-user services which make meaningful use of it”* and that there are *“no semantic web users.”* One particularly critical respondent felt that there was a *“lack of functional benefits”*.

‘Would you do anything differently in future?’

Prior knowledge

A good number of participants suggested that they would have benefited from additional knowledge before commencing: *“[It] would have been good for me to understand these technologies better before we started (and now for that matter)”*, *“[I would] get a semantic technologies specialist to advise on adoption of some in-house decisions.”* An important twist on this theme was the need to *“focus more effort on institutional buy-in. While the project is viewed as an overwhelming success by its intended (external) users, it is frequently viewed as a failure by museum staff who have expectations of 100% data accuracy and of all conceivable functionality”*.

Tools and Resources

Several participants wished they had had *“more project resource on semantic issues”* such as *“a larger team to assist with creating the semantic output”*, a *“tool to map dataset on the ontology”* and even *“a budget (this was done for \$0 with volunteered time and effort).”*

Born Semantic

Three participants indicated that it is much easier to *“try to be born semantic”* than convert legacy data, as *“solving interoperability problems is much harder afterwards and data quality is lower.”*

Reduced Complexity

Three respondents suggested that they would *“focus on getting data used before worry about modeling too much”* and *“trade ‘complete’ for ‘good enough’ earlier.”* One of these *“Wouldn’t use full blown CRM! Things have moved on a lot since the project with simpler reference models e.g. CRM Core probably yielding a better return on investment.”*

‘Any other comments?’

The majority of responses to this question were very project specific but it is interesting to note that, from those who answered, there seems to be a very wide spectrum of opinion as to the utility of semantic technologies. At one end were criticisms such as:

- *“It was hard work being at the bleeding edge of semantic web R&D as we were at the time.”*
- *“For the neophyte semantic modelling techniques can seem impractical, rigid and unrealistic. Implementation issues often descend into ethereal discussions about meaning itself, with no clear usable outcome.”*
- *“In terms of using semantic technologies at a local level, most needs can be more efficiently met by using simpler data structures and relationships. Or such has been our experience.”*

On a more positive note, others stated that:

- *“The project was lots of fun.”*
- *“From my perspective, this project has been an overwhelming success...I would definitely recommend semantic technologies to anyone looking to improve access to specialized, obscure, jargonized, and/or sparse data.”*
- *“It was very exciting to work with semantic technologies: it opens up a world that goes much beyond simply syntactic technologies such as XML.”*

3.4 Conclusions

We ended the previous chapter having established that two visions of the Semantic Web — Linked Open Data and Mixed-Source Knowledge Representation (MSKR) — appeared to exist, but with less clarity as to whether either of these views exercised particular influence on Cultural Heritage projects. In fact it seems that both perspectives are present, but that they are associated with a very specific variable: whether or not the project concerned is fixed-term or open-ended. Table 3.28 summarizes the results we have been discussing, ranking several key variables in order of their prevalence in association with each type.

Fixed-term projects predominantly use heavyweight ontological apparatus (such as OWL and the CIDOC-CRM), involve a domain specialist in the RDF conversion process, and have a restricted target audience. They are also unlikely to provide resolvable or persistent URIs, grant public access to the data, or (unsurprisingly) see consumption by external parties. For open-ended projects the situation is entirely reversed. This seems as clear an indicator as we might hope for, that the majority of fixed-term projects are interested in MSKR, and the majority of open-ended projects take a Linked Open Data perspective. Of course, these are not entirely discrete classifications. As we have seen, few projects will exhibit all the attributes associated with either of these visions, but a plausible narrative for this division does seem available.

Fixed-term projects typically appear to be concerned with a traditional Humanities research model that necessitates the production of research output before the termination of funding. They must demonstrate a high level of data integrity and a finished product, often a report, database or website. Ontologies, in particular, are seen as a way to bring heterogeneous data sources together into a new, semantically harmonious form. In contrast, open-ended projects are more commonly led by computer scientists or those with technical responsibility for a institutional database or website. They view semantic

TABLE 3.28: STCH Survey: Variables associated with Fixed-term and Open-ended projects

Variable	Fixed-term	Open-ended	Difference
OWL	70%	30%	40
CIDOC CRM	65%	30%	35
Conversion by Humanities specialist	60%	25%	35
Restricted target group	50%	15%	35
External consumption	20%	50%	30
Public access	45%	75%	30
Resolvable URIs	45%	75%	30
Persistent URIs	35%	80%	45

technologies as a way of making their resources more useful by providing a stable Web presence for both incoming and outgoing links.

Let us conclude this first part of the thesis by returning to our original questions:

1. What are the benefits to archaeologists of using semantic technologies to express their data?
2. What are the social and technical requirements for expressing archaeological data with semantic technologies?
3. To what extent is contemporary archaeological practice able and willing to meet the social and technical requirements of expressing data with semantic technologies?

It is clear from our results that the answer to the first question will depend on what the archaeologist in question wishes to achieve. If their work is open-ended they are more likely to be interested in the benefits of Linked Open Data, whereas if it is fixed-term then the possibilities afforded by MSKR may seem closer to their goals. This in turn will have a substantial impact on the answers to Questions 2 and 3. Unfortunately, it is beyond the limitations of this research to undertake the experimentation required to explore both perspectives. Instead it will focus on just one perspective, MSKR, in line with the goals of the Roman Port Networks project. The experiments and results of this case study will be discussed over the next three chapters. We will return to the topic of Linked Open Data, however — and speculate further on what role it might play within Archeology — when we reach our final conclusions in Chapter 7.

Chapter 4

MSKR in Archaeology: Building a Framework

4.1 Overview

In the second part of this thesis we turn from abstract considerations to practical ones. We have identified that semantic technologies do indeed seem to offer potential benefits to Archaeology — two different kinds in fact — and we must now establish the costs of implementing them. In order to do so we will be required to work not only with real archaeological data but also with real archaeologists, and therefore in collaboration with an ongoing project. This in turn determines which Semantic Web model we are able to investigate, MSKR or Linked Open Data, for there are regrettably not the resources available to do both. The case study we will be working with is Roman Port Networks,¹ an international fixed-term research project funded by the British Academy and the University of Southampton, and directed by the British School at Rome. As we shall see, the nature of the research, and the goals it aims to achieve, lead us strongly towards an MSKR approach and therefore the following three chapters will focus on the concerns associated with this perspective.

In this chapter we will discuss some of the features of Roman Port Networks and how its decentralized and voluntary (but closed) partnership structure — which we shall refer to as microprovision — creates a specific set of circumstances to be addressed. We will then go on to describe an infrastructural framework to which each content partner can contribute their data. Once this basic framework has been established we shall be led to investigate ways in which project partners can convert legacy data into a semantic format that is compliant with it. Chapters 5 and 6 will evaluate two alternative approaches to this problem.

¹<http://www.romanportnetworks.org/>

4.2 Case Study: Roman Port Networks

Roman Ports Networks is an investigation into the relationship of Portus — the main port of Rome throughout most of the Imperial era (27 BCE–476 CE) — to other ports in the Western Mediterranean and beyond (Keay, 2009). By establishing the co-presence of specific ceramic and marble types at a range of key maritime sites, it aims to gauge fluctuating levels of trans-Mediterranean interaction during the Roman period. Source data comprise large quantities of published and unpublished harbour and shipwreck excavation databases from a variety of academic and research institutions in different countries.

While the datasets all pertain to the same domain, they frequently employ mixed taxonomies and are heterogeneously structured, containing both raw and summary data. In addition to multiple data formats, normalization is not guaranteed, uncertainty frequent and variant spellings common. Different recording methodologies have also given rise to alternative quantification and dating strategies, and the multilingual nature of the project — with resources in Catalan, English, French, Italian and Spanish² — further adds to its complexity. In other words, it is a typical real-world, mixed-context situation. As an international endeavour, requiring the synthesis of large quantities of data with varying format but restricted scope, it provides an ideal opportunity to work through the issues involved in applying semantic technologies to Archaeology.

At the commencement of the project, the institutional partners agreed to contribute relevant datasets according to two pre-established chronological horizons associated with the Roman era in the West: the Republican-Imperial transition (100 BCE–50 CE) and the post-Imperial period (475 CE–550 CE). In order to maintain realistic working conditions, original data formats were not to be adjusted in any way prior to their conversion to RDF. They also agreed to volunteer a limited amount of time to trial any software or methodologies developed for the project. Such methods should assist users in creating a mapping between local data, a domain ontology and a canonical URI thesaurus for shared concepts such as taxonomies. Once this process is complete, the archaeologist should be able to export data as ontology-compliant RDF that can easily be integrated with data provided by other project partners in order to create visualizations and perform analyses.

As project data is compiled from the voluntary contributions of many partners, much of which forms the basis of their current research, an Open Content policy was not adopted. For the sake of project collaboration it was agreed that both input data and output RDF would be held within a password protected and centrally managed datastore.³ This effectively prohibits the possibility of a Linked Open Data approach and, in combination with the project's goals of data integration, determines MSKR as

²A complete list of project partners is given in Appendix G.

³<http://romanportnetworks.pbworks.com/>

being the perspective most in line with its aims. Nevertheless, it is hoped that much of the data will ultimately be made publicly available and therefore some consideration must still be given to the potential advantages of using globally shared vocabularies, albeit on a provisional basis.

Amphorae

Although Roman Port Networks is concerned with the distribution of a range of materials, for the purposes of investigation into semantic technologies the study is limited to just one: amphorae (Figure 4.1). Originally invented by the Phoenicians, and subsequently adopted by Greeks, Romans and other cultures engaged in the maritime economy of Antiquity, amphorae are standardized ceramic transport vessels used for the long-distance transport of goods. As containers with a standard form and volume, they not only assisted in the physical transmission of liquids and perishables, but also catered to the fiscal concerns of the Roman taxman. Of equal importance to their producers, traders and consumers was the regional stylistic variation that helped those at both the quayside and marketplace decide what to pick up and what to leave behind.

This diversity in size, shape and material — as well as inscribed, stamped and printed text and symbols, content residues, remarkable durability and widespread distribution —



FIGURE 4.1: A *Dressel 20* Roman amphora (Keay and Williams, 2005)

make amphorae one of the most important artefact types available to archaeologists when considering trade and exchange within the Roman economy (Peacock and Williams, 1986). That very importance has led to an extremely broad array of typological systems in order to categorize them, however. Some of these systems are limited to forms found on a single site (e.g. Dressel (1899)) while others provide syntheses of amphorae found across an entire region (e.g. Keay (1984)). Fortunately, the online database *Roman Amphorae: A Digital Resource*⁴ (Keay and Williams, 2005) provides a helpful, if by no means exhaustive, overview of many of the main amphora types and type series.

Focussing on this particular class of material will be sufficient for identifying and addressing many of the theoretical and practical issues we are likely to face when applying semantic technologies in the archaeological domain. However, it will also be important to bear in mind the possibility of including other materials, such as marble and metal ingots, in order to ensure that any infrastructural and methodological proposals we make can be adapted to alternative forms of material evidence with the least amount of effort possible.

4.3 Requirements

While Roman Port Networks clearly has similarities to many other projects that adopt an MSKR philosophy, that group is not homogeneous. A particularly important distinction to consider is that of **macro-** versus **microprovision**. Many Semantic Web projects are not only centralized but also have considerable resources available for investment in technical staffing and resources. This is most frequently the case with GLAM projects which often (although by no means always) have a dedicated IT department and budget. In contrast, archaeological academics often work either independently or within small teams with little if any technical support beyond maintenance of hardware and generic software. They are therefore representative of a ‘long tail’ of potential Semantic Web stakeholders with comparatively small datasets and limited resources (whether technical, temporal, or financial). The absence of this community in much of the literature to date is of particular concern to proponents of Linked Open Data, given the Web’s heavy reliance on large numbers of small contributions, making it a ‘Social Machine’ (Hendler et al., 2008). While our focus here is on MSKR, lessons derived from working with this group are likely to extend to Linked Open Data and, indeed, well beyond the domain of Archaeology.

A further feature of the project to bear in mind is that the Roman Port Networks partnership is based on voluntary contribution rather than contractual obligation. Although this is perhaps more common among Linked Open Data than MSKR projects,

⁴http://archaeologydataservice.ac.uk/archives/view/amphora_ahrb_2005/
(DOI:10.5284/1000021)

it increases the need to make the process of generating semantic data as effective and efficient as possible. A method that is perceived to be either too time-consuming or too difficult will simply not be adopted. It must additionally be able to cope with the diverse range of spreadsheet and database formats employed, as insisting upon a single digital standard is equally likely to act as a deterrent.

So what options are available to us? Sahoo et al. (2009) discuss 15 projects responding to the challenge of mapping relational databases to RDF (a process known as **RDB2RDF**), yet virtually all of them are too generic, too complex, and require too much prerequisite knowledge of the Semantic Web theory and content, to be of use to microproviders. Those that are fully automated (especially Byrne and Klein (2009)) are too inaccurate for use as MSKR data sources although they show greater promise for text mining and Linked Open Data. The report provides no standard formalism for RDB2RDF mappings. The *Requirements for Relational to RDF Mapping* report (Erling, 2008) notes a number of important requirements for a relational to RDF mapping language, but also provides no guidance as to how such mappings should be created or by whom. The *RDB2RDF Incubator Group Final Report* likewise mentions only one Use Case that involves large numbers of datasets and users with low tech-literacy (Malhotra, 2009), but even this assumes multiple instances of a standard schema such as WordPress⁵ and is still only suitable for Web Application developers (Auer et al., 2009). A W3C Working Draft of **R2RML**, an RDB2RDF mapping language, was published in September 2011 (Das et al., 2011), but is unfortunately too recent to have been incorporated into the present work.

Given, therefore, the restrictions on time and money faced by the Roman Port Networks partners in pursuit of their work, a great deal needs to be done to facilitate this production process. In more concrete terms, a successful mapping and export process must aim to be:

Quick – There is clearly no universal definition of ‘quick enough’ but a process taking longer than a working day to complete (from commencement to visualisation) is likely to be perceived as a project rather than a task and correspondingly more burdensome on the producer.

Cost-effective – It must use easily available (ideally Open Source) software and require minimal technical support.

Accurate – It must produce RDF at a level of accuracy limited only by the source data. Note that this does not imply the same level of completeness or precision as the source, but the output should not introduce false information.

Transparent – The archaeologist must understand enough of the production process to feel confident in its output.

⁵<http://www.wordpress.com>

A production process that does not fulfil these criteria is unlikely to support the number of contributors, and thus datasets, requisite to catalyse a positive feedback loop within the project, and arguably, most projects that operate on a decentralized, voluntary basis. The goal is therefore to develop a prototype system sufficiently intuitive and well-documented to be usable without a significant level assistance, while generic enough to be adaptable across a range of archaeological materials. Any proposed solution can be broken down into two elements:

1. Infrastructure: The specification and instantiation of a common ontology and thesauri in RDF.
2. Support: The implementation of a workflow process that allows data holders to export their data as ontology- and thesauri-compliant RDF.

4.4 Infrastructure

The ability to generate RDF cannot reside solely in a standalone application. It requires a service-level framework that provides URIs for specialized common vocabularies such as the ontology and thesauri. Only once this is in place is it possible to map between **local terminology** (i.e., terms used by the datasource) and **canonical URIs**. There are two distinct features of these global identifiers to consider. The first (which we might call **Term Standardisation**) is their ability to provide a *de facto* common vocabulary of common URI concepts. This may derive from either regulation or consensus — typically the latter, as neither the Internet nor archaeological communities have historically responded well to externally imposed standards. Nevertheless, as it is in everyone’s general interest to share terminology, conventional standards emerge over time and by providing canonical URI terminologies where none currently exist, it can be made more convenient for new contributors to adopt them rather than develop their own. The second aspect (which we might call **Term Contextualisation**) is to provide canonical URIs with additional background information using RDF. Unfortunately, the highly interpretive nature of Archaeology means that such ‘descriptions’ can be highly contentious, but as URIs are themselves opaque (i.e. have no semantic content) it is vital to provide at least enough information to unambiguously identify the resource they refer to. Bearing these considerations in mind there are three tasks to accomplish:

1. Establish a domain ontology for the project.
2. Establish thesauri for the project.
3. Arrange hosting and maintenance of the above.

4.4.1 Domain Ontology

A domain ontology serves two purposes. First it provides a common reference point for both data contributors and consumers, enabling them to understand the relationships between the canonical URIs they are mapping to and thus eliminate some (if not all) potential misconceptions about them. This is most easily achieved by having a publicly accessible diagrammatic and text representation of the key concepts (classes and their properties). Second, it allows machines to inference across data, permitting logical conclusions to be derived that may not be explicitly expressed in the data itself. This requires the RDF to be accessible in a machine-readable notation such as RDF/XML or Turtle.

It is clearly best practice to adopt pre-existing ontologies if possible in order to diminish the Co-reference Problem and reinforce the *de facto* terminology standards discussed earlier. In the case of Archaeology, an ontology already exists for archaeological fieldwork known as CIDOC CRM-EH (Cripps et al., 2004). We have decided not to use this ontology which therefore requires some justification. As discussed in Chapter 2, CIDOC CRM-EH was developed as an extension to the CIDOC CRM, a more generic ‘core ontology’ developed for the Cultural Heritage sector. While the potential informative power of the CIDOC CRM is great, it is felt by many to be too abstract and too complex for adoption without considerable expert assistance (Cripps et al., 2004; Addis et al., 2005; Stein et al., 2005). As the principal goal of this research is to identify the benefits and costs specifically associated with semantic technologies — and these were themselves frequently described as ‘difficult’, ‘complex’ and ‘hard’ by the STCH survey participants — it was decided that there was too great a risk of the complexity associated with the CIDOC CRM overshadowing the challenges presented solely by semantic technologies themselves. Nonetheless, interesting work is being done in this space, in particular by the STAR and STELLAR projects (see 2.3.3). Should such research demonstrate that the CIDOC CRM can be made more easily comprehensible to non-specialists, it will certainly be worth reappraisal.

Returning to the drawing board on this basis, it is clear that the ontology must be kept as simple as possible. In order to do this, the principle was maintained that only data which could be used for meaningful *inter-site* comparison should be represented. For example, the number of rims, bases and handles may be important to determine the approximate overall quantity of amphorae, but across multiple sites this level of granularity is superfluous and thus excluded from the ontology. While such an approach is particularly suited to the broad-scale analyses of Roman Port Networks it may also provide a sound universal principle for Archaeology in general. Semantic technologies are not a replacement for individual data repositories and, for the foreseeable future at least, *intra-site* analysis is almost certainly better carried out using a conventional site database which will have been developed to accommodate the specific needs of the

excavator's methodology. There are therefore no immediate grounds for increasing the complexity of an ontology by introducing elements that have no relevance beyond their original dataset.

The ontology used, known as **ArchVocab**,⁶ was designed with assistance from a number of ceramics and marble experts in the UK, France, Spain and Italy, in order to ensure that key data necessary for a comprehensive inter-site summary can be described adequately and that strategically useful research questions can be addressed. Figure 4.2 provides a visual rendering. The ontology is separated into **Classification** and **Instance Data** layers so that independent datasets are linked by canonical URIs signifying classificatory and singleton concepts. These URIs provide a vocabulary of terms that are common to any instance data set: typology, location, period, form or material. The classification layer also makes deliberate reuse of external vocabularies, including SKOS and Hempl.⁷ The Instance Data layer represents resources that are specific to a given dataset and require the creation of new URIs. The overall design is intended to be simple and stable enough for archaeologists to easily interpret in relation to their own data.

It is worth observing that the 'type' of an artefact may refer to entirely different qualities, depending on the nature of the material under consideration. For instance, while ceramics experts generally emphasise the shape of an amphora over its fabric when classifying, marble specialists tend to focus on the material first. As a result, the verb 'type' has been avoided and separate properties `archvocab:ofForm` and `archvocab:ofMaterial` created so that both can be described without ambiguity. In accordance with our principle of recording data relevant to inter-site analysis, material is here used to specify the region of origin, rather than other defining features — such as colour or consistency — which have limited relevance when compared across excavations.

An amphora's 'class' is the combination of both form and material (Peacock and Williams, 1986) and this combination forms the basis of what is perhaps the most 'atomic' entity in the ontology: `archvocab:Find`. A **Find** is the quantity of all finds of a specific amphora class within a given archaeological context (represented by `archvocab:Context`). The **Context** in turn records the *terminus post* and *ante quem*, area, and type of the context, as well as the `archvocab:Excavation` to which it belongs. Finally, the **Excavation** is associated with a specific geographical place in the modern or ancient world (and thus spatial location as well). These instance concepts are associated with both canonical URIs and Literal values. In the latter case, the varying kinds of quantity metric are represented by a set of subProperties of the Property `archvocab:hasQuantity`.⁸

⁶<http://archvocab.net/excavation/ontology/>

⁷<http://www.hempl.org/rdf/2003-09-17/hempl>

⁸For legacy reasons, `archvocab:hasQuantityNMI` follows the Spanish *Número Mínimo de Individuos* rather than the English Minimum Number of Individuals (MNI).

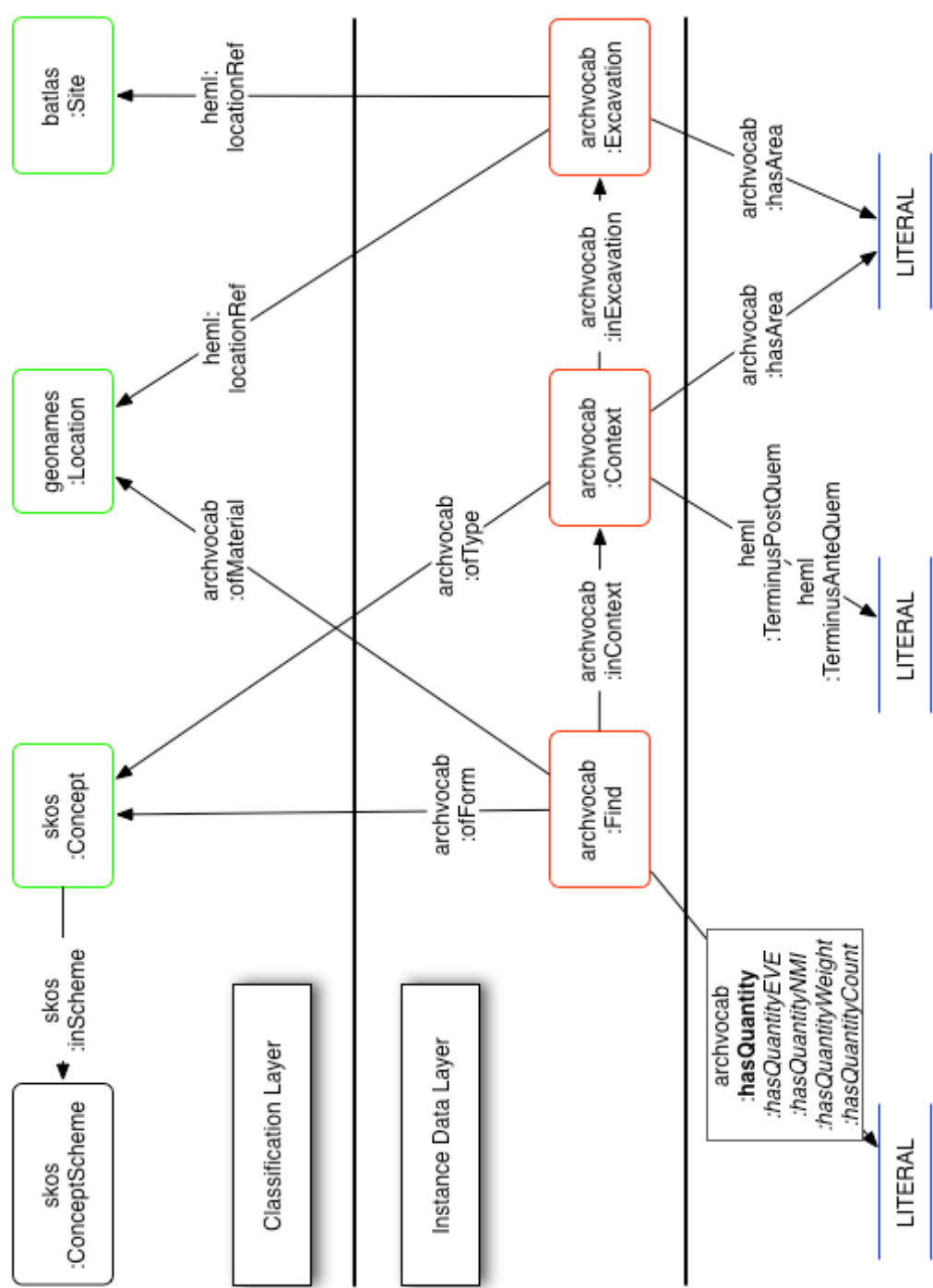


FIGURE 4.2: ArchVocab Excavation Ontology diagram

4.4.2 Thesauri

The most natural axis on which to combine datasets is across category distinctions. While the ontology makes it possible to map the structural semantics of a database or table, it does not contain the typological terms by which to classify instance data. In order to achieve this, a URI term-list of appropriate classifications must be used. As we have seen, the challenge is significant as there are numerous typology systems for amphora forms with a large amount of overlap between them. An average site record is likely to refer to forms from multiple typology systems and there may even be disagreement as to whether two forms are in fact identical. Any attempt to create a new typology system, or impose the use of one, will likely deter curators from contributing for fear of introducing inaccuracies into their data. Therefore the only practical solution is to create URIs for all ceramics types and use RDF to associate those that are considered to be the same. Fortunately *Roman Amphora: A Digital Resource* can greatly aid in this task. Although it does not list every amphora form — identical forms having been distilled into a single category — it does provide a strong platform for further development and is certainly sufficient for experimental purposes.

Thesauri were initially implemented in a bespoke SKOS file. SKOS, as we have seen, is an RDF ontology that provides a standard URI vocabulary for dealing with Knowledge Organisation Systems such as classification thesauri (Miles and Bechhofer, 2009). Two of its core classes, `skos:Concept` and `skos:ConceptScheme`, along with the Property `skos:inScheme` permit us to define an amphora form as belonging to a specific typology system in RDF. Additional information, such as alternative names or the source of its definition, can also be given. Each typology system in the database was converted into SKOS-compliant RDF files with a separate index of resources and corresponding HTML description.⁹ The advantage of this approach, to be discussed later, is that such a scheme can be managed centrally, and additional information, such as the synonymy or subclassing of different amphora forms, can also be expressed using Properties such as `skos:exactMatch` and `skos:narrowMatch`. In order to help query data by the environment in which amphorae were found, a second SKOS thesaurus of archaeological context types was also created.¹⁰

In order to deal with find materials other than amphorae — such as the Marble Catalogue of the *Institut Català d'Arqueologia Clàssica* (ICAC) — it would naturally be necessary to create further thesauri. Generic tools for this kind of work, such as Protégé,¹¹ are one way of adapting pre-established classification schemes. However, the technical challenges associated with this process, especially when updates may be required from multiple

⁹<http://archvocab.net/amphora/>

¹⁰<http://archvocab.net/excavation/context/>

¹¹<http://protege.stanford.edu/>

contributors, may make it more practical to use an open URI service such as Freebase.¹² A more detailed discussion of the issues involved is provided in Sections 6.1.2 and 6.3.2.

External Services

As well as finding typological correspondences across datasets, it is extremely helpful to see spatial ones as well. Traditionally this has been done using Geographic Information Systems (GIS) that use coordinates to locate points or areas against a common geographic datum. Spatial data of this nature can be extremely useful, but its abstract geometry is frequently unsuitable for discovering the conceptual and mereological associations between places. Rather than associate excavations or regions of origin with spatial coordinates, it is sometimes more helpful to associate them with place concepts, such as settlements and territories, especially when dealing with finds from an era in which the historical placenames are known.

The GeoNames¹³ service, a longstanding occupant of the Linked Open Data Cloud diagram, takes precisely this approach to spatial data. It bundles together synonymous placenames, and by giving each bundle a URI, it is then able to provide a great deal of additional information about each place, including alternative names, approximate coordinates and feature type. Although it draws on a wide range of public datasources, it also solicits **Volunteered Geographic Information (VGI)** with new places being continuously added by members (these are vetted to avoid duplication). This important benefit allows for the introduction of absent ancient sites, although a good number of classical locations are already incorporated. GeoNames also provides both a search API and the ability to request RDF based on an HTTP request to the place URI. Although similar large-scale gazetteer services exist (e.g. Yahoo! (2010)), none of them yet provide such an open, URI-based framework. A parallel development, the Pleiades Project (Elliott, 2009) has been developing an online gazetteer of ancient toponyms along similar lines, providing URIs for all the places listed in the standard reference work for ancient place locations: the *Barrington Atlas of the Greek and Roman World* (Talbert, 2000). Unfortunately Pleiades did not provide spatial information at the time the experimental work described in the following chapters took place and so was not directly used. This has since become available however, and provides an ideal resource for classical archaeologists.

As well as space, it would be equally useful to compare data along the temporal dimension. The most advanced work done on this topic is that of Binding (2010) based on the **English Heritage Timelines Thesaurus** (English Heritage, 2000). Unfortunately this is restricted to dates relevant to the UK and its license stipulates that the thesaurus is “available for incorporation into systems subject to license prohibiting unauthorised

¹²<http://www.freebase.com/>

¹³<http://www.geonames.org/>

passing on of the reference data set to third parties” thus rendering it unsuitable for multi-institutional projects such as Roman Port Networks. The additional advantage of linking to such URI vocabularies, especially where they are shared publicly, is that in time it should be possible to draw on a wealth of contextual data without being limited to traditional disciplinary boundaries (Harris et al., 2010; Barker, 2011).

4.4.3 Hosting and Maintenance

In contrast to the instance data itself, there is no need to restrict access to the URI infrastructure described above. Indeed, to do so reduces its utility by preventing the serendipitous associations that can arise when third parties make use of it. On the other hand, there is a concomitant risk that failure to maintain the service can leave those who use it with broken links. So long as a domain remains within the hands of a responsible community, a defunct service can potentially be revived. However, URIs with unrestricted top-level domains (such as `.com`, `.org` or `.net`) are susceptible to domain-squatting if the license should lapse and the domain be re-registered. It is therefore preferable for important services to be provided under government or academic top-level domains (such as `.edu` or `.gov`) where this is less likely to occur. The use of `http://archvocab.net` as the central domain for the Roman Port Networks URIs seems, in retrospect, a less robust choice.

Ontologies and thesauri were instantiated in RDF/XML and HTML, including a graph representation. They are served from an Apache Web Server¹⁴ running on a Linux virtual machine hosted by the University of Southampton ECS. The server was also configured to automatically perform appropriate **HTTP 303 Redirects**, dependent on the client request (Heath and Bizer, 2011, Section 2.3.1).

4.5 Conclusions

With a stable URI infrastructure in place to link to, the next objective is to provide tools and a workflow by which Roman Port Networks partners can map and export their holdings as RDF with minimal support. Such a delegated methodology is vital to prevent the creation of bottlenecks within the project workflow, or risk project failure, if technical support can no longer be provided at any point.¹⁵ In order for ontology and thesauri-compliant RDF to be generated by those without an informatics background, a number of practical and theoretical tasks need to be accomplished (Table 4.1).

The following chapters discuss two alternative approaches to this process. Chapter 5 discusses TRANSLATION, a Wizard-based toolkit that uses Natural Language Processing

¹⁴<http://httpd.apache.org/>

¹⁵By avoiding a so-called ‘High Bus Factor’ (Malik, 2005).

TABLE 4.1: RDF generation support tasks

Task	Description
Comprehension of task/ workflow/ontology	The archaeologist has to understand the nature of the abstract ontology and thesauri to which they are mapping and tasks which they are to undertake.
Schema-to-Ontology mapping	Local relational table and column schemas must be mapped to the concepts represented in the domain ontology.
Canonical URI Mapping	Locally-used terminology must be mapped to pre-existent canonical URIs.
Literal Standardization	Literal values, such as numbers and text, may need to be normalised and/or converted in order to comply with the Property concepts within the domain ontology.
Instance URI generation	URIs for instances specific to the dataset must be created based on an appropriate namespace.
RDF generation	RDF compliant with both the ontology and the implicit structure of the data must be generated.
Enrichment	Important additional information that may not be part of the source data — such as the excavation name or data provenance — must be included.
Validation	The archaeologist must be able to establish whether the process has been successfully completed.

techniques to predict likely matches between local terms and canonical URIs. Chapter 6 describes *Introducing Semantics*, a ‘Cookbook’ method that strikes a very different balance in terms of user engagement and technological requirements. Both approaches are then evaluated, based on experience and feedback from the Roman Port Networks partners.

Chapter 5

TRANSLATION — A Toolkit for Generating RDF

5.1 Overview

Having established the necessary infrastructure for providing canonical URIs, it is now possible to develop the tools required to generate ontology and thesauri-compliant RDF about instance data from real-world datasets. Fully automated tools, such as D2R (Bizer and Cyganiak, 2006), are not suitable for this process due to the lack of normalization across datasets and the high level of domain knowledge typically required to understand the diverse codes and abbreviations employed. This chapter describes the first of two possible approaches, known as TRANSLATION.

TRANSLATION is a Wizard-based toolkit that attempts to facilitate an end-to-end mapping process — from source data to ontology-compliant RDF — by inspecting a tabular dataset and suggesting likely mappings. Two processes are involved. In the first, which is undertaken only once, a data contributor uses the Wizard to create a unique XML configuration file that expresses mappings from the local datasource schema and terminology to the canonical URIs. Using basic Natural Language Processing, it predicts probable mappings that the user can accept or correct using a **Graphical User Interface (GUI)**. It also creates (‘mints’) standardised URIs for instances, such as excavation and context, for which a canonical URI is not available. In the second stage, which may be executed multiple times, a fully automated process uses the configuration file in conjunction with the original dataset to generate and export RDF. Future updates require no additional work (and could even be run nightly, if desired) assuming the datasource schema is not altered and new local terms are not introduced.

The software for both processes was prototyped as a standalone application written in the **Java**¹ programming language and is described in the following sections. The write-once-run-anywhere nature of Java applications makes them ideal for these kinds of utility applications and provides a great deal of flexibility in the prototyping phase. As a compiled programming language however, it is somewhat unsuited to adaptation by those without experience in software development. This issue will be discussed further at the end of this chapter. The following sections outline the key functionality of TRANSLATION before proceeding to a general evaluation.

5.2 Stage 1: Mapping

The **Data Inspector Wizard** is a standalone Java application that can be pointed at a digital resource containing a tabular dataset, such as a database, spreadsheet or CSV file. The Wizard takes the user through a pre-established workflow determined by the kind of find they are importing. Each step of the workflow deals with a specific aspect of the excavation ontology, reading from and updating the central configuration file. It also makes consistency checks to prevent the user from committing logical errors during the mapping process. Although the workflows are ‘hardwired’ in the prototype, such software could be further extended in order to allow advanced users to create new workflows using pluggable components.

5.2.1 Configuration File

The end product of Stage 1 is simply an XML file, the main elements of which are described in Table 5.1. The elements are ontology specific, and thus only able to fulfil the first three requirements in the W3C report *Requirements for Relational to RDF Mapping*: Representation Neutrality, Simplicity, and Readability (Erling, 2008). The three further requirements (Machine Processable Representations, a Smooth Learning Curve and Comprehensiveness) are beyond its limitations. Adaptation to a generic mapping language, such as R2RML, might be envisaged however, now that the W3C RDB2RDF Working Group² has published its recommendations.

¹<http://java.com/>

²<http://www.w3.org/2009/08/rdb2rdf-charter.html>

TABLE 5.1: Principal Elements of XML configuration file

Element	Description
uri	The URI for the RDF graph to be generated
namespace	The URI namespace to be used for new instances
file	The file name and location
dbconfig	Connection parameters for the datasource
concept-mapping	Mappings from excavation ontology Resources to source table and columns
excavations	URIs for each excavation (with additional RDF data where available) and mappings from local terms
contexts	URIs for each context (with additional RDF data where available) and mappings from local terms
context-tpqs	<i>Terminus post quem</i> dates derived from local context dating terms
context-taqs	<i>Terminus ante quem</i> dates derived from local context dating terms
series-mapping	Mapping of local terms used for find typology systems
form-mapping	Mapping of local terms used for find form types
materials	Mapping of local terms used for find material types

5.2.2 Architecture

The Data Inspector Wizard makes use of a variety of Open Source or freely available libraries (Figure 5.1).

The GeoNames Java API³ is provided by the GeoNames project to provide programmatic access to the online gazetteer.

JUNG⁴ is a Java API for graph visualisation.

Jena⁵ is a Java framework for building Semantic Web applications. In particular its API provides for the programmatic management of:

- RDF and RDFS,
- In-memory and persistent storage,
- Reading and writing RDF in various notations,
- A SPARQL query engine.

JDOM⁶ is a parser API for accessing, manipulating and outputting XML.

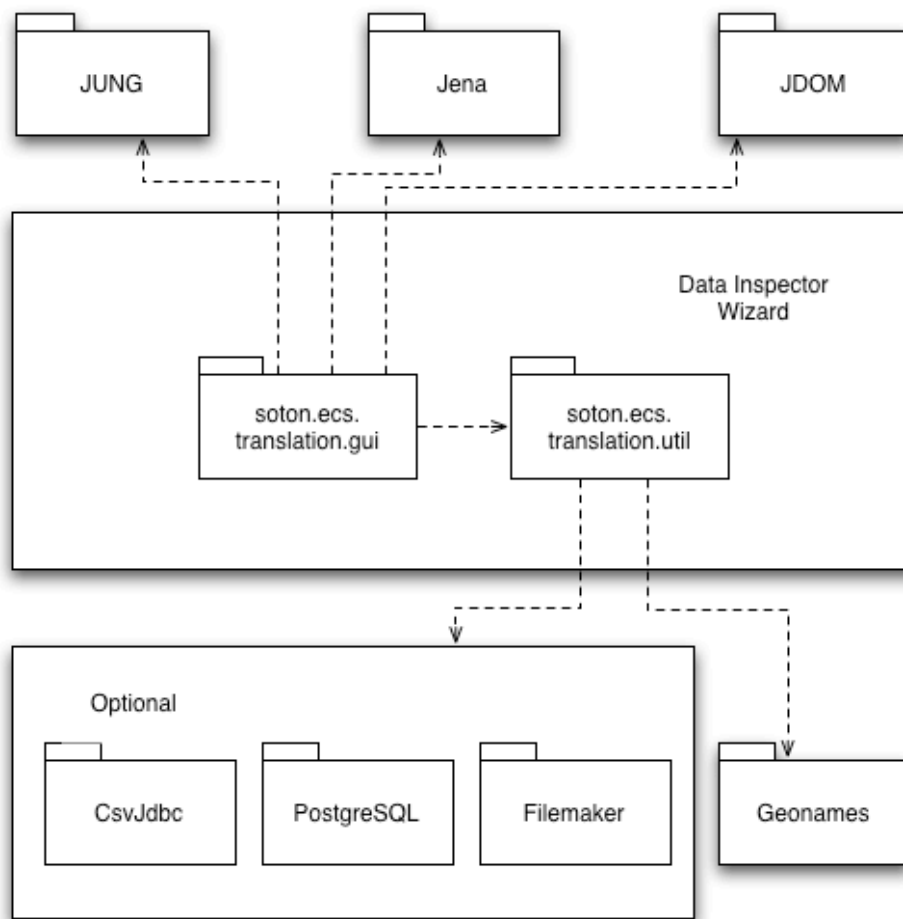


FIGURE 5.1: UML Package Diagram of the Data Inspector Wizard

The application uses the Java Swing⁷ libraries to construct its GUI, making the user step through a series of separate subclasses of `WorkflowPanel` each of which is dedicated to mapping an element of the ontology. The main class, `DataInspectorWizard`, instantiates these classes, depending on the workflow selected by the user. At the start of each step, the panel loads information from the JDOM model. Once the user has finished adding data it then checks the data for any consistency problems before writing back to the configuration file.

5.2.3 Configuration details

The initial step requires the user to define a number of key parameters related to the datasource — including connection and logon details — and the relevant workflow depending on the material type (Figure 5.2). Beyond this, the user must also specify several values specific to the desired output. The first is a URI for the new RDF graph itself. This makes combined RDF data easier to query and filter because the provenance

⁷<http://download.oracle.com/javase/tutorial/uiswing/>

Set Datasource

Please enter the Database connection parameters:

Config ID:

Namespace:

Config File:

☒ PostgreSQL ☐ FileMaker ☐ CSV

☒ Amphorae ☐ Marble

Connection URL:

Username:

Password:

Not connected

FIGURE 5.2: Data Inspector Wizard: Basic configuration information

of each triple is known. It is also important for attaching metadata that could include the contributor’s institution, contact details and licensing requirements. The second is a namespace for the instance URIs the mapping software will create. Ideally this would be a registered domain name owned by the data curator so that RDF/XML output could be hosted locally as Linked Open Data. If this is not possible or desired, the data can still be held in another RDF data repository but the URIs will not be dereferencable.

5.2.4 Schema Mapping

The next step identifies which columns in the source refer to each Resource type in the excavation ontology. The user is presented with both a graph and pseudo-triple representation of the ontology and can then use a drop down list in order to correct the appropriate column name in the ‘Object’ Field (Figure 5.3). Once selected, the graph automatically updates to show how the columns relate to one another. The wizard requires a unique column for each Resource type, so if multiple fields are relevant it may be necessary for the contributor to create a SQL VIEW over the table. In contrast, the

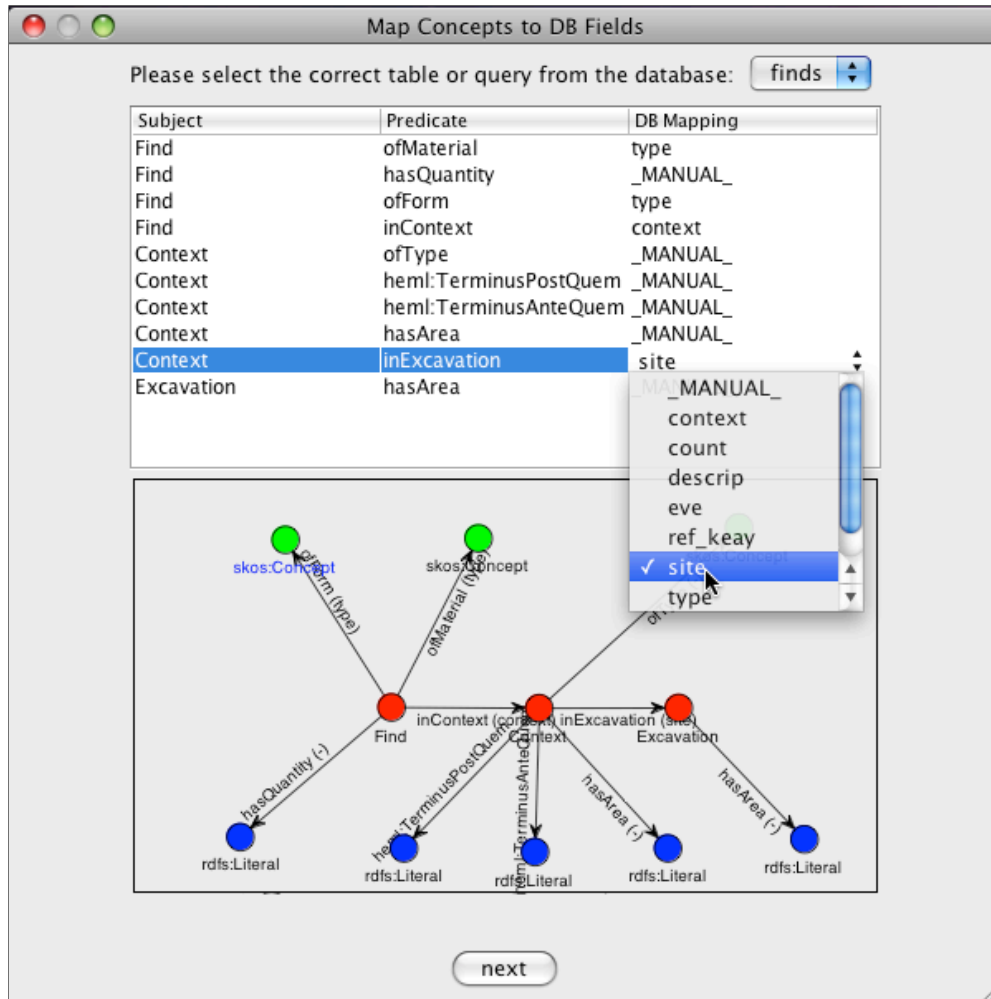


FIGURE 5.3: Data Inspector Wizard: Ontology-to-column schema mapping

same column may be mapped to multiple data Resources although the value extracted from each field will not necessarily be the same (see Section 5.2.7). Occasionally, the relevant data may not be available within the datasource itself (for instance, many databases will not contain a reference to the excavation as this is considered implicit). Should this be the case, the user can define the property as `_MANUAL_` and define it by hand during the appropriate step.

The succeeding steps form a modular workflow that is dependent upon the nature of the local repository. It starts by defining the URI(s) for the excavation(s) the datasource refers to, before proceeding along the dependency chain. While the specifics required to map local database terms to URIs may differ slightly depending on the nature of the concept, this process can be separated into three separate kinds of task: **URI Minting**, **URI Search**, and **Literal Manipulation**.

5.2.5 URI Minting

As every excavation is unique, it will inevitably be necessary to create new URIs representing concepts (including the excavation itself) that do not already exist in the pre-established URI infrastructure, a process known as ‘URI minting’. These may be created manually or from a datasource column. If the creation is manual, the user will be prompted for an input string that names the new Resource. This is converted to lowercase and URL-encoded (in order to escape problematic characters) before being appended to the namespace with an appropriate subdomain (e.g. ‘excavation’). If the creation is automatic, the application will use a `DISTINCT SQL` query to determine individual instances and then repeat the same process for each.

A special case is the generation of `archvocab:Find` URIs which are too numerous to create manually but are not likely to have a unique identifier in the datasource (see Section 4.4.1). These are constructed by concatenating the context, form and material type of the `archvocab:Find`. Once URIs have been constructed, they form the basis of selection queries in subsequent steps. In order to improve user comprehension, only their `rdfs:label` is displayed in application fields but the full URI can be prompted in a tooltip by mousing over it (Figure 5.4).

Context	inExcavation	hasArea (m2)	ofType
11006	Plaza Encarnacion		Construction
11007	Plaza Encarnacion		Public
11010	Plaza Encarnacion		Public
11012	Plaza Encarnacion		Public
11014	Plaza Encarnacion		Public
11015	Plaza Encarnacion		Dump
11019	Plaza Encarnacion		Funerary
11024	Plaza Encarnacion		Funerary
11026	Plaza Encarnacion		Domestic
11035	Plaza Encarnacion		Transport
11050	Plaza Encarnacion		Transport
11078	Plaza Encarnacion		
11094	http://departamento.us.es/dpreyarq/web/anforas/excavation/plaza+encarnacion		
11114	Plaza Encarnacion		
11116	Plaza Encarnacion		
11124	Plaza Encarnacion		
11133	Plaza Encarnacion		
11134	Plaza Encarnacion		
11138	Plaza Encarnacion		
11201	Plaza Encarnacion		
11205	Plaza Encarnacion		
11211	Plaza Encarnacion		
11213	Plaza Encarnacion		
11215	Plaza Encarnacion		
11219	Plaza Encarnacion		

FIGURE 5.4: Data Inspector Wizard: URIs for instance data

5.2.6 URI Search

In cases where the local term refers to a concept that is likely to exist in the URI infrastructure the problem is one of helping the user to discover and identify it. Having pre-established the nature of the Resource in the second step, this process is greatly facilitated because the software can immediately query an appropriate vocabulary service. Two examples are given below.

Simple URI Search Example: Location

The excavation ontology allows both `archvocab:Excavation` and `archvocab:Find` Resources to be assigned properties that take a GeoNames URI as their object. The GeoNames Java API can deliver a set of possible matches against a search criterion entered by the user. Early attempts at fully-automated location assignment proved to have an unacceptably high level of inaccuracy due to the large number of topographic homonyms, but by filtering over additional search criteria, such as category and country, it is usually possible to reduce the possibilities to a very small set (and frequently just a single option). These are presented to the user in the form of a drop down menu that specifies the name, type, and country associated with each URI. (Figure 5.5). One of the advantages of using GeoNames is that it also stores alternative language toponyms for URIs (sometimes including ancient names) so it is possible both to search and represent places in any language. As GeoNames is also a community-based service, if the URI does not exist in the gazetteer it is possible to add it by hand, thus improving the service for all GeoNames users, including Roman Port Networks partners. This is a topic we shall return to later.

Complex URI Search Example: Amphora Form Mapping

While some URIs can be identified by simple string matching, others require more sophisticated approaches. For example, local terms for amphora forms are complex descriptors, generally divided into up to four ordered elements, none of which are mandatory.

1. A reference — often abbreviated — to the typology system, such as ‘Dressel’ or ‘dr.’,
2. A reference to the amphora form by a name or number such as ‘20’ (but occasionally in roman numerals or non-numeric),
3. Additional information such as the material type or an alternative identification,
4. A marker of uncertainty such as a question mark (‘?’) or disjunction (‘ x or y ’).

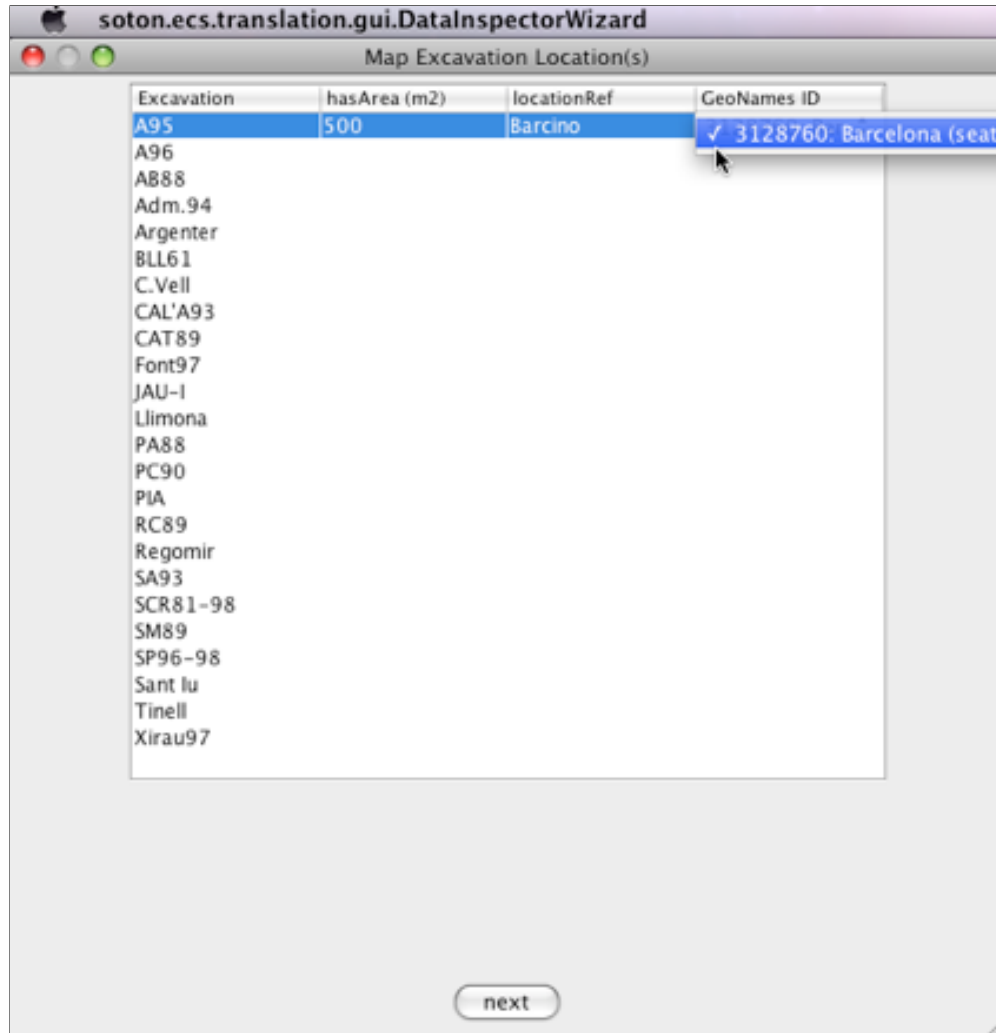
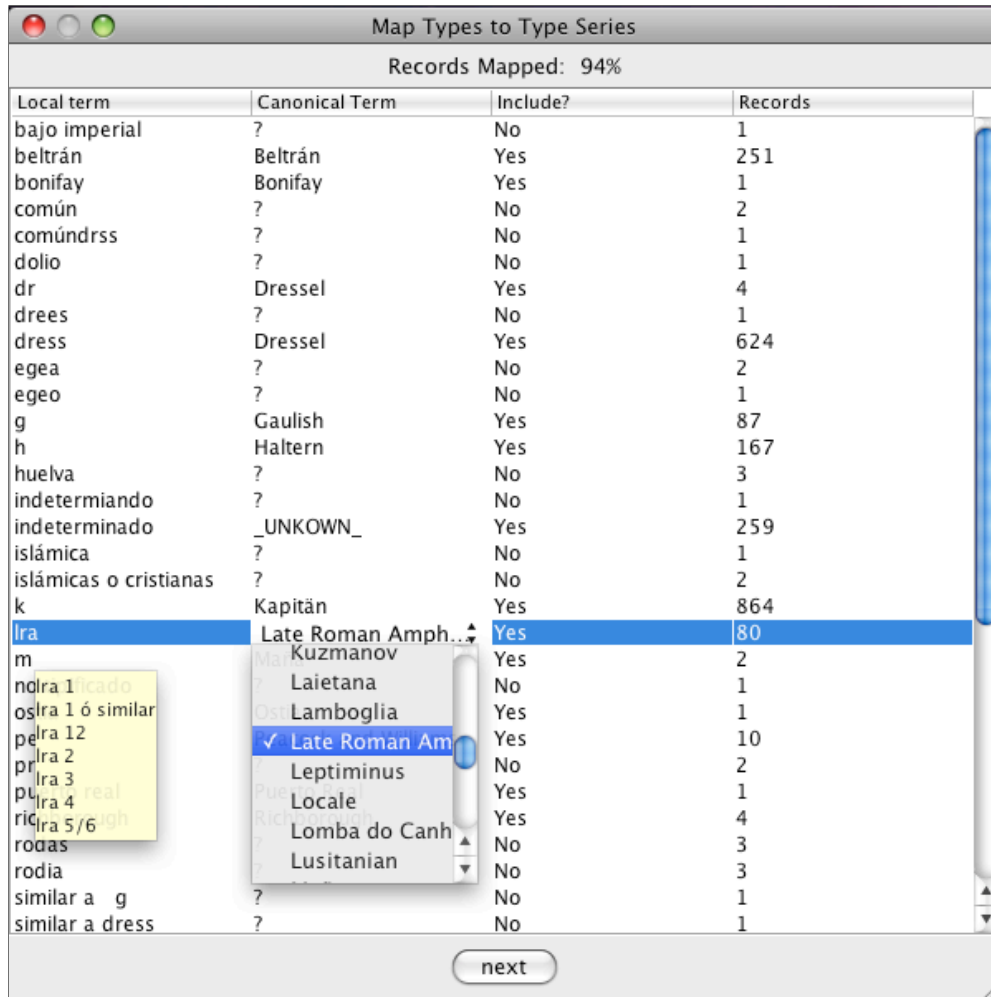


FIGURE 5.5: Data Inspector Wizard: Mapping excavations to GeoNames URIs

The result is that the following two entries could both refer to the same amphora form and even come from the same database:

1. Dressel XX with *tituli picti*
2. Dr. 20?

As it is much easier to identify an amphora form once the typology system is known, the software breaks the identification process down into two parts. First it tokenizes the local term, using the first numeric value (whether arabic or roman) it identifies. The token prior to the numeric value is presumed to be the local term for a typology system. If there is no number, it is assumed to be a typology system with a single amphora form. The Wizard then aggregates all instances in the dataset with the same typology system values and predicts the `skos:ConceptScheme` to which it refers based on string matching. These results are presented to the user for correction (Figure 5.6). From the testing carried out on several real-world datasets it seems that, although prediction may



Local term	Canonical Term	Include?	Records
bajo imperial	?	No	1
beltrán	Beltrán	Yes	251
bonifay	Bonifay	Yes	1
común	?	No	2
comúndrss	?	No	1
dolio	?	No	1
dr	Dressel	Yes	4
drees	?	No	1
dress	Dressel	Yes	624
egea	?	No	2
egeo	?	No	1
g	Gaulish	Yes	87
h	Haltern	Yes	167
huelva	?	No	3
indetermiando	?	No	1
indeterminado	_UNKOWN_	Yes	259
islámica	?	No	1
islámicas o cristianas	?	No	2
k	Kapitän	Yes	864
Lra	Late Roman Amph...	Yes	80
m	Kuzmanov	Yes	2
nolra 1	Laietana	No	1
oslra 1 ó similar	Lamboglia	Yes	1
pe	Lra 12	Yes	10
pl	Lra 2	No	2
pl	Lra 3	Yes	1
pl	Lra 4	Yes	4
ric	Lra 5/6	No	3
rodas	Lusitanian	No	3
rodia	?	No	1
similar a g	?	No	1
similar a dress	?	No	1

FIGURE 5.6: Data Inspector Wizard: Amphora typology system mapping

be quite low across all *terms* used in a dataset (often below 50%), the proportion of *records* mapped correctly without user intervention is generally very high: often 90% or above. This is because deviation from an easily predictable norm is most frequently due to typographical errors in un-normalized source data.

With this done, the Wizard uses the corrected `skos:ConceptScheme` mapping in order to predict the actual `skos:Concept` amphora form for each local database term. Results in field trials from six different institutions have shown it to be fairly accurate (often above 90%) as the estimation process chiefly relies on number-matching. Once again, the user is able to correct mis-assignments or expunge problematic instances (Figure 5.7). As synonymy tends to occur but homonymy does not — or is at least undetectable either manually or automatically — the final output will often map multiple local terms to a single canonical term. For example, ‘Dr. XX’, and ‘Dressel 20’ might both refer to `amphora:dressel/20`, whereas ‘K. 2’ might refer to *either* `amphora:keay/2` or `amphora:kapitan/2` but not both.

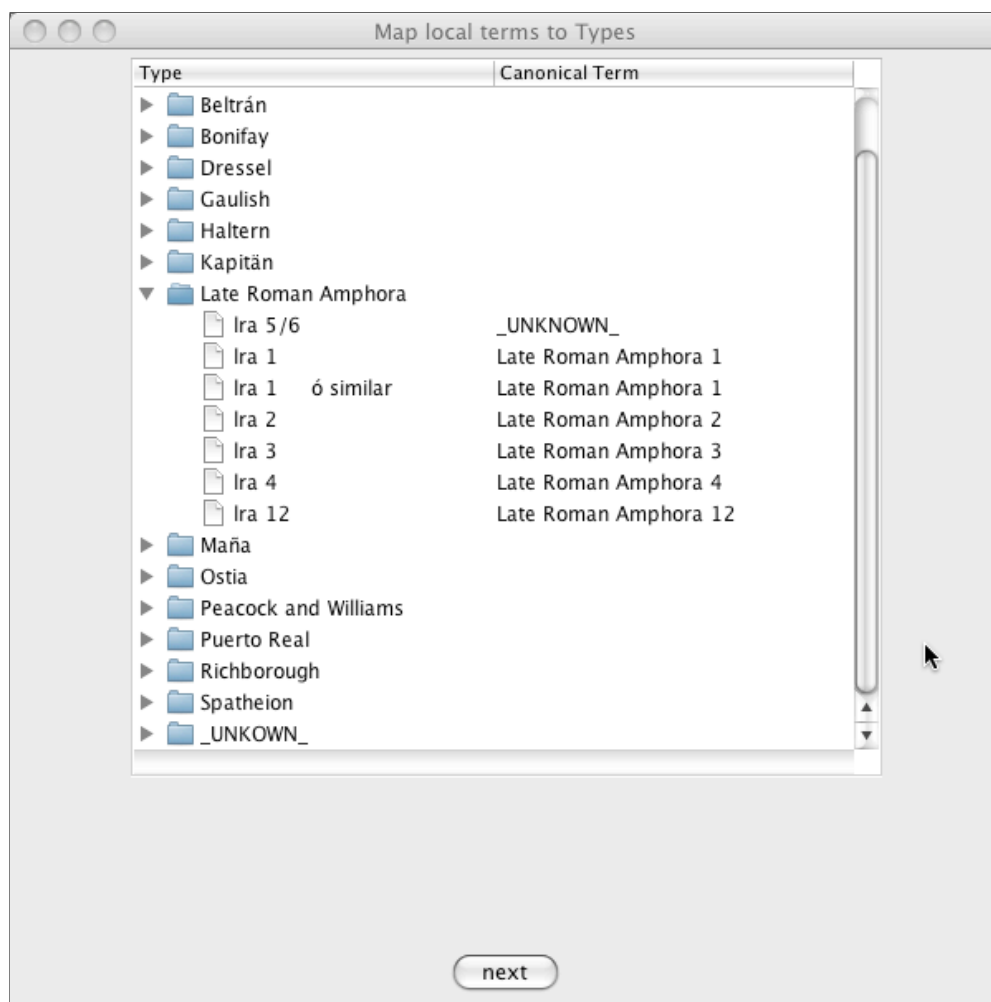


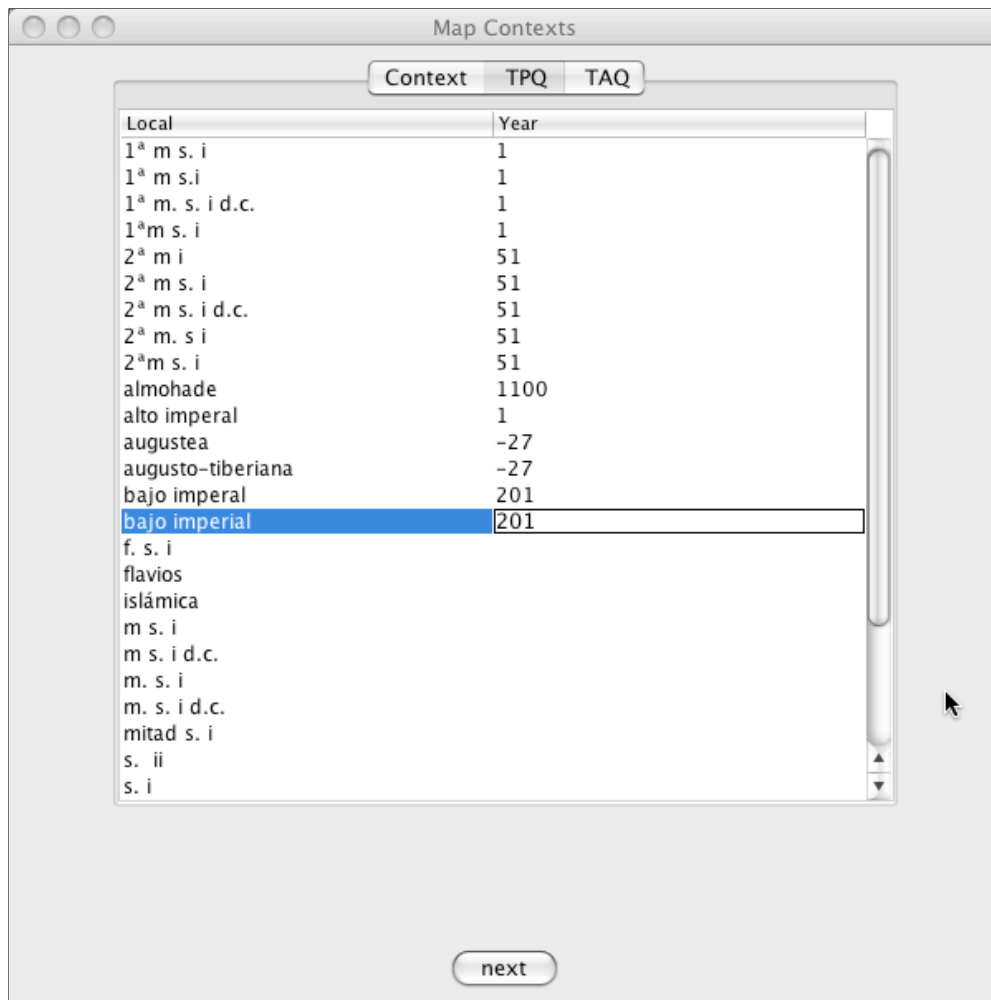
FIGURE 5.7: Data Inspector Wizard: Amphora type mapping

5.2.7 Literal Standardization

While URIs are a powerful means of specifying categories or instances, they are less useful for representing serial concepts (like numbers or dates), or arbitrary strings of text. At this information level it is better to use Literal values. As RDF only supports the use of Literals as the Object of a triple, they can only form endpoints in the aggregated graph. A number of the Properties expressed in the excavation ontology expect a Literal Object, but these will not necessarily be present in a useful way within the dataset. In these cases the Wizard may need to transform available values or request user input.

Example: Dates

Dating in the excavation ontology is associated with individual contexts in an excavation. A context may have a *terminus post quem* (the earliest possible date in which the creation of a context began) and a *terminus ante quem* (the latest possible date at which its creation ended). The Wizard allows the user to associate these

FIGURE 5.8: Data Inspector Wizard: Entering *terminus post quem* Literals

dates with the context using the Hempl RDF properties `heml:TerminusPostQuem` and `heml:TerminusAnteQuem`⁸ (Figure 5.8). While the source may give these in calendar years, more frequently the context will be associated with an historical period, such as ‘Augustan’ or ‘Late 4th century’. In these cases, the Wizard needs to derive the relevant date from the local term and this will vary depending on the nature of the property. For instance, the term ‘4th century AD’ would give a *terminus post quem* value of 301 CE but a *terminus ante quem* value of 400 CE.

Example: Quantity

Quantification is an extremely diverse, but important, element of archaeological recording. As different metrics will bias the observation in different ways, it is not possible to standardise this process. The different factors involved in quantification — including

⁸These are defined as an `rdfs:Property` in the HEML RDF description despite the initial capitalisation.

the method, unit and value — can be represented by creating URIs for each measurement, but over large bodies of material this greatly increases the complexity of the RDF produced. The solution presented here is to create Properties that are the `rdfs:subPropertyOf` `archvocab:hasQuantity`. Each of these subproperties is specific to individual quantification techniques such as **Estimated Vessel Equivalent (EVE)** or **Minimum Number of Individuals (MNI)**. This allows data consumers to quickly identify the quantification metrics available and, if necessary, filter and deal with them separately.

5.3 Stage 2: Export

On completion of the Data Inspector Wizard, an XML configuration file will have been fully generated sufficient for a fully-automated **Extract-Transform-Load (ETL)** RDF generation process to be undertaken. This contains mappings between:

- The source schema and the excavation ontology Resource types.
- Relevant local terms and canonical URIs or appropriate Literal values.

A second prototype Java tool, the *Data Importer*, automatically generates RDF from the database taking the configuration file and any datasource logon details as its parameters. Minor database changes, such as new records using the same local classification terms, can be handled without any changes to the configuration file being necessary. Structural changes, or the introduction of new local terms, can easily be taken account of by editing the configuration file. Currently this must be done with a text editor but it is envisaged that the Data Inspector Wizard could be further developed to permit reloading and editing of configuration files. In either case, maintenance is minimal — a vital consideration for archaeologists.

The RDF generated is in two forms. The basic output is an RDF/XML document immediately available to the data providers themselves. If the namespace provided to the Data Inspector Wizard is the same as a website they control, the document can be posted just as one would post a webpage. This makes it instantly accessible to other researchers who can then dereference the URIs for each context or find. As Roman Port Networks is generally a closed environment, the RDF is imported into a central Jena SDB⁹ triplestore using a PostgreSQL¹⁰ database backend, which provides enhanced performance, security and querying functionality for project partners. Each dataset is tagged with the Configuration ID provided to the Data Inspector Wizard. This makes updates simply a case of deleting all the triples in one such subgraph and replacing it

⁹<http://jena.sourceforge.net/SDB/>

¹⁰<http://www.postgresql.org/>

with the new ones. As long as the Configuration ID is changed, the file could also be used to import data from other databases that have the same schema and local terminology (as might be the case in, say, a commercial archaeology practice).

5.4 Stage 3: Representation

Consumption and representation is a particularly important topic because researchers who are unable to benefit from RDF have no incentive to generate it. Although research work is being undertaken in this area by another PhD researcher at the University of Southampton Archaeology Department, it was necessary to demonstrate to Roman Port Networks partners the potential benefits of using semantic technologies. In order to do so, a third Java tool was created to convert RDF into formats with which archaeologists are likely to be better acquainted. The tool produces two separate outputs. The first is a tabular representation of the data exported as a CSV file. This benefits the researcher in two ways — first, it is a format that is much more suitable for comparison with the original datasource. Second, and in contrast to the source data, it can be instantly combined with the same tabular output from any other datasource that has been processed by TRANSLATION.

In order to demonstrate some of the additional benefits of RDF, the tool also generates a JSON file of finds from each excavation. As it is able to retrieve the coordinates for each place automatically from the GeoNames service, these can be added to the dataset. The JSON file can then be displayed as a webmap using the SIMILE project Exhibit API¹¹ which in turn provides the additional benefits of filtering locations by amphora form (Figure 5.9). A temporary demonstration of this has been set up at <http://archvocab.net/demo/romanports.html>.

5.5 Evaluation

The software was trialled throughout the course of a series of visits to Roman Port Networks partners¹² in the summer of 2009. The author worked in conjunction with the archaeologists in order to explain the tool suite's purpose and functionality and then allow them to apply it to real data. The process was typically very successful from a technical point of view. Certainly it was possible to convert all the datasets provided into an RDF format with only a small number of records proving to be intractable. This was followed by a lengthier conversation in which specific shortcomings were identified, some of which were relatively minor :

¹¹<http://www.simile-widgets.org/exhibit/>

¹²In Aix, Barcelona, London, Naples, Pisa, Rome, Seville and Tarragona.

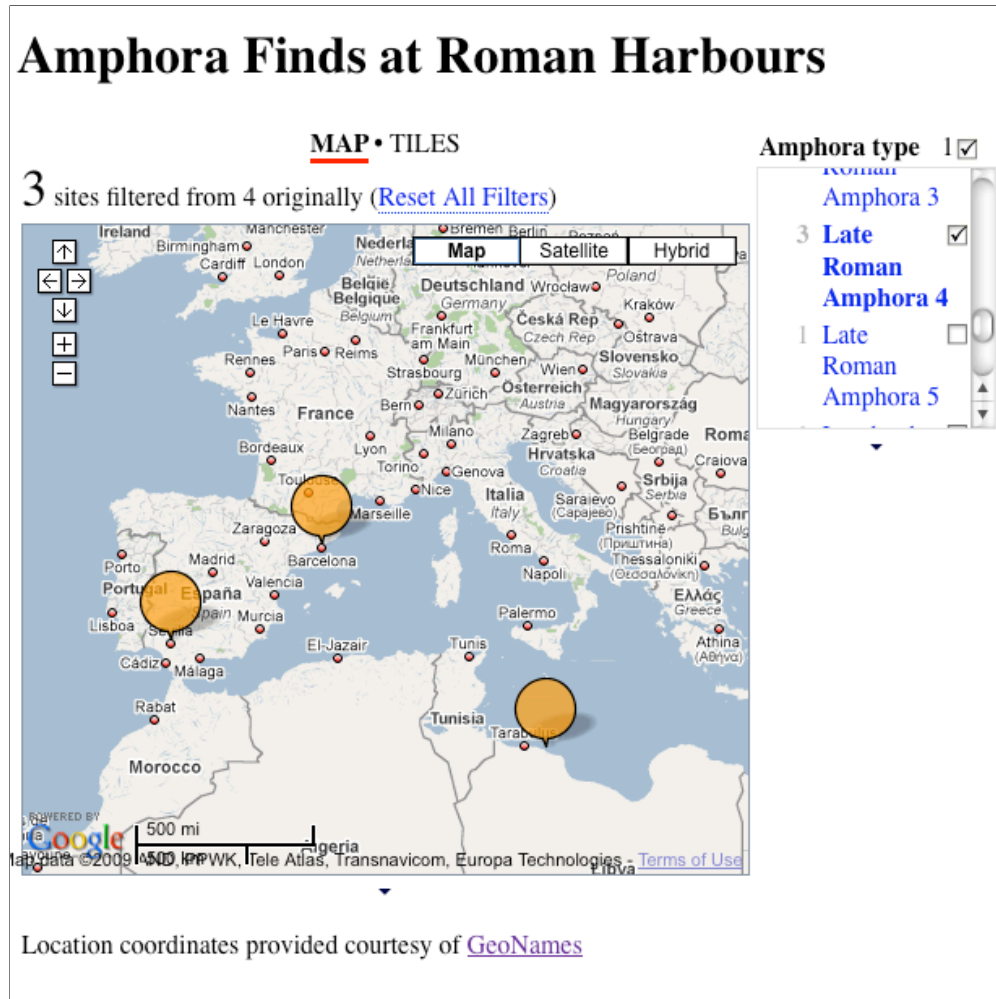


FIGURE 5.9: Results of RDF generation displayed in SIMILE Exhibit

- The need to implement additional quantity types.
- Adding descriptive comments into the XML config file to assist editing.
- Adding missing thesaurus terms.

However, these seemed to be symptomatic of a number of wider issues which proved to make the TRANSLATION approach ultimately unviable in its current form. Most importantly, these were overwhelmingly *social* issues rather than technical ones. Indeed, very few technical problems arose that did not appear susceptible to a moderate level of technical intervention. In contrast, it was highly apparent that TRANSLATION suffered from several usability problems that made its practical adoption, even within a fixed-term MSKR project, highly unlikely.

Learning vs. Outcome imbalance

A particular problem with the tool-suite was the failure to determine at the outset that it essentially performs a ‘one-off’ function and is largely project-specific. This

became particularly clear when, having demonstrated its use to the partners using their own data there was no additional data to practice on, nor, indeed, was there any need to practice! This illustrated a fundamental imbalance in the approach: even with a sophisticated tool to assist them, the user has a considerable amount to learn about both the basic concepts and the tool itself, yet they will rarely need to use it more than once. As there is no inherent benefit in learning how to use the tool itself, a heavy reliance is placed on making its output useful to the user.

Lack of ‘traditional’ benefits

Unfortunately there were also very few immediate and personal benefits to the partners, although all of them could see the potential advantage to the project as a whole. Converting the data to RDF created a form of ‘black box’ from which it was possible only to either a) convert the data back into a CSV file, which eliminates the majority of benefits of RDF, or b) use generic tools with superficially interesting functionality but little practical value for real-world analysis. Few, if any, of the archaeologists had the technical background that would enable them to utilise RDF directly. In short, the process takes human-readable data with which the archaeologist is comfortable and converts it into a machine-readable format with which they are not.

Overly rigid infrastructure

As noted above, there were occasions when additional amphora forms (and other categories) had to be added to the SKOS thesauri. Although any given partner was unlikely to introduce more than two or three additional categories, they were often important to their own dataset. Furthermore, because only a limited number of partners are involved in the project, it does not seem likely that a point would be reached in which new partners benefit from the investment of earlier contributors, as might be the case in a Linked Open Data environment. Although an additional tool could be developed to edit the SKOS files, this would require yet further investment in learning on the part of the user. The alternative is to centralise the process through the involvement of a digital ‘super-user’ who is responsible for maintaining the thesauri. However the need to contact such a person in the middle of the mapping exercise each time a new amphora form was encountered could potentially draw the process out enormously, greatly reducing the benefits of having a compact workflow. At a more general level, it is clear that the Wizard itself cannot easily be adjusted to deal with other material types by anyone without experience in software development.

5.6 Conclusions

A workflow that imposes considerable effort on people to produce outcomes they cannot work with is clearly not viable in a collaborative and voluntary environment. The question to be answered is whether these problems are due to the TRANSLATION approach or semantic technologies themselves. There is unquestionably inherent complexity in the process that cannot be simply be laid at the door of the CIDOC CRM or other complex ontologies. A number of independent and non-trivial aspects of the mapping task need to be undertaken and these require direct engagement from a domain expert. While there are clearly plenty of ways in which TRANSLATION might be improved over time — in terms of both usability and documentation — there is no escaping Berners-Lee’s observation that *“it involves asking people to make the extra effort”*. If we cannot avoid making this extra effort then the alternative must be to try to balance the other side of the equation and make the output more beneficial.

One way to do this is to develop more powerful visualization and analysis tools. That is beyond the scope of this thesis, but it is worth remarking that there are still limits to what such tools can achieve. At one end of the spectrum are tools that provide a small number of highly specific functions. In such a case they may be simple to use but require technical expertise to develop and will have a limited capacity for research. This is problematic in an environment in which there is a need to open up avenues of research broad enough to support more than a handful of people. There is also a tendency for the limits of the tool to be equated with the limits of the new data format which will not help foster enthusiasm for it. Alternatively tools can be made more flexible and powerful, but at the cost of increasing development time and the steepness of the learning curve. End-user tools are therefore part of the solution but unlikely, in themselves, to fully redress the balance. So is there any way of improving things at the publication end?

Shipman and Marshall (1999) identify five principles by which the difficult leap from unstructured to structured knowledge can be assisted:

1. That designers work with users to understand the nature of the usage situation,
2. That designers evaluate the cost/benefit trade-off of introducing any new formalism,
3. That designers should support incremental transition to formalisms for more complex tasks,
4. That designers provide facilities that use automatically recognized (but undeclared and potentially inaccurate) structures within data to assist the user,
5. That training and facilitation are a vital part of user adoption.

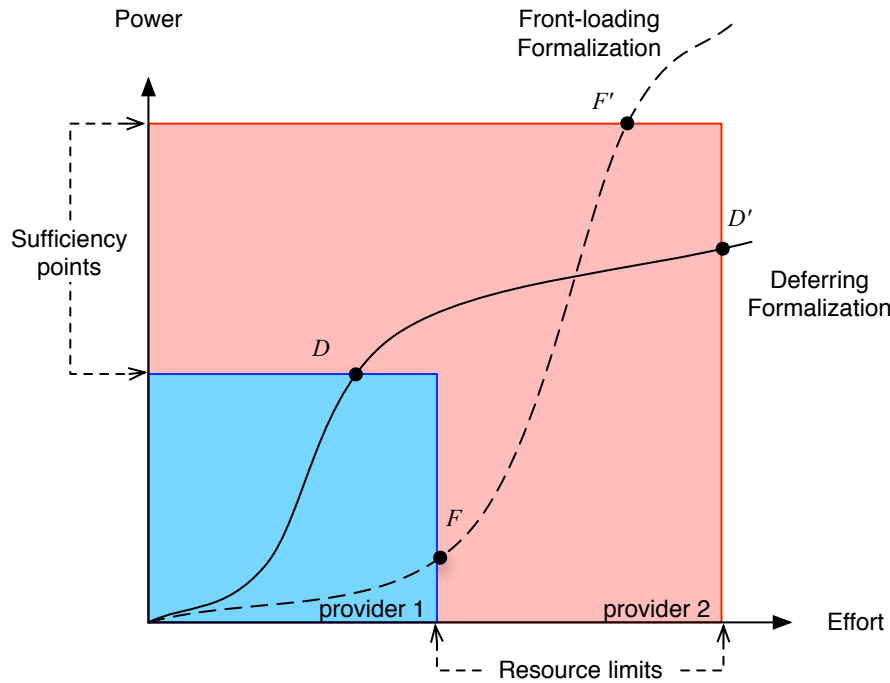


FIGURE 5.10: Formalization benefit variance across different users

So far, we have tried to assist users by identifying automatically recognized structures in their data (#4). Unfortunately, in working with users to understand the nature of the usage situation (#1) we have also discovered that the cost/benefit trade-off is heavily in deficit! (#2). Training and facilitation (#5), while clearly vital, seem equally unlikely to sufficiently mitigate the problem, as the payback for them largely accrues over repeated use which is not the case in this kind of process. The one principle which we have not hitherto considered is the third one: *“that designers should support incremental transition to formalisms for more complex tasks.”* This was an important limitation of the TRANSLATION approach. By attempting to get the user from A to B as quickly as possible, it cut out any potentially useful secondary benefits along the way and, to some degree, reduced the user’s engagement and understanding of the process as well.

We can think of this problem in a more abstract way. Every provider makes a cost-benefit decision based on the realities of their situation — the effort they are willing or able to expend versus the computational power required to achieve their individual goals. Effort beyond one’s means is impossible, and redundant computational power is undesirable if achieved at additional cost. Ideally, a provider will expend the least level of effort required to reach their sufficiency point so different formalization technologies have different utility curves in order to suit different user needs. **Formalization deferring** technologies (such as spreadsheets) have a high power-to-effort ratio early on which then rapidly degrades. **Formalization front-loading** technologies (such as RDF) require high initial investment in order to improve computational power later. This play-off between effort and power is shown as a diagram in Figure 5.10.

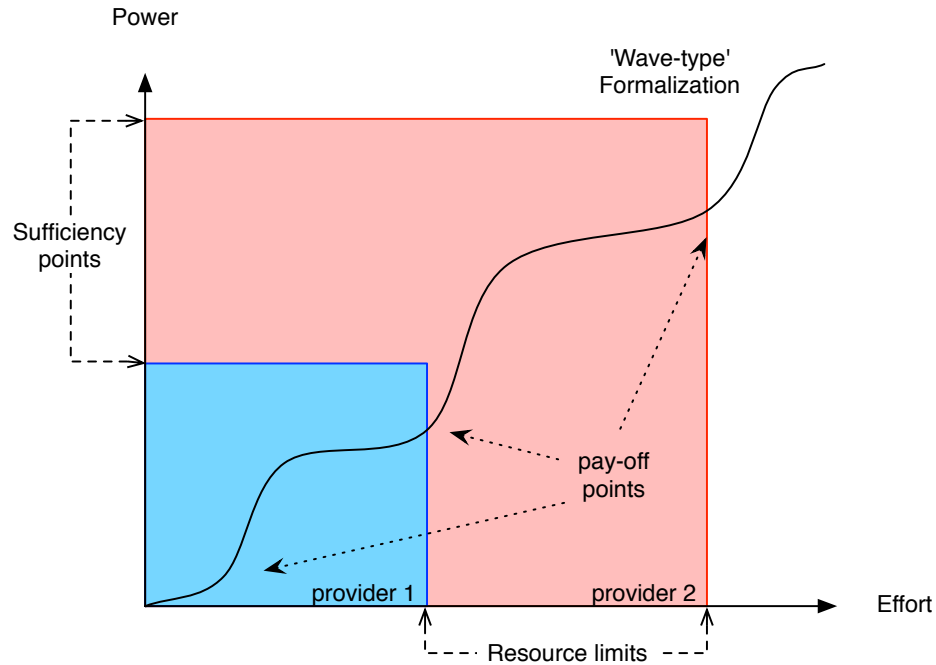


FIGURE 5.11: Wave-type formalization process

Provider 1 (blue) has few resources and low requirements. A formalization front-loading process will not meet their sufficiency criteria before their resources are exhausted (F) whereas a formalization deferring process will meet their needs at low cost (D). Provider 2 (pink) has both greater needs and resources. A formalization deferring process will ultimately become too intractable before sufficiency is reached (D') whereas a formalization front-loading process is both achievable and has longer-term potential (F').

When only a single provider is involved they can sometimes find a solution which is ‘just right’ for them, unanticipated future needs notwithstanding (Brooks, 1987). In a multi-contributor context the challenge is to accommodate the varying resource levels and requirements of each data provider. As essential formal complexity itself cannot be reduced, the only way to do this is to provide multiple pay-off points, creating a ‘wave-type’ formalization process (Figure 5.11). While this imposes long-term overhead on all contributors, it helps to maintain a preferable cost-benefit ratio throughout the whole formalization cycle, thereby helping both to ‘bootstrap’ the early process and maintain the benefits of scale. In the following chapter we describe a radically different approach to TRANSLATION — entitled *Introducing Semantics* — that attempts to test this principle.

Chapter 6

Introducing Semantics — A Semantic Cookbook

6.1 Overview

Despite the fact that TRANSLATION had only limited success for semantic mapping, the experiments conducted in Chapter 5 clearly demonstrate the need to very carefully balance the complementary, but limited, resources available within most MSKR contexts. These can be identified at three different levels: **URI infrastructure**, **local computing experts** and **local domain experts**. The URI infrastructure provides the vocabularies, services and tools necessary to unify heterogeneous data. While it is possible to create this infrastructure for individual projects — in terms of both content and functionality — sharing and reusing it across projects has the advantage of creating a point of both technical and methodological intersection which helps focus future discussion on areas of common interest. Shared URIs also act as the ‘semantic glue’ that makes the Web of Linked Open Data possible. Local computing expertise consists of those people involved in a project that can creatively make use of digital techniques to solve domain-specific problems. Such people are inevitably in very short supply so it is therefore imperative to focus their effort wherever possible on those tasks which cannot be undertaken by others. Domain experts are those who provide the theoretical and procedural knowledge required for a project. The discipline-specific nature of most funding programmes means that they usually form the largest contingent of a project’s staff. The ubiquity of desktop computing — at least in economically developed countries — means that basic computer literacy (i.e. the ability to type and interact with a human-readable interface) is commonplace. However, additional computing skills, such as the ability to write computer scripts, interpret low-level output or anticipate and solve complex digital problems are much rarer.

Thus, a well-balanced socio-technical system will use local computing expertise to facilitate the interaction between local domain experts and the global infrastructure as much as possible, while minimizing dependency. As this will almost certainly require closer interaction with digital technology than the domain experts are accustomed to (perhaps even comfortable with), every effort must be made to keep incentives clear. Furthermore, as many of the tasks required of semantic mapping are often carried out only once (or rarely), tasks for domain experts cannot impose a steep learning curve. This is best achieved by a staged process that is broken down into individual tasks, each with their own clear purpose. Tools and processes should be intuitive enough that minor deviations from the (inevitably) generic approach should be possible without advanced computing knowledge.

As a number of different domain experts are likely to have to work on similar tasks, a common digital workspace is helpful for providing instruction, soliciting feedback and uploading results. This also allows for a fast feedback cycle in which comments and suggestions can be rapidly incorporated into the workflow. While Web-based systems are naturally ideal for distributed project teams, there may be the additional necessity of maintaining such information in an environment with restricted access. Finally, it is necessary to provide the domain experts with the tools to make some use of the new data format as soon as it has been converted. This should ideally be possible at each stage of development as it is primarily through interacting with the data that they will best be able gauge whether the conversion process has been successful.

6.1.1 *Introducing Semantics*

In an attempt to meet the criteria cited above, a radically different method, entitled *Introducing Semantics*, was trialled, inspired by the ‘cookbook’ genre popular in Computer Science literature. The frequent need for computer programmers to undertake one-off tasks quickly when coding led to the development of this textbook format, and some — notably *Refactoring* by Martin Fowler et al. (1999) and *The Linux Cookbook* by Michael Stutz (2004) — have become classics in the field. The central principle is to deal with common tasks or problems in individual sections (‘recipes’) that explain the central issues to be considered, describe relevant use cases and provide a set of concrete steps by which they can be addressed. The name *Introducing Semantics* is deliberately ambiguous, implying both an expectation that the user will be expected to have no prior knowledge of semantic technologies, but also that conversion will take place one stage at a time rather than *en bloc*.

There are a number of advantages to such an approach. The first is that users have immediate practical instructions for dealing with the given task without requiring extensive theoretical discussion beforehand. This greatly reduces preparation time and potential confusion for the reader. Second, the independent and explicit description of

each task makes it more susceptible to automated processing. When done well, this can prompt very rapid development of complementary toolkits. The publication of *Refactoring*, for instance, prompted many software **Interactive Development Environments (IDEs)** to incorporate functions that explicitly mirrored the book’s recipes. Tools developed in this manner need not provide all the functionality described in the cookbook precisely because it can act as a fallback for unimplemented functions. Third, a cookbook enables users to understand why they are undertaking a specific task, over and above the instructions for how to perform it. This makes it much easier for them to adapt or even ignore it all together, depending on their individual circumstances. Given the diverse nature of Archaeology and the materials it encounters, this is an extremely important advantage. In contrast, it is likely to be beyond the capability of a domain expert to make even small changes to a fully digital tool.

Introducing Semantics is not quite a true ‘cookbook’ insofar as the tasks themselves follow a natural order and are not entirely independent of one another. It is split into three consecutive but self-contained **Semantic Levels**. Each introduces different aspects of semantic technologies into the dataset and makes it easier to undertake new analyses and visualizations. The process is not intended to change the original data and operations are performed on a copy of the original dataset. Furthermore, the process is language independent (an important consideration in Archaeology) but for the sake of simplicity, English terminology is used throughout the guides by way of example. The three levels are:

Semantic Level 1: Literal Standardization

Semantic Level 2: Introducing URIs

Semantic Level 3: Introducing RDF

Over the entire course of recipes, a user should be able to convert almost any conventional tabular record of Roman amphorae data to Turtle RDF, using canonical URIs for both the amphora form, origin and excavation location. Each stage achieves a subgoal however, building upon the stages that precede it. The first Semantic Level, *Literal Standardization*, prepares the ground for semantic technologies, addressing many of the problems which make the semantic conversion process difficult. Neither URIs nor RDF are employed at this stage. The second Semantic Level, *Introducing URIs*, explains why URIs are useful for interconnecting data and provides recipes to both a) discover canonical URIs for common concepts, and b) mint URIs for local concepts. The URIs remain embedded in a familiar spreadsheet format however. The final Semantic Level, *Introducing RDF*, enables the user to convert their URI-enhanced spreadsheet to RDF for hosting or ingestion into a triplestore. At the completion of each Semantic Level, suggestions are given for novel ways the user can interact with the data that were not previously possible.

6.1.2 Infrastructure

Introducing Semantics makes extensive use of several freely available tools and services in order to facilitate a domain expert's interaction with the data and infrastructure. The working platform used is **PBWorks**,¹ an online collaboration environment with extensive wiki and filehosting functionality. This provides a secure environment in which domain experts can follow the recipes, give feedback and upload datasets. Versioning of both the wiki and filesystem also means that users are able to make changes with confidence. In the case of Roman Port Networks, a project workspace was created and all members of the project were given access rights.² Subdirectories for the original data files, and all output from each Semantic Level, were also created so that project partners could see their own work in comparison with that produced by others.

Spreadsheet editing is done using the **Calc**³ software package that forms part of the OpenOffice⁴ suite of productivity applications. Although the great majority of archaeologists use Microsoft Excel⁵, two factors prohibited its use. The first is that using an Open Source alternative means that *Introducing Semantics* remains open to anyone at no cost. The second is that macros are used by several recipes and Excel does not provide cross-platform support for these. Using Calc made it possible to provide macro scripts that could be used by any archaeologist within the project. Unfortunately, the switch from Excel to Calc that most archaeologists will experience does add an additional level of cognitive overhead.

Two inter-related products are used for managing URIs and aligning data with them: **Freebase**⁶ and **Google Refine**.⁷ Freebase, discussed briefly in Section 2.2.2, is an online semantic encyclopaedia that provides an ideal platform on which to host canonical URIs for abstract concepts such as amphora forms and typology systems. It combines a relatively intuitive human interface for importing and editing data with a sophisticated query API and direct access to the raw RDF. Google Refine (formerly Gridworks) is a tool that serves a dual function. It is able to normalize tabular data very efficiently (several of the recipes for Semantic Level 1 can be carried out extremely easily once the data is loaded into it) and also provides a user-friendly interface for aligning local terms with URIs in Freebase. The ability to leverage such functionality greatly reduces much of the work that would otherwise be necessary for this process and is another major reason for using Freebase as a repository for abstract concepts.

¹<http://pbworks.com/>

²<http://romanportnetworks.pbworks.com/>

³<http://www.openoffice.org/product/calc.html>

⁴<http://www.openoffice.org>

⁵<http://office.microsoft.com/en-gb/excel/>

⁶<http://www.freebase.com/>

⁷<http://code.google.com/p/google-refine/>

6.1.3 Layout

Each Semantic Level begins with an overview that makes clear the overall goal and lists the recipes in their natural sequence. Each recipe follows the same format: First a **Time** indicator informs the user how long the task might be expected to take.⁸ The **Purpose** section explains the reasons for carrying out the task. After this comes an **Example** of how the data might look before and after processing. Finally, the **Recipe** gives the sequence of steps to be followed. Figure 6.1 shows a typical example. The following sections will describe the three Semantic Levels, with a brief discussion of each of their recipes. This will be followed by a general evaluation of the entire approach.

One of the principle reasons for dividing conversion into separate Semantic Levels is that it enables the user to stop at a point suited to their own needs. Unfortunately, as this process is inevitably a trade off between the rich, human-readable content of the input and the impoverished, but machine-readable nature of the output, losses are inevitably more conspicuous to the user than the gains. Concrete examples of visualization and analysis techniques opened up by the conversion process are therefore provided at the end of each Semantic Level in order to help address this tendency. These visualization recipes prompt the user to consider new ways in which the data is amenable to exploration and investigation but are not intended to be an exhaustive list of the possibilities.

⁸This is largely for motivational purposes as users may expect the work to be arduous and lengthy which is rarely, if ever, the case.

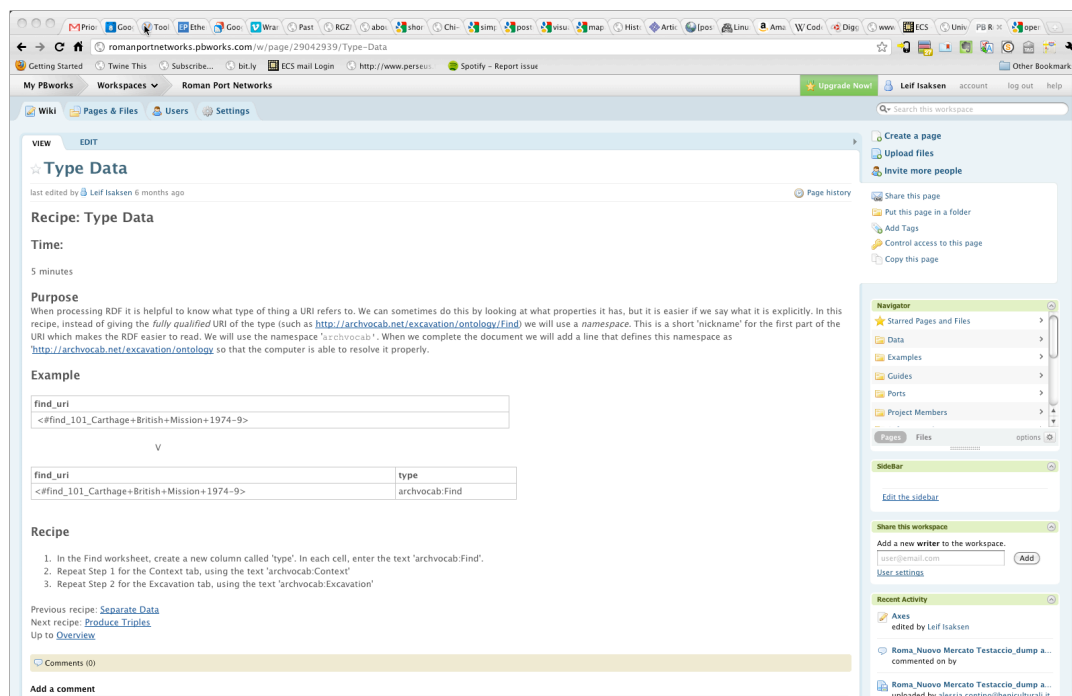


FIGURE 6.1: An example of a recipe in PBWorks

6.2 Semantic Level 1: Literal Standardization

6.2.1 Overview and Aims

Semantic Level 1 is predominantly intended to enhance the consistency of *local* data. That is, to assist analysis that does not require the consideration of other datasets. It aims to achieve two important goals:

1. Increase the internal consistency of the dataset,
2. Standardize aspects of the dataset structure to facilitate further semantic processing.

More specifically, it will:

- Ensure that only one term is used for a given concept,
- Standardize the structure by which Context dates are recorded (but not the dates themselves),
- Ensure that only one ‘Find’ is described per table row,
- Prepare the data for Semantic Level 2.

The primary benefits are:

- Quick visualization of a single dataset using visualization software and services such as spreadsheet graphs or Many Eyes,⁹
- Easier comparison with datasets that have also been processed in this way.

This Semantic Level makes use of spreadsheet software with which most archaeologists are already likely to be familiar. Microsoft Excel, OpenOffice Calc and Google Spreadsheets are all viable tools for the recipes in Semantic Level 1. None of the recipes should be difficult for those acquainted with spreadsheets but the amount of time required will largely depend on the complexity of the original data. A typical timeframe would be 1–2 hours from start to finish.

⁹<http://www-958.ibm.com/software/data/cognos/manyeyes/>

6.2.2 Recipes

Recipe 1: Create a Copy

None of the recipes are intended to change the original data and recipes should only be performed on a copy of the dataset. This recipe creates a copy of the data and uploads it to a specific directory of the PBworks website, which has an extremely simple interface for this function. The original data is typically, although not always, in the form of a spreadsheet. In cases where the data is held in a database, it is recommended that it be exported as a CSV file before importing into a spreadsheet package. The user is also cautioned to check for common problems such as missing column names, character encoding issues (especially for accented or non-latin letters) and the conversion of decimal numbers to integers (or vice versa). Naturally it is impossible to provide concrete guidance for the wide variety of data formats in which the data may originate, so the main intention is to raise awareness of such problems and, if encountered, direct the user to consult the appropriate software documentation. Before uploading, the user is encouraged to use an appropriate naming convention for the file. Note that additional metadata is not incorporated here as the goal is ultimately to encourage users to create self-describing data as efficiently as possible. Extensive discussion of what to include, and under what circumstances, would inevitably add considerable additional complexity from the outset. However, users are expected to follow appropriate metadata procedures for both their original dataset and any outputs they intend to make public.

Recipe 2: Table Axes

In order to have a common structure for processing it is necessary to ensure that the axes of the spreadsheet are such that finds — the fundamental entity to which other information will be attached — are associated with rows rather than columns. Finds, discussed in Section 4.4.1, are the smallest unit which it makes sense to compare across sites. In order to mitigate the inevitable ambiguity of using common archaeological terms across different disciplinary traditions, the following definitions are also provided:

Find The aggregation of all fragments of one amphora Class within one Context

Class The combination of an Amphora's Form and Fabric

Form The physical shape of an amphora, specified by category in a typology system

Fabric The geographic origin of the material with which an amphora is composed

Context A unit of an archaeological Excavation

Excavation A season of archaeological excavation at a particular site

The subsequent recipes require that there should be one row in the spreadsheet for each Find or Context and its associated information (amphora Form, Fabric, Context, dating, location and so on). Some datasets, especially summary tables, list the amphora Forms as rows, and the Contexts as columns, for concise viewing. If this is the case with the user's data then the tables need to be transposed (i.e. their axes need to be inverted) by following this recipe. Sometimes category labels will only be found in the first row and implicit for successive rows. If this is the case, users are also instructed to fill in such rows explicitly.

Example Input¹⁰

form	context_1	context_2	context_3
Keay 5	5	3	
Dressel 1		1	7
LRA 3	2		

Example Output

context	keay_5	dressel_1	lra_3
Context 1	5		2
Context 2	3	1	
Context 3		7	

Recipe 3: Filter Undesired Content

As the amount of time needed to complete each Semantic Level depends heavily on the amount of variation in the data, rather than the amount of content, it is helpful to remove irrelevant material as early as possible. The nature of such extraneous content is of course impossible to predict but is typically Find material that is not a subset of the data category of interest. In the case of Roman Port Networks this might be Finds other than amphorae. It is important to re-emphasise that the goal here is not to semantically describe the complete original dataset, but only those parts which are relevant to the specific research agenda. This recipe requires the user to filter and delete rows appropriately (or columns in cases where Find classifications are separated this way). Summary data, such as count totals, is also removed.

¹⁰In the interests of simplicity, input and output examples only show information specifically relevant to the recipe under discussion.

Example Input

context	category	type	count
1	amphora	Ostia V	3
1	fineware	ARS	6
2	amphora	Keay 20	2

Example Output

context	category	type	count
1	amphora	Ostia V	3
2	amphora	Keay 20	2

Recipe 4: Add and Normalize Context Dates

Time is a longstanding issue of debate in archaeological database management and there is a great deal of variation in the way that Context dating is recorded. Frequently an approximate era will be given, or a *terminus post* and *ante quem*. In order to facilitate the production of time-series graphs and broad comparisons with other data, it is necessary to produce year-based dates for each Context. The only way to move from an imprecise dating term (such as ‘third century’) to a specific year without introducing inaccuracy is to extract the widest date range possible. The two essential dates are:

Terminus post quem (required where available): The earliest date at which a Context’s creation began.

Terminus ante quem (required where available): The latest date at which a Context’s creation ended.

If it is important to capture the original level of precision the following dates may also be included:

Inner terminus post quem (optional): The latest date at which a Context’s creation began.

Inner terminus ante quem (optional): The earliest date at which a Context’s creation ended.

It is rarely possible to perform meaningful analyses over large numbers of Contexts in this way however, so the external temporal bounds are most useful, practically speaking, even while being the most vague. An important issue highlighted here is that a period date (e.g. 2nd century CE) will give different values dependent on whether it is being considered as a *terminus post quem* (101 CE) or *terminus ante quem* (200 CE). No instruction is given as to the precise definitions of period terms ('Late Middle Bronze Age', 'mid-sixth century', etc.). This is because the user will inevitably be in the best position to judge the definition of those phrases in their own context. Imposing arbitrary generic definitions neither clarifies the original data compiler's intention or avoids ambiguity during analysis.

Unfortunately there are few time visualization technologies with good support for BCE dates, but by setting these as negative integer values it is possible to plot them adequately enough on a graph. Using integers also effectively converts years into a point date rather than a period. This is (almost) always sufficient for archaeological use. Date formats, in contrast, tend to lead to additional complexity in how they are recorded and interpreted because they must ultimately be reduced to a specific calendar-independent moment in time. This high level of granularity is seldom of use to archaeologists when comparing inter-site distributions over large time periods.

Example Input

context	earliest	latest
1	1st C.	Late 2nd C.
2	Late 2nd BCE	Early 1st C.

Example Output

context	earliest	latest	tpq	inner_tpq	inner_taq	taq
1	1st C.	Late 2nd C.	1	100	150	200
2	Late 2nd BCE	Early 1st C.	-150	-100	1	50

Recipe 5: Make Each Row Equivalent to a Single Find

Although useful for visualization in summary tables, it is very difficult to combine tables in which Finds from different Classes of amphora are given in separate columns and thus referred to in the same row. Where this is the case, it is necessary to produce a new row for each Class of amphora. This is, unfortunately, a relatively convoluted process, involving a considerable amount of copying and pasting. Fortunately summary tables of this nature tend to be comparatively small as they are generally laid out for visualization

across a single page. At the end of this recipe, any row with no content (such as that of Late Roman Amphora 3 in Context B in the example below) are deleted. Likewise, any border formatting that may have been present is also removed.

Example Input

context	key_5	lra_3
A	1	2
B	1	

Example Output

context	count	form
A	1	Key 5
A	2	LRA 3
B	1	Key 5

Recipe 6: Consolidate Counts

Raw archaeological data will often be recorded at the level of individual rims, handles, bases and sherds, or small assemblages of them. Although this can be very important for intrasite analysis it is at too high a level of granularity for inter-site analysis. As a Find is the sum of all fragments of a certain ceramics Class we only need totals.

Two recipes are given here, depending on the nature of the data. If sherd types are separated into separate columns, and there is only a single row for each Find, the process is relatively straightforward, and can be done using the `SUM()` function found in all spreadsheet packages. The second recipe, in cases for which there are multiple rows that need to be consolidated for each Find, is a little more complicated and time consuming, and uses several spreadsheet functions. The essential principle of consolidating counts should be relatively easy for the user to grasp however. In either case, the archaeologist may alternatively wish to calculate a standard metric, such as Estimated Vessel Equivalent, and use that figure in place of the raw count.

Example Input I

context	rim	base	handle	sherd	form
A	1	3	0	5	LRA 3

Example Output I

context	count	form
A	9	LRA 3

Example Input II

context	form	fabric	fragment_type	fragment_count
A	Dressel 20	Baetica	handle	3
A	Dressel 20	Baetica	rim	1
B	Dressel 2-4	Gaul	handle	3

Example Output II

context	form	fabric	fragment_count
A	Dressel 20	Baetica	4
B	Dressel 2-4	Gaul	3

Recipe 7: Uncertainty

Uncertainty, like time, is another complex topic when it comes to digital representation. Modeling uncertainty is extremely difficult largely because there are so many kinds and degrees of it. Most frequent, however, are i) possibility: an indication that a value may or may not be correct, and ii) disjunction: an indication that an attribute has either one value or another (but not both).

The first is often expressed with a modal operator such as a question mark (‘?’) or adverb (‘perhaps’, ‘possibly’, etc.) added to the description. As these operators cannot be distinguished by the computer as being independent of the values themselves, it is necessary to record them separately. The simplest way to deal with uncertainty of this nature is to create a new ‘uncertain’ field which can be either true or false. This way we can choose to filter out uncertain Finds if we wish to. We must also remove the question mark from the description so that it will not remain distinct from other Finds of the

same Class. This is by no means a perfect solution as by separating the uncertainty from a specific field it raises it to the level of the entire entity. For example, the semantics of the input may state that ‘3 amphorae were found in Context A. It is possible that they are of Form *Keay 5*’. The output, in contrast, must only be interpreted as: ‘It is possible that 3 amphorae of type *Keay 5* were found in Context A.’ Nevertheless this greatly simplifies the filtering process later and is in line with our philosophy of prioritizing accuracy over precision.

Multiple possibilities (e.g. ‘*Keay 13* or *Dressel 20*’) are in some ways more problematic. Representing them as separate statements would make it likely that they will be interpreted as two separate Finds unless complicated logical operators are introduced. Again, following the priority of accuracy over precision, the user may choose one of two options:

1. They may choose one of the alternatives for the value and mark it as uncertain (in the manner described above).
2. If the value is categorical they may change to ‘unidentified’ or a similarly ‘null’ value. In this case the **Normalize Terms** recipe (Recipe 9) should be followed.

The first option is generally preferable, as less information is lost, but psychologically the user may feel more comfortable giving no value at all than ‘hazarding a guess’.

Example Input

context	form
A	LRA 3?
A	Keay 5

Example Output

context	form	uncertain
A	LRA 3	TRUE
A	Keay 5	FALSE

Recipe 8: Separate Form and Fabric

In some datasets the description of Form and Fabric is recorded in the same column. These need to be separated into individual columns so that they can be interpreted separately. This recipe simply replicates the column — new terminology is established in the following recipe (‘Normalize Terms’).

Example Input

context	class
A	Dressel 2-4 (Gaul)
A	Dressel 20 (Baetica)

Example Output

context	form	fabric
A	Dressel 2-4 (Gaul)	Dressel 2-4 (Gaul)
A	Dressel 20 (Baetica)	Dressel 20 (Baetica))

Recipe 9: Normalize Terms

This is perhaps the most fundamental recipe in Semantic Level 1. A great deal of legacy data, and raw data in particular, uses varying terms to describe the same concept, often without realising it. This can be caused by a range of factors which may or may not be apparent to the human eye. Examples include:

- Capitalization ('beltrán 2')
- Numerals ('Beltrán II')
- Abbreviation ('Bel. 2')
- Whitespace ('Beltrán2')
- Accents and character encodings ('Beltran 2')
- Typographical errors ('Bletrán 2')
- Qualifying information ('Beltrán 2 (local)')

Machine-readability, with or without URIs, is inherently based on symbol-matching, i.e. the assumption that identical strings of characters refer to the same value. It is therefore necessary to ensure that terms are normalized so that there is only one term for each concept within the dataset. Note that for Semantic Level 1 it does not matter what term is used as long as consistency is maintained. Depending on the complexity and 'messiness' of the data this may be easy to do within a spreadsheet package or not. In simple cases the user is given instructions to sort the data based on each category field, checking to ensure that there are not multiple terms for the same concept. Single instances of terms are often a good indicator of spelling mistakes and similar anomalies.

This process is also used to define relevant terms for Form and Fabric if these have been separated out in Recipe 8.

If the data seems particularly complex it is recommended that the user download and make use of the Google Refine package which is explicitly intended for such work. Although alternative software packages exist, such as the Stanford DataWrangler,¹¹ Refine will also be used for URI mapping to Freebase in Semantic Level 2, so the additional time spent learning the software here is not lost. Specific instructions are given for a number of common tasks, including: installation; creating a new Refine project; undo and redoing actions; filtering and sorting; creating ‘Text Facets’; clustering terms together; removing whitespace and capitalization; editing multiple terms; exporting back to a spreadsheet. Refine is a relatively sophisticated tool, however, and providing a complete tutorial to the user would be beyond scope of the Cookbook. The user is therefore also directed to helpful online documentation¹² as well as the download site.¹³

Recipe 10: Add Find ID

At this stage the data should now share a common structure in which each row is equivalent to a Find and there are not multiple local terms for the same concept. This recipe requires the user to make a final consistency check to make sure that no duplicate records have been created and adds a unique local Find identifier to each row.

Example Input

context	form
A	Dressel 20
A	Dressel 20
A	LRA 3

Example Output

find	context	form
1023	A	Dressel 20
1024	A	LRA 3

¹¹<http://vis.stanford.edu/wrangler/>

¹²<http://code.google.com/p/google-refine/wiki/GettingStarted>

¹³<http://code.google.com/p/google-refine/downloads/list>

Recipe 11: Clean Up and Upload

The final recipe is a health check that the data now conforms to Semantic Level 1 and provides instructions for uploading it to the appropriate directory on the PBWorks server. The user should clear any extraneous formatting and ensure that they only have one row of headings. The following columns should remain, although the labels used for them may vary depending on the language and conventions of the user.

REQUIRED:

- `find_id` (integer)
- `context` (charvar)
- `terminus_post_quem` (integer)
- `terminus_ante_quem` (integer)
- `form` (charvar)
- `fabric` (charvar)

ONE OR MORE OF:

- `fragment_count` (integer)
- `weight` (integer)
- `minimum_number_of_individuals` (integer)
- `estimated_vessel_equivalence` (integer)

OPTIONALLY:

- `uncertain` (boolean)
- `inner_terminus_post_quem` (integer)
- `inner_terminus_ante_quem` (integer)

The file is then saved using a filenames template similar to that in Recipe 1 and uploaded to a directory of other datasets that conform to Semantic Level 1.

6.2.3 Visualization and Analysis

On completion of Semantic Level 1 the data is now considerably easier to visualize. Combining it with other datasets that have been processed to this level is also easier, although still not possible to do automatically. Two approaches are suggested for visualization. The first is the creation of an accumulation graph using spreadsheet software. The second is the use of freely available online visualization toolkits such as Many Eyes.

Visualization and Analysis 1: Accumulation graph

Although the data in Semantic Level 1 is much more consistent in its layout and terminology it still presents a number of challenges for graphing temporal trends. Most automated graphing techniques are not good for visualizing irregular time intervals — especially those with uncertain dates. There are at least two problems to tackle:

1. Most graphs require a value for each amphora Class at each time interval.
2. Most graphs expect equal intervals between time points.

The combination of these requirements make temporal uncertainty very difficult to represent on most graphs. However a **linear scatterplot graph** can provide a reasonable approximation to a time curve for multiple irregular series of values (Figure 6.2). With this method we can display both the *terminus post* and *ante quem* times for Contexts containing specific categories of amphora and project an approximate ‘time corridor’ for their deposition. This shows the accumulation of particular amphora Forms over time periods of increased or decreased deposition. By using normalized Context dates

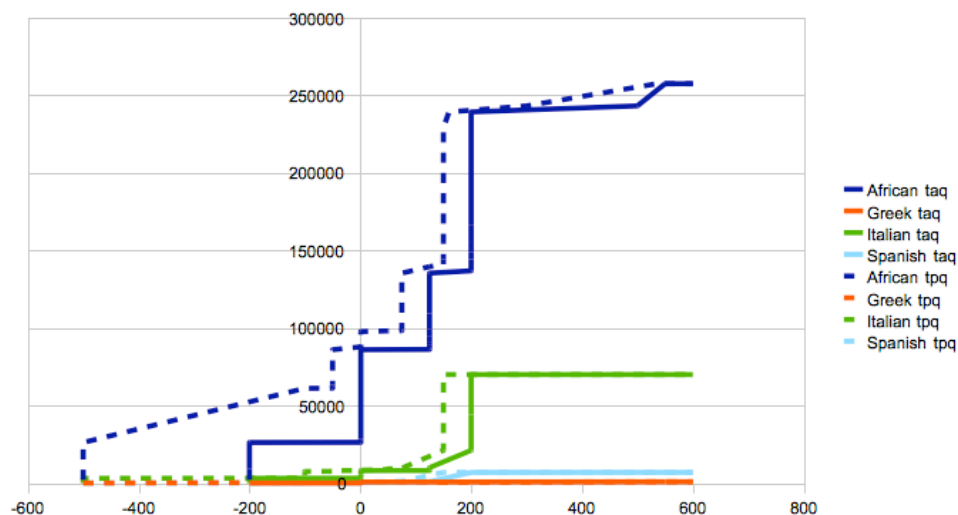


FIGURE 6.2: Example Accumulation Graph

and normalized weights and/or fragment counts, it is also possible to compare temporal distributions across different Excavations, even within the same graph. Caveats must be borne in mind however: First that this is of course just the deposition time — the amphorae may have arrived at the site much earlier and attempting to use these dates as a proxy for the transport period should be done with great caution. Second, the dataset will naturally be incomplete and there is no guarantee of how representative it is of the entire range of on-site deposits. The only the way to test this effectively is to compare distributions from different deposit series, which is precisely what such visualizations are designed to facilitate. In other words it is their inherent comparability with other data — rather than their immediate epistemic value — that makes them useful to the researcher.

Visualization and Analysis 2: Many Eyes

While most domain experts will be used to plotting data in a spreadsheet using a variety of graphing techniques, fewer are aware of the increasing number of **Information Visualization (InfoViz)** services available online. These vary considerably in terms of complexity and conditions of use and the intention here is merely to make users aware of them and their potential, especially as they are greatly easier to use with pre-normalized data. Many Eyes is a free online service that provides a wide range of such visualization tools. It is possible to upload a dataset and visualize it for free (Figure 6.3), but uploaded data is made visible to anyone on the Web.¹⁴

¹⁴Examples of amphorae from Carthage visualized in Many Eyes are available at:

<http://manyeyes.alphaworks.ibm.com/manyeyes/visualizations/amphora-types-by-date>

<http://manyeyes.alphaworks.ibm.com/manyeyes/visualizations/amphora-treemap>

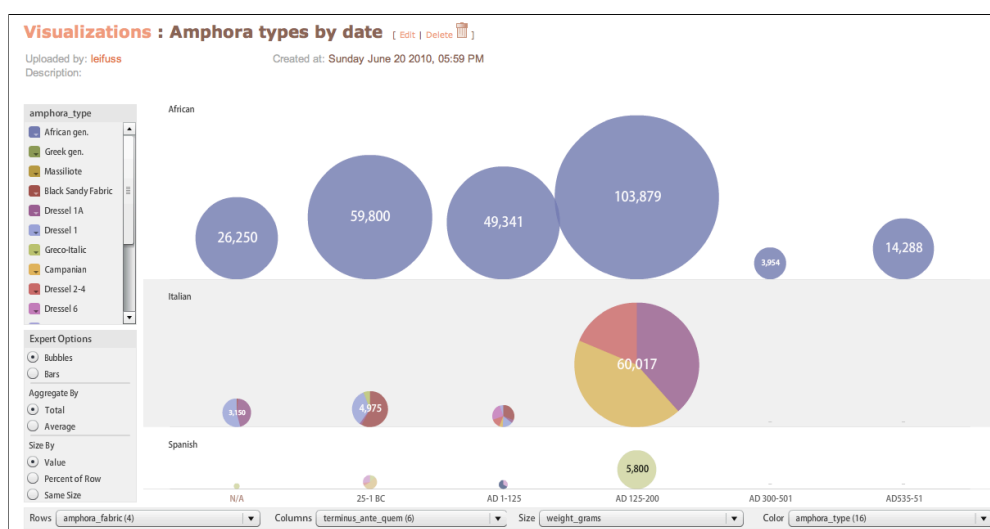


FIGURE 6.3: Amphora data from Carthage visualized in Many Eyes

6.3 Semantic Level 2: Introducing URIs

6.3.1 Overview and Aims

Semantic Level 2 is aimed at producing data for *collaborative* use. That is, data that can be used in conjunction with other known data sets. It aims to achieve one important goal:

1. Add HTTP URIs to a normalized spreadsheet document conformant with Semantic Level 2.

More specifically, it will:

- Attach pre-defined HTTP URIs to modern geographical terms,
- Attach pre-defined HTTP URIs to ancient geographical terms,
- Attach pre-defined HTTP URIs to standard archaeological category terms,
- Attach pre-defined HTTP URIs to column headings specifying the ontological category of their content,
- Prepare the data for Semantic Level 3.

The primary benefits are:

- Common identifiers for columns and standard terms make automated merging of datasets with other Semantic Level 2 datasets possible,
- HTTP URIs mean that data from external sources, such as images or geographic coordinates, can be introduced automatically to the data.

One of the challenges of using publicly available tools is that the technical landscape shifts over time and we are forced to adapt our approach to whatever functionality is available. This can happen rapidly, even over the course of a fixed-term project, and indeed the tools used here were not available until well after this research began. Semantic Level 2 uses Google Refine, an Open Source software package with which archaeologists are unlikely to be acquainted (unless they have used it for cleaning data in Semantic Level 1). The software is designed to be user-friendly to install and use, however. It also uses Freebase, an online, editable semantic encyclopaedia that provides an easily editable platform with which it is relatively simple for anyone to create persistent URIs and then edit the data associated with them. It is expected that both Refine and Freebase will change, and in all likelihood be superceded, in due course. The wiki format is therefore

ideal for this kind of approach as it enables such changes to be incorporated gracefully by domain experts as better practices become available. The following discussion will attempt to focus on the theoretical concerns of each recipe although inevitably it will be influenced to some degree by the practical concerns of working with these specific tools. The length of the entire process will, as always, depend on the complexity of the dataset, but assuming the data fully conforms with Semantic Level 1, it should take no longer than one or two hours to work through.

6.3.2 Recipes

Recipe 1: Add Excavation and Location

Most spreadsheets and databases do not record either the name of the Excavation or the place with which it is associated but this is important information for inter-site analysis. First, it enables us to distinguish between Contexts which may have the same name in different Excavations. As most Contexts are simply given a numeric or alphanumeric label, the chances of overlap are high. Second, it enables us to map the data geographically. This is because the great majority of both modern and ancient places now have their approximate geographical coordinates freely available online. These are not likely to be precise enough for local analysis but when combining datasets across a wide geographical area they can be extremely useful.

Would it be simpler to just record the coordinates of the Excavation instead? The answer to this question is no, for two reasons. First, coordinates themselves can only be understood in terms of a datum, such as OSGB36 or WGS84 (Survey, 2010). Recording and interpreting this information is a non-trivial process that typically requires a significant level of cartographic knowledge. Secondly, although precise coordinates are useful at a local level they can be less useful for global comparisons. For instance, it is not possible to deduce that two places are in fact the same, based on their coordinates, without relying on arbitrary buffering parameters that draw associations based on distance. If Place B is close to Places A and C, but A and C are both far from each other, one can still only assume *either* that they are all distinct *or* that they are all the same. There is no automated way to establish that A and B are the same, but that C is not, for instance. Even well defined spatial polygons can be problematic. The burials of ancient Rome are extra-mural (and thus beyond its geographic footprint) but this does not make them ‘non-Roman’. By associating Excavations with a place (whether ancient or modern) it is much easier to establish when they should be meaningfully associated with one another.

The purpose of this recipe is not to create or associate the data with a URI but simply prepare for that process. Arguably it is a task that could be included with Semantic Level 1, but as it has only marginal (if any) benefits when considering data locally, it is

more pragmatic, and better fits the cost-benefit equation, to include it here. The user is simply required to create two new columns which specify 1) the Excavation itself, and 2) the nearest known place with which it is associated.

Example Input

find	context
1023	A
1024	B

Example Output

find	context	excavation	location
1023	A	Keay-Earl Portus 20011	Fiumicino, Italy
1024	B	Keay-Earl Portus 20011	Fiumicino, Italy

Recipe 2: Search for Topic URI

One of the key processes in Semantic Level 2 is the discovery of relevant URIs in the Freebase vocabulary. Freebase URIs are assigned to **Topics**, essentially the equivalent of an encyclopaedia article. In turn, these Topics are classified under one or more **Types** which have specific properties associated with them that can be edited. This recipe is intended to acquaint the user with this process as well as some of its underlying principles. Of particular significance to these recipes is the fact that much of the following URI identification process is done with Google Refine. Although this software greatly enhances efficiency it clearly cannot find URIs which are not already present in Freebase and provides limited contextualising information. As a result it is helpful to give users a walkthrough of the Freebase website and its essential components before progressing.

The recipe takes the user to the main portal at <http://www.freebase.com> and asks them to search for the place(s) they have listed in Recipe 1 (Figure 6.4). It then explains where entity attributes such as Type, latitude and longitude can be found and edited if necessary. If either a URI for the place or its geographic coordinates are missing, the relevant recipes to rectify this (**‘Create a Topic URI’** (Recipe 3) and **‘Add Data to a Topic URI’** (Recipe 4)) are flagged up.

Using Freebase is not without its quirks and two are identified within the recipe, both caused by a mismatch between the flexible, graph structure of the data model, and the more restrictive nature of the interface. The first is that because Topics can be associated with multiple Types, but only one Type is shown in drop down menus, it is

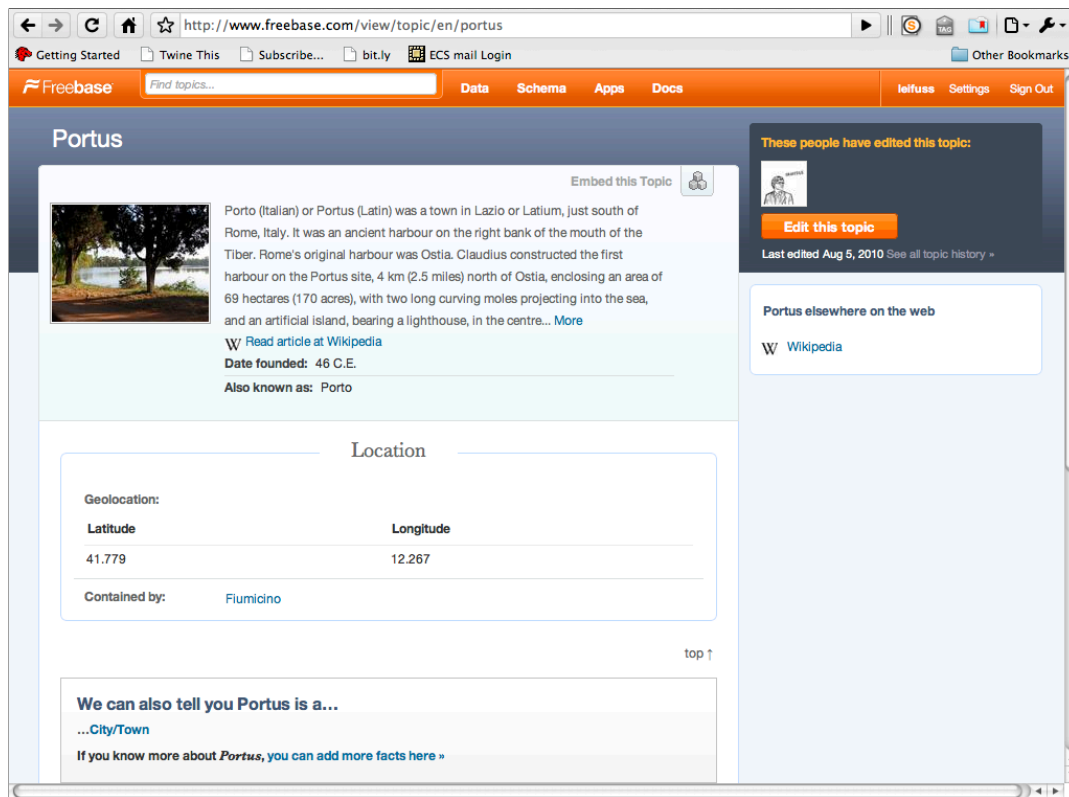


FIGURE 6.4: A Roman excavation site in Freebase

not always apparent that the Topic in question is the desired one. `Freebase:London` might be displayed as a `FreebaseVisualArt:art_subject`, for example. Occasionally the relevant entity can only be established by resolving the URIs themselves. Similarly, although Freebase supports the use of multiple names in multiple languages, in practice most of its place URIs only have an English toponym as much of the source data has come from the English edition of Wikipedia. This means that searching in English is almost always necessary, which can be a hindrance to speakers of other languages. If the user does not know the English name they are directed to GeoNames where it is usually easy to find the location in another language and then derive the English toponym for querying Freebase. Although this process sounds lengthy, it is unlikely that the user will be processing more than a handful of places and thus it does not have a significant impact in terms of time.

Recipe 3: Create a Topic URI

This recipe give users step-by-step instructions for adding a new Topic URI, such as an amphora Form, to Freebase. Freebase itself makes this process extremely easy with its online web interface, although the actual steps to be taken are not highly intuitive (thus the need for instructions). For instance, a user could create a Topic representing the Form *Dressel 20* and then specify that it was of Type `freebaseAmphora:amphora_type`.

A complete list of amphora Form Topics can be retrieved by simply resolving the URI of `freebaseAmphora:amphora_type`. Furthermore, because Freebase then knows that the Topic is a type of amphora Form it will automatically provide a set of corresponding properties — such as origin and contents — that can be filled in by the user.

The recipe does not cover bulk uploads of entities which are unlikely to be necessary for most users of the cookbook. However this is an issue which anyone intending to extend the process to cover other archaeological materials will be interested in. Freebase provides an interface for this called the **Freebase Loader**.¹⁵ A large number of `freebaseAmphora:amphora_type` Topics were pre-loaded into Freebase in this way based on data from the *Roman Amphora: A Digital Resource* database (with permission from the authors).

Recipe 4: Add Data to a Topic URI

Recipe 4 provides the user with guidance on adding additional information to a Topic, such as Type or geographic coordinates. One of the interesting issues this brings up is that of authority. While the creation of a new Topic URI may be seen as relatively innocuous because it did not previously exist, editing Topics which have been created by others may challenge people's conceptions of ownership, authority and provenance. Likewise, there is no 'sandpit' within which the user can trial their changes. Changes made within Freebase, like those in Wikipedia, will immediately impact anyone else making use of that URI. This level of exposure can not only be disconcerting to the unfamiliar, but in turn raises a further question: is information added by the user also open to editing and change? As Freebase is an Open Data environment it is not possible to restrict editing rights. However, it does maintain a 'Page History' for every Topic which is accessible from each page.¹⁶ Likewise, users must log on either with a Google, Yahoo!, or Freebase account, so edits cannot be made anonymously. For the time being it is unclear whether malicious or controversial edits will become a problem but there is as yet no evidence to suggest that it will.

Related issues of consideration are those of acknowledgement, copyright and academic credit. These issues work at two levels:

1. Adding data generated by the user
2. Adding data from a third party

In the first case, it might be questioned whether adding data to Freebase is essentially providing free labour to Google that in other contexts one would be acknowledged for.

¹⁵<http://data.labs.freebase.com/loader/>. Documentation available at: http://wiki.freebase.com/wiki/Freebase_Loader

¹⁶For example, <http://www.freebase.com/history/view/m/0c24fx6>

The answer to this is somewhat dependent on the user's perspective. Freebase is ultimately a resource that provides generic data for public use and thus the level of detail typically required to justify scholarly attribution may well be above and beyond what is required here. Likewise, there is something of an expectation that the user will benefit from the large quantity of freely provided data as a *quid pro quo* for contributing something in return. Whether the user should consider this sufficient reward for their labour is a question that clearly lies beyond the scope of this thesis.

Adding data from third parties is an important issue when contributing to a public repository. Freebase requires that the user make a statement to the effect that copyright third-party data is available under a license which permits its deposition, and the license holder must be cited so that details can be displayed alongside the content. It is the user's responsibility to establish the licensing arrangements for any third party information they provide. Users are unlikely to fall foul of copyright law accidentally however. US Copyright Law¹⁷ (under which jurisdiction Freebase falls) does not cover 'facts', and thus the type of information with which we are largely concerned, such as geographic position, can be added freely. One possible candidate for copyright are images of any kind. One example of this is the photographs and section drawings associated with the `freebaseAmphora:amphora_type` Topics. In such a case the user is asked by the Freebase service itself to directly assign both the appropriate license and the source of the content. Additionally the user may want to point to the source of the data by means of adding a URL to the 'Other related websites' attribute. A final point of interest is that Freebase content is itself released under a **Creative Commons Attribution License (CC-BY)**¹⁸ and so cannot be displayed in a public context without providing a direct attribution to Freebase.¹⁹ This doesn't preclude additionally attributing content to the original author as well of course.

Recipe 5: Import the Dataset into Google Refine

This recipe asks the user to install Google Refine (if they have not already done so for Semantic Level 1) and import their data into it (Figure 6.5). The issues here are principally practical ones, but there are one or two aspects worthy of comment. The first is that Refine runs as a local service accessed through a Web Browser. Archaeologists may be more familiar with traditional standalone software and the use of a browser may lead them to assume that by uploading content they are uploading it to the Web. It is therefore important to stress that this is not the case (for the same reasons it is important to emphasise that they *are* doing so with Many Eyes). A second issue specific to Refine is that it works best with MS Excel or CSV files, rather than OpenOffice Calc, and so users may be required to export their data to CSV before importing. Finally,

¹⁷<http://www.copyright.gov/help/faq/faq-general.html>

¹⁸<http://creativecommons.org/licenses/by/2.5/>

¹⁹<http://www.freebase.com/policies/attribution>

The screenshot shows the Google Refine interface for a dataset named 'encarnacion'. The top bar includes the Google Refine logo, the dataset name, and buttons for 'Open...', 'Export', and 'Help'. Below the top bar, there are controls for 'Facet / Filter' (Undo / Redo 15), 'Refresh', 'Reset All', and 'Remove All'. The left sidebar shows two active facets: 'TIPO' with 305 choices and 'FORMA' with 29 choices. The main table displays 2786 rows with columns: FORMA, TIPO, FRAGMTO, PASTA, and GRADOS BORDE. The table is sorted by 'name count'.

FORMA	TIPO	FRAGMTO	PASTA	GRADOS BORDE
ánfora	DRESS. 7/11	borde	BC	23
ánfora	BELTRÁN II a	borde	BC	45
ánfora	INDETERMINADO	pivote macizo	lusitana	
ánfora	K 23	borde	bética	58
ánfora	DRESS. 20	asa		
ánfora	spatheion	borde y cuello	amarilla	360
ánfora	DRESS. 20	borde		70
ánfora	k 19	pivote	bética	
ánfora	lra 1	asa		
ánfora	k 23	asa	bética	
ánfora	LRA 1	cuello		
ánfora	ÁGORA M273	borde		20
ánfora	K 19 Ó 21	asa	lusitana	
ánfora	K 23	borde y asa	bética	
ánfora	INDETERMINADO	asa		

FIGURE 6.5: A dataset in Google Refine

and as with introducing any new data management program, the recipe immediately makes reference to the **Undo/Redo** controls so that users feel more comfortable making changes.

Recipe 6: Search for Excavation Location URI

Recipe 6 introduces the process of mapping local terms in the dataset to persistent Freebase URIs. It starts by using the simple case of the place identified in Recipe 1, and which we know to be present in Freebase thanks to Recipes 2–4. This process is known as **reconciliation**. Users create a new column within the dataset for the results based upon the freetext name of the place. The Reconciliation Service then compares all local terms in this column (in this case there is only like to be one distinct term) and attempts to identify first the Type and then the specific Topic to which they refer. Users can either accept Refine’s suggestion or access a more sophisticated interface with which to search manually (this process is typically very fast assuming one knows that the URI exists). Refine enhances efficiency by mapping all identical local terms to the same canonical URI, thus the normalization process carried out in Semantic Level 1 pays large dividends here. Once the data has been reconciled it is represented as a hyperlink which takes the user directly to the appropriate Freebase Topic page in their browser.

Recipe 7: Search for Fabric Location URIs

This recipe introduces a more complex URI process — mapping the probable source region of the amphorae to Freebase URIs for Roman provinces. As discussed in Section 4.4.1, at inter-site level the Fabric of amphorae is primarily (if not entirely) of interest to archaeologists in establishing geographic provenance. As a result, mapping it directly to a place reduces complexity enormously and avoids the difficulties of comparing attributes such as ‘brown’ or ‘sandy’ which are rarely possible to compare meaningfully across geographically dispersed datasets. TRANSLATION used the GeoNames gazetteer which makes varying levels of geographic granularity possible. This is both a strength and a weakness for it greatly increases the likelihood that different URIs will be selected. For instance, the source of a particular Find might equally be associated with *Hispalis* (Seville), the River *Baetis* (Guadalquivir), *Baetica* (Andalucia) or *Hispania* (Iberia).

For several reasons it was therefore decided to use Freebase URIs for Roman provinces. The first was the pragmatic fact that Refine does not support direct reconciliation with GeoNames. As a result, to do otherwise would require considerable extra manual work. The second was that, when looking at the global picture, it is generally most helpful to be able to compare datasets at a similar level of granularity. The third is that GeoNames only contains contemporary locations, and not Roman provinces which are somewhat more meaningful as regional units. This is not to suggest that Roman provinces were either culturally or economically homogenous, but using the proxy of modern regional administrative units (such as Andalucia for *Baetica*) adds inaccuracy to the data. Topics for the Roman provinces already existed within the Freebase repository but have since been associated with the `freebaseRomanEmpire:roman_region` Type.²⁰ These have also been given approximate centroids to facilitate mapping. This is relatively unproblematic although it occasionally raises issues for geographically dispersed provinces such as *Creta et Cyrenaica*.²¹

Although a significant improvement over GeoNames for this task, Freebase does not always contain Late Imperial provinces and dioceses or other historically significant roman political boundaries such as *Hispania Citerior* and *Ulterior*. Granularity can also vary as there are Topics on Provinces (such as *Baetica*), regions (such as *Hispania*), and even supra regional blocs (such as the Western Roman Empire). However, mereological relations, such as ‘Contained By’ can be included in the data which may make limited inferencing possible. Another issue to bear in mind is that the naming conventions are not nearly as standardized as might be expected with modern names. *Armenia* is listed under its Hellenistic title, the Kingdom of Commagene, for example. Likewise the Roman province of *Asia* cannot easily be distinguished from its continental namesake by

²⁰They can be viewed at http://www.freebase.com/view/user/robert/roman_empire/roman_region

²¹http://www.freebase.com/view/en/creta_et_cyrenaica

name alone. For these more problematic cases it is easiest for users to identify the correct URI on the Freebase website before reconciling the data in Refine. Nevertheless, as each Fabric type must only be reconciled once the process is not overly time-consuming. It is also worth noting that the reconciliation process does not delete the original freetext field which remains available for examination and (where suitable) additional processing.

Recipe 8: Search for Amphora Form URIs

The third recipe for adding URIs to the dataset concerns the Forms of the amphorae. A large number of these Forms have been uploaded to Freebase, using the Freebase Loader, and have been given the Type `freebaseAmphora:amphora_type`. These are derived from the *Roman Amphorae: A Digital Resource* database. This database is itself a summary of the very wide range of possible amphora Forms and therefore does not include every name for `freebaseAmphora:amphora_types`, let alone variations in foreign languages.²²

Users are able to use the same process described in Recipe 7 to reconcile local terms with the URIs in Freebase. Assuming well-normalized data this is not a highly time-consuming task but is likely to be heavily reliant on the user knowing alternative names for the `freebaseAmphora:type_series` they use, as there is a strong possibility that some of their local terms will differ from the names in Freebase. This in turn will require them to consult the list of `FreebaseAmphora:amphora_types`. This is fairly lengthy, containing some 512 amphora types. It is also possible for the user to add a new Topic URI where one does not seem to be available. Naturally this risks increasing the probability that users will use separate URIs for what are in fact the same amphora Form classified under different schemes. This problem is better addressed by SKOS alignment however, and so the process of aligning `FreebaseAmphora:amphora_types` from different `FreebaseAmphora:type_series` is not considered in this recipe and will be addressed in Semantic Level 3. This recipe assumes that each `FreebaseAmphora:amphora_type` is distinct.

The `FreebaseAmphora:amphora_types` have had a small amount of additional data included from the source material where available such as likely sources of origin and contents. All the data should be considered indicative of the potential such an approach brings however, rather than an exhaustive resource. It is greatly hoped that the utility of making a core body of data available in this way will encourage those with expert knowledge to supplement it with further details over time.

²² Although on occasion non-English names are used in place of the English one (e.g. ‘Gauloise’ instead of ‘Gaulish’).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	context	tpq	taq	form	fabric	eve	excavation site	location					
2	B	550	650	Aegean amphora CC*	Asia	35	2007-Cirelli Ravenna	Ravenna					
3	A	400	550	Aegean amphora CC*	Asia	30	2007-Cirelli Ravenna	Ravenna					
4	B	550	650	Agora M273	Asia	71	2007-Cirelli Ravenna	Ravenna					
5	A	400	550	Agora M273	Asia	42	2007-Cirelli Ravenna	Ravenna					
6	B	550	650	LRA2	Asia	107	2007-Cirelli Ravenna	Ravenna					
7	A	400	550	LRA2	Asia	178	2007-Cirelli Ravenna	Ravenna					
8	B	550	650	Samo's cistern type	Asia	36	2007-Cirelli Ravenna	Ravenna					
9	A	400	550	Samo's cistern type	Asia	0	2007-Cirelli Ravenna	Ravenna					
10	B	550	650	African Big Cylindrical	Africa Proconsularis	998	2007-Cirelli Ravenna	Ravenna					
11	A	400	550	African Big Cylindrical	Africa Proconsularis	1070	2007-Cirelli Ravenna	Ravenna					
12	B	550	650	African Spatheia	Africa Proconsularis	143	2007-Cirelli Ravenna	Ravenna					
13	A	400	550	African Spatheia	Africa Proconsularis	750	2007-Cirelli Ravenna	Ravenna					
14	B	550	650	LRA3	Asia	624	2007-Cirelli Ravenna	Ravenna					
15	A	400	550	LRA3	Asia	34	2007-Cirelli Ravenna	Ravenna					
16	B	550	650	LRA7	Aegyptus	20	2007-Cirelli Ravenna	Ravenna					
17	A	400	550	LRA7	Aegyptus	140	2007-Cirelli Ravenna	Ravenna					
18	B	550	650	South Italy (K52)	Italian Peninsula	175	2007-Cirelli Ravenna	Ravenna					
19	A	400	550	South Italy (K52)	Italian Peninsula	445	2007-Cirelli Ravenna	Ravenna					
20	B	550	650	LRA1	Syria	374	2007-Cirelli Ravenna	Ravenna					
21	A	400	550	LRA1	Syria	32	2007-Cirelli Ravenna	Ravenna					
22	B	550	650	Nubian Amphora	Nubia	17	2007-Cirelli Ravenna	Ravenna					
23	A	400	550	Nubian Amphora	Nubia	71	2007-Cirelli Ravenna	Ravenna					
24	B	550	650	Agora M334	Iudaea	89	2007-Cirelli Ravenna	Ravenna					
25	A	400	550	Agora M334	Iudaea	535	2007-Cirelli Ravenna	Ravenna					
26	B	550	650	LRA4	Iudaea	892	2007-Cirelli Ravenna	Ravenna					
27	A	400	550	LRA4	Iudaea	125	2007-Cirelli Ravenna	Ravenna					
28	B	550	650	LRA5-6	Iudaea	267	2007-Cirelli Ravenna	Ravenna					
29	A	400	550	LRA5-6	Iudaea								

FIGURE 6.6: A spreadsheet exported from Google Refine showing hyperlinked contents

Recipe 9: Export Spreadsheet from Google Refine

This recipe provides instructions to export the data out of Google Refine and back into a spreadsheet so that it can be easily used again. The new spreadsheet will now contain hyperlinks to the URIs identified in the previous recipes (Figure 6.6). While this process is very simple, it is an important aspect of Semantic Level 2 that the data is not left in a format or data management system with which the user is unfamiliar.

Recipe 10: Add URIs to Header Row

Although local terms have now been mapped to Freebase URIs, it is still not possible to automatically determine how they relate to one another. This recipe gives instructions on how to embed hyperlinks that contain the ArchVocab URIs to which the columns relate. The URIs themselves are a mixture of both `rdf:Types` for Resources specific to ArchVocab (`archvocab:Finds`, `archvocab:Contexts` and `archvocab:Excavations`) and `rdf:Properties` for Literals and external URIs. Table 6.1 lists the mappings expressed as unqualified URLs so that a prefix definition is not required in the spreadsheet. Following the schema described in the last recipe of Semantic Level 1, not all columns themselves are mandatory and additional columns may be left in without URIs (although they will be ignored when being interpreted by machine).

TABLE 6.1: Semantic Level 2 Column URI mappings

Column	URI
find_id	http://archvocab.net/excavation/ontology/Find
context	http://archvocab.net/excavation/ontology/Context
terminus post quem	http://www.heml.org/rdf/2003-09-17/heml#TerminusPostQuem
terminus ante quemd	http://www.heml.org/rdf/2003-09-17/heml#TerminusAnteQuem
excavation	http://archvocab.net/excavation/ontology/Excavation
location	http://archvocab.net/excavation/ontology/atLocation_label
location uri	http://www.heml.org/rdf/2003-09-17/heml#locationRef
form	http://archvocab.net/excavation/ontology/ofForm_label
form uri	http://archvocab.net/excavation/ontology/ofForm
fabric	http://archvocab.net/excavation/ontology/ofMaterial_label
fabric uri	http://archvocab.net/excavation/ontology/ofMaterial
fragment count	http://archvocab.net/excavation/ontology/hasQuantityCount
weight	http://archvocab.net/excavation/ontology/hasQuantityWeight
m_n_i	http://archvocab.net/excavation/ontology/hasQuantityNMI
e_v_e	http://archvocab.net/excavation/ontology/hasQuantityEVE
uncertainty	http://archvocab.net/excavation/ontology/uncertain
inner_t_p_q	http://archvocab.net/excavation/ontology/innerTPQ
inner_t_a_q	http://archvocab.net/excavation/ontology/innerTAQ

Recipe 11: Check and Upload SL2

This recipe is similar to the final recipe of Semantic Level 1. It confirms the nature of the resulting dataset, and specifically that the appropriate columns are now identified by URIs and that the fields of the `location uri`, `form uri` and `fabric uri` columns now have contents hyperlinked to Freebase.

6.3.3 Visualisation and Analysis

Although the introduction of URIs makes little difference to the human-readability of the spreadsheet — many of the columns will continue to look much the same — it has a profound impact on its machine-readability. In particular, it is now possible to automatically:

- Combine datasets,
- Include external data,
- Produce geospatial visualizations.

Visualization and Analysis 1: Automatically Combining Datasets

One of the immediate advantages of using URIs to identify columns is that it makes it possible to automatically merge together datasets that conform to Semantic Level

2. Indeed the same process could be adapted to import them directly into a database for additional processing. Although the process will not be able to align substantively different kinds of information (such as different quantification methods, for instance), it will at least be able to keep them in separate columns. The ability to do this quickly and with minimal effort is clearly of great benefit to a researcher.

While the processing required is relatively simple and can be achieved using a spreadsheet macro, spreadsheet macros are themselves a complex topic. The advantage of using them is essentially that they can a) be written once and then easily downloaded and run by others, and b) modifications can be made by those with sufficient technical knowledge. Unfortunately, the most recent version of the MS Excel macro language, Visual Basic, is not compatible with operating systems other than Windows. As a result, an example Macro has been written for Open Office which can be run on any machine. There is no reason however why a similar macro cannot also be written for Excel. The Macro itself allows the user to select a directory containing Semantic Level 2-compliant spreadsheets and generates a new spreadsheet containing the combined data.

Example Input A

context	excavation	type	eve
1004	Carthage	Greco-Italic	17

Example Input B

nmi	u_e	sitio	tipo
4	356A	Hispalis	Dressel 20

Example Output

Context	ofForm	Excavation	hasQuantityNMI	hasQuantityEVE
1004	Greco-Italic	Carthage		17
356A	Dressel 20	Hispalis	4	

Visualization and Analysis 2: Automatically Including External Data

An additional benefit of using Google Refine is that it is able to directly import additional information from Freebase. In particular, the graph-like nature of Freebase allows for chains of content to be produced. For instance, the user can automatically create a new column (or columns) that displays the typical contents or geographic origin of each

`FreebaseAmphora:amphora_type` (where this data is available). This data only needs to be entered into Freebase once and can thereafter be easily incorporated by any other user. This ability to automatically augment a relatively small amount of initial data is a good example of how the power of Open Data can provide benefits which are truly beyond those which would be possible within a conventional closed database system.

Example Input

context	ofForm
1004	Dressel 20

Example Output

Context	ofForm	contents	origin
1004	Dressel 20	Olive oil	Hispania Baetica

Visualization and Analysis 3: Spatial and Temporal Visualization

Of even greater significance is that the data is part of a graph which can be traversed to discover other valuable information. For instance, an amphora Class's place of origin may also have latitudinal and longitudinal coordinates. These can be derived using exactly the same process.

Example Input

Context	ofForm	contents	origin
1004	Dressel 20	Olive oil	Hispania Baetica

Example Output

Context	ofForm	contents	origin	latitude	longitude
1004	Dressel 20	Olive oil	Hispania Baetica	37.9	-4.8

Depending on the technical skills of the archaeologist, powerful additional possibilities also present themselves. For instance, such data can be immediately imported into a **Geographical Information System (GIS)** for further analysis. **Web Mapping** tools

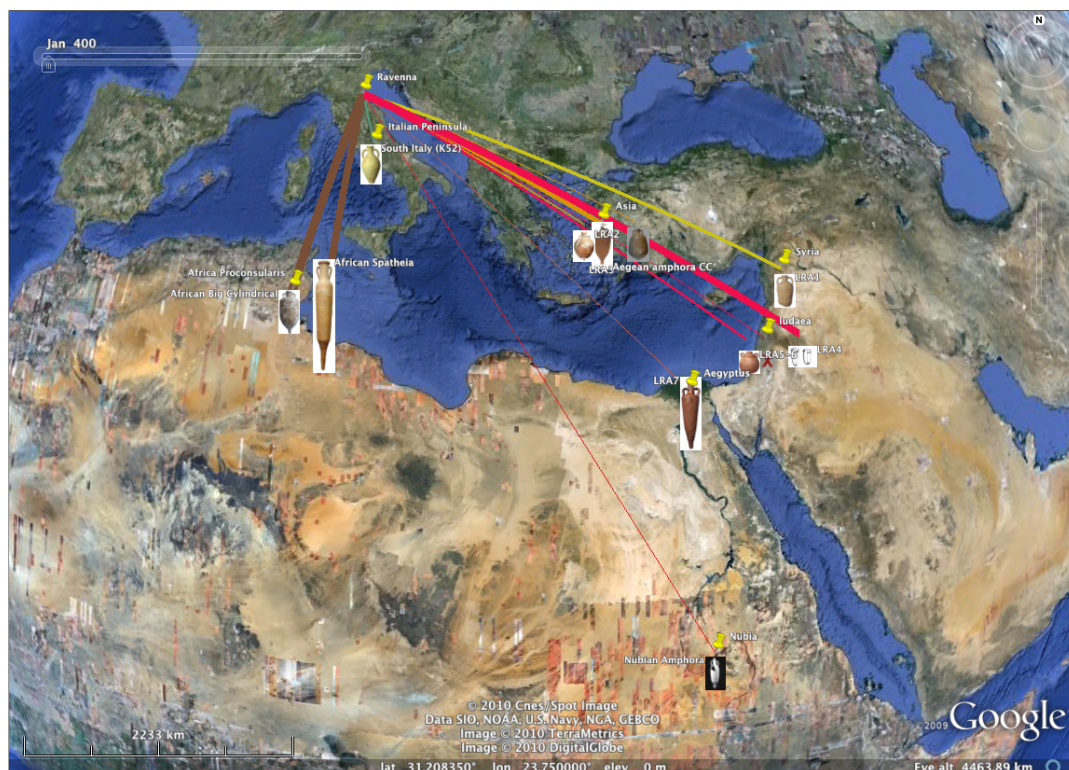


FIGURE 6.7: A dynamic network diagram showing the probable origins of ceramic finds from a Roman harbour

such as **Google Earth**²³ and **TimeMap.js**²⁴ can display both spatial and temporal dimensions. With data even in the limited machine-readable format created by Semantic Level 2, it is possible to write software that will convert it automatically into a **KML** file storing an interactive time-map such as that shown in Figure 6.7. This map displays a dynamic network diagram associating amphora Finds with their probable place of origin. Lines appear and disappear dependent on the timeframe of interest, the thickness represents volume, and icons represent images retrieved automatically from Freebase based on the Find's `FreebaseAmphora:amphora_type`. A prototype application for generating such files has been made available for download by the user.

Such visualization technologies are powerful but also have their limitations. Most of all their very intuitiveness can make it easy to forget what they represent. For example, the date range of the network links could be wrongly interpreted as the period in which the amphorae were transported, when of course this may be far from the case. Likewise, the use of image icons can give a misleading sense of the highly variable size of amphorae, as well as their shape.²⁵ Even the use of straight lines may suggest direct exchange rather than the possibility of a multi-stage process. There is one major benefit to such an approach however, which is that it greatly improves with increasing volumes of

²³<http://www.google.com/earth/>

²⁴<http://code.google.com/p/timemap/>

²⁵Those new to ceramology are often surprised to see how much variation there is within a single amphora Form

data. Whereas tabular data becomes increasingly difficult to interpret across multiple datasets, increasing quantities of data compliant with Semantic Level 2 should be able to highlight common phenomena more readily. While individual Finds may well be misleading, the cumulative effect of combining many datasets is to isolate those cases which diverge from the historical trend. Thus, such tools may be of limited help to scholars working solely with their own data, but are potentially of much greater use to large collaborative programmes and open research environments.

6.4 Semantic Level 3: Introducing RDF

6.4.1 Overview and Aims

Semantic Level 3 is aimed at producing data for *global* use: data that can be easily combined with other datasets without pre-arranged cooperation. It aims to achieve one goal:

1. Produce an RDF file describing the data in a way that can be instantly combined with other RDF files.

More specifically, it will:

- Map the column names defined in Semantic Level 2 to the concepts in the ArchVocab ontology,
- Mint HTTP URIs for concepts that are not defined by other data resources (such as Finds, Contexts and Excavations),
- Convert the data from a table structure into a graph structure,
- Save the data as a file describing the RDF graph.

The primary benefits are:

- RDF files can be instantly combined together into larger datasets,
- Powerful queries can be run over multiple datasets using the SPARQL query language,
- Concepts which are the same but have been classified differently (e.g. *Haltern 70* and *Camulodunum 185*) can be merged together easily.

While Semantic Level 3 can be accomplished with tools that are familiar to archaeologists (spreadsheet packages) and in a relatively short-timeframe (1–2 hours), the challenges it raises are in some ways the most significant of any of the Semantic Levels. First, although the data is left in a form that is ultimately more flexible, it is difficult to develop simple tools that make use of this flexibility without imposing new restrictions on the data. This in turn re-raises the issue of the benefit to the archaeologist. At the end of Semantic Level 3, two tools have been created that provide some idea of what might be done with the data. They do, however, only open up a small window on how it might ultimately be used.

Second, although Semantic Level 3 does not require the data to be hosted online, not doing so negates one of the primary benefits of converting data into RDF: the ability to seamlessly navigate between different data sources. Making data available online is not something that archaeologists are used to however, in terms of the technical requirements, issues of sustainability, or the ethos of direct public access. The latter in particular is an area in which archaeologists have many misgivings. Roman Port Networks, like many other projects in this domain, has found a half-way house in the use of a restricted-access website, which functions as an adequate repository for an MSKR project. Nonetheless, it needs to be flagged up that the benefits of Semantic Level 3 are in closer alignment with a Linked Open Data approach than MSKR, and the cost-benefit analysis for any specific project must be evaluated accordingly. While archaeological practice remains focussed on individual and collaborative projects however, it is as much as we can do to demonstrate those benefits on a reduced scale.

6.4.2 Recipes

Recipe 1: Mint URIs

For Semantic Level 2, external URIs were used to identify concepts shared in common. This makes it possible to combine different datasets together because all Semantic Level 2-compliant datasets use the opaque URI `freebaseRDF:m/0c24fx6`²⁶ to refer to *Dressel 20* amphorae, for example. We can also automatically gather additional information about Finds from the Web — such as place of production or contents — by dereferencing the URI (either in a browser or through an **Application Programming Interface (API)**). This capability exists because open vocabularies have already been created that contains URIs for all these concepts. In Semantic Level 3 users go a step further by minting new URIs for concepts which do not exist on the web already. In particular they create URIs for the `archvocab:Finds`, `archvocab:Contexts` and `archvocab:Excavations` referred to in the dataset. The result is that others can

²⁶The namespace `freebaseRDF:` resolves to `http://www.freebase.com/view/`. `http://rdf.freebase.com/ns/` would be preferable but Google Refine reconciles to the former.

i) combine datasets that refers to these URIs and ii) get additional information about them more easily.

In order to facilitate hosting, URIs are minted as **fragment identifiers** — the end-part of the URI. The namespace itself will depend on where the file is hosted. The full URI is the combination of the file URI + the fragment. For example, the full URI of a fragment called `<#context_123>` in the file `http://romanportnetworks.pbworks.com/f/excavation_abc.ttl` would be:

```
<http://romanportnetworks.pbworks.com/f/excavation_abc.ttl#context_123>
```

Recipe 1 uses the local terms in the dataset as a basis for the URIs. This is done by using a spreadsheet macro for OpenOffice Calc which the user can download. Thereafter template functions are given which generate the URIs by concatenating together URL-encoded values from the `archvocab:Find`, `archvocab:Context` and `archvocab:Excavation` fields. If opaque URIs are desired (i.e. those in which the meaning is not expressed within the URI itself, such as `FreebaseRDF:m/0c24fx6`), these could similarly be generated with a hashing algorithm, although there are no simple functions for this available in Open Office BASIC to the author's knowledge.

Example Input

context	excavation
4.16(a)	Carthage British Mission 1974-9

Example Output

context	excavation	excavation_uri
4.16(a)	Carthage.....	<#context_4.16%28a%29_Carthage+British+Mission+1974-9>

Recipe 2: Extract URIs from Hyperlinks

Recipe 2 is essentially a practical one as the spreadsheet has hyperlinks to external URIs embedded behind a human-readable label. It is necessary to extract the URI into a separate field so that both the URI and the label can be recorded in the RDF. Once again, this process is undertaken using a custom-written macro that is made available for download.

Example Input

form
Dressel 1

Example Output

form_label	form_uri
Dressel 1	<http://www.freebase.com/view/m/0c24_xn>

Recipe 3: Escape Literals

Recipe 3 is another straightforward recipe that wraps String Literals in inverted commas. Once again a spreadsheet function is used although this time no specialised macro is required.

Example Input

excavation_label
Carthage

Example Output

excavation_label
“Carthage”

Recipe 4: Separate Data

Every RDF statement requires a specific Subject. In the case of the Archvocab ontology they are `archvocab:Find`, `archvocab:Context` and `archvocab:Excavation`. It is therefore important to separate out the information that is relevant to each type of Subject (Table 6.2). This process is very similar to normalizing the structure of a relational database. The end result is:

1. A table of columns related to `archvocab:Finds`,
2. A table of columns related to `archvocab:Contexts`,

3. A table of columns related to `archvocab:Excavation(s)`.

This process reveals the high level of redundancy caused by representing it in a single table, so the recipe also removes any duplicate records created by the normalization process. It is quite possible for there to be as many records in the `archvocab:Find` table as there were in the original dataset, for instance, whereas there may only be a single record in the `archvocab:Excavation` table (assuming it contains data from only a single Excavation).

TABLE 6.2: Archvocab RDF Object Columns divided by Subject

Find Data	Context Data	Excavation Data
find_uri	context_uri	excavation_uri
find_id	context_label	excavation_label
fabric_extracted	tpq	location_extracted
fabric_label	taq	location_label
form_extracted	inner_tpq	
form_label	inner_taq	
hasQuantityWeight	excavation_uri	
hasQuantityEVE		
hasQuantityNMI		
hasQuantityCount		
uncertain		
context_uri		

Example Input

find_uri	context_uri	tpq	taq	form_uri	fab_uri	excav_uri	loc_uri
<#fin...	<#con...	150	250	<http...	<http...	<#exc...	<http...
<#fin...	<#con...	150	250	<http...	<http...	<#exc...	<http...
<#fin...	<#con...	200	350	<http...	<http...	<#exc...	<http...

Example Output A

excav_uri	loc_uri
<#excavation.../portus>	<http...

Example Output B

context_uri	tpq	taq	excav_uri
<#context...1>	150	250	<#excavation.../portus>
<#context...2>	200	350	<#excavation.../portus>

Example Output C

find_uri	context_uri	form_uri	fab_uri
<#find...a>	<#context...1>	<http...>	<http...>
<#find...b>	<#context...1>	<http...>	<http...>
<#find...c>	<#context...2>	<http...>	<http...>

Recipe 5: Type Data

The `rdf:type` of the Subject is not yet explicitly identified in every row. By including this information in the data it is easier for those consuming it to quickly identify triples of relevance to them. This recipe is very quick and simply adds canonical URIs representing the Subject's `rdf:type` to each row. The namespaced version used here is only to facilitate readability although it does require the addition of a prefix definition in Recipe 7. In practice this could be done as easily using fully qualified URIs.

Example Input

find_uri
<#find_101_Carthage+British+Mission+1974-9>

Example Output

find_uri	type
<#find_101_Carthage+British+Mission+1974-9>	archvocab:Find

Recipe 6: Produce Triples

Recipe 6 changes the layout of each table into a triple format. In other words, while it remains a spreadsheet, it is converted to a table of four columns in which each row consists of a Subject, a Predicate and an Object followed by a period ('.'). In principle

this process could be automated but much would depend on how strictly the user has followed the naming conventions for columns created in earlier chapters. In practice, the work is likely to be relatively brief as it predominantly requires cutting and pasting spreadsheet columns. The `rdf:Predicates` associated with each column are given in Table 6.3.

TABLE 6.3: Semantic Level 3 `rdf:Predicate` mappings

<code>rdf:subject</code>	<code>rdf:predicate</code>	<code>column</code>
<code>archvocab:Find</code>	<code>rdfs:label</code>	<code>find_label</code>
<code>archvocab:Find</code>	<code>archvocab:inContext</code>	<code>context_uri</code>
<code>archvocab:Find</code>	<code>archvocab:hasQuantityWeight</code>	<code>quantity_weight</code>
<code>archvocab:Find</code>	<code>archvocab:hasQuantityCount</code>	<code>quantity_count</code>
<code>archvocab:Find</code>	<code>archvocab:hasQuantityNMI</code>	<code>quantity_nmi</code>
<code>archvocab:Find</code>	<code>archvocab:hasQuantityEVE</code>	<code>quantity_eve</code>
<code>archvocab:Find</code>	<code>archvocab:ofForm</code>	<code>form_extracted</code>
<code>archvocab:Find</code>	<code>archvocab:ofForm_Label</code>	<code>form_label</code>
<code>archvocab:Find</code>	<code>archvocab:ofMaterial</code>	<code>fabric_extracted</code>
<code>archvocab:Find</code>	<code>archvocab:ofMaterial_Label</code>	<code>fabric_label</code>
<code>archvocab:Find</code>	<code>archvocab:uncertain</code>	<code>uncertain</code>
<code>archvocab:Find</code>	<code>rdf:type</code>	<code>type</code>
<code>archvocab:Context</code>	<code>rdfs:label</code>	<code>context_label</code>
<code>archvocab:Context</code>	<code>heml:TerminusPostQuem</code>	<code>tpq</code>
<code>archvocab:Context</code>	<code>heml:TerminusAnteQuem</code>	<code>taq</code>
<code>archvocab:Context</code>	<code>archvocab:innerTPQ</code>	<code>inner_tpq</code>
<code>archvocab:Context</code>	<code>archvocab:innerTAQ</code>	<code>inner_taq</code>
<code>archvocab:Context</code>	<code>archvocab:inExcavation</code>	<code>excavation_uri</code>
<code>archvocab:Context</code>	<code>rdf:type</code>	<code>type</code>
<code>archvocab:Excavation</code>	<code>rdfs:label</code>	<code>excavation_label</code>
<code>archvocab:Excavation</code>	<code>heml:locationRef</code>	<code>location_extracted</code>
<code>archvocab:Excavation</code>	<code>archvocab:atLocation_Label</code>	<code>location_label</code>
<code>archvocab:Excavation</code>	<code>rdf:type</code>	<code>type</code>

Example Input

<code>excavation_uri</code>	<code>excavation_label</code>	<code>location_uri</code>	<code>location_label</code>
<code><#excavation.Car...></code>	<code>“Carthage Bri...”</code>	<code><http://...></code>	<code>“Carthage”</code>

Example Output

<code><#excavation.Car...></code>	<code><rdfs:label></code>	<code>“Carthage British Mis...”</code>	<code>.</code>
<code><#excavation.Car...></code>	<code><heml:locationRef></code>	<code><http://www.freebas...></code>	<code>.</code>
<code><#excavation.Car...></code>	<code><archvocab:atLoc...></code>	<code>“Carthage”</code>	<code>.</code>

Recipe 7: Save as RDF

The final recipe exports the tables as textfiles which effectively constitute Turtle RDF. The contents of these files are aggregated together and a standard header of namespace prefixes is prepended. This data can now be converted to any other RDF format, such as RDF/XML, or ingested into a triplestore. Likewise, additional triples could be added providing metadata such as who created the document, licensing terms, and so on.

Example Output

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix heml: <http://www.heml.org/rdf/2003-09-17/heml#> .
@prefix archvocab: <http://archvocab.net/excavation/ontology/> .

<#find_100_Carthage+British+Mission+1974-9> rdfs:label "100" .
<#find_100_Carthage+British+Mission+1974-9> archvocab:inContext
<#context_1.2_Carthage+British+Mission+1974-9>.
<#find_100_Carthage+British+Mission+1974-9> archvocab:hasQuantityWeight
16800 .
<#find_100_Carthage+British+Mission+1974-9> archvocab:ofForm_Label
"African gen." .
<#find_100_Carthage+British+Mission+1974-9> archvocab:ofMaterial_Label
"African" .
<#find_100_Carthage+British+Mission+1974-9> archvocab:ofMaterial
<http://www.freebase.com/view/en/africa_province>.
<#find_101_Carthage+British+Mission+1974-9> rdfs:label "101" .
...
```

6.4.3 Visualisation and Analysis

As noted above, there are significant challenges involved in making RDF data directly usable by mainstream archaeologists. As a text format it is extremely difficult to visualize. Currently very few tools are capable of reading it, and those that can are so generic as to be of limited use for the specific requirements to which archaeologists would wish to put it. In short there is still considerable reliance on the development of bespoke tools that either represent it in new ways, or convert it back to a format that is better known to archaeologists.

The demonstration tools described here are an initial attempt at providing such functionality, although it should once again be stressed that they are not the principal focus

of this research and should be seen as indicative of the possibilities, rather than fully-fledged software solutions. In particular, while they have been developed in the Java programming language, it may well be more appropriate to create similar products in a scripting language such as **Python** so that they can be modified more easily. Together the tools allow the user to automatically:

1. Combine multiple datasets together so that they can be filtered and exported to a table or map representation.
2. Merge identical amphora Forms which have been classified using different typology systems.

Visualization and Analysis 1: Combine, Filter and Export Data

The RDF Explorer is a bespoke tool developed specifically to merge together RDF datasets created as part of Roman Port Networks. It consists of six separate panels (Figure 6.8).

1. The **Source Browser** allows the user to load or unload any of the RDF datasets that have been uploaded to the Roman Port Network PBWiki site.
2. The **Faceted Browser** enables the user to filter the data. The facets themselves are auto-generated and relate to a particular Subject (`archvocab:Find`, `archvocab:Context` or `archvocab:Excavation`) which can be selected from a drop down menu below.
3. The **Console** displays a system log reporting on which files have been loaded or unloaded and any errors encountered.
4. The **Data Panel** shows information about any Resources which have been selected in the Faceted Browser. The blue table generates a row for each matching Resource with columns based on triples in which the Resource is the Subject. The red table lists triples in which the Resource is the Object.
5. The **Management Panel** allows the user to filter content based on its `rdf:type`, switch SKOS alignment on and off (see next section), and export the data to other formats.
6. The **Info Panel** provides a more human readable version of the data selected in the Data Panel.

Data loaded within the RDF Explorer can be exported as a CSV file that can in turn be imported into a spreadsheet. The basic exported record unit is always a `archvocab:Find`

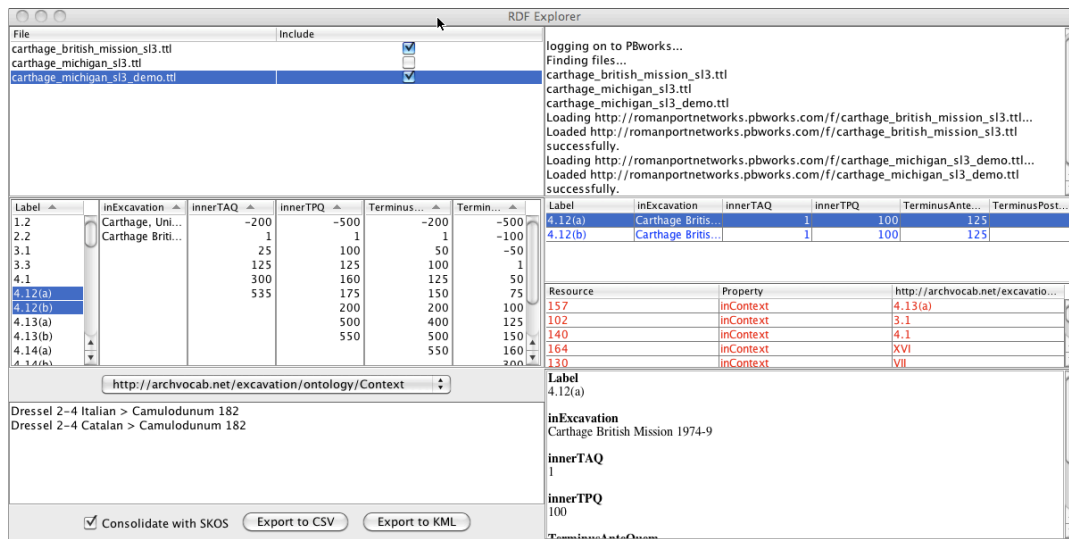


FIGURE 6.8: The RDF Explorer

(with its associated `archvocab:Context` and `archvocab:Excavation` information) regardless of the Resource type shown in the RDF Explorer. Thus, if the explorer is displaying `archvocab:Excavations` it will export a row for each `archvocab:Find` from those Excavations. Alternatively the data can be exported as a KML file similar to that shown in Figure 6.7 but displaying the user’s data of interest.

Visualization and Analysis 2: Consolidate using SKOS

Although all the local terms for amphora types have been mapped to a Freebase URI there remains the difficulty that many of the `FreebaseAmphora:amphora_types` in different `FreebaseAmphora:type_series` are in fact the same. SKOS helps us to resolve this problem by providing a vocabulary for mapping between the URIs. The Predicates `skos:exactMatch`, `skos:closeMatch`, `skos:broadMatch` and `skos:narrowMatch` allow us to say that two categories are either i) the same, ii) similar, iii) parent and child or iv) child and parent. By hosting a separate file that contains these mappings we can automatically merge them in the RDF Explorer with the click of a mouse. The file itself is hosted on the Roman Port Networks PBwiki site. It uses the `skos:narrowMatch` property because its asymmetry makes automated processing easier. Not only does it avoid problems of transitivity — whereby ‘chains’ of equivalence statements can have unexpected results — but even in cases of true equivalence it provides a preferred label to use (which otherwise cannot be inferred automatically). The format is:

```
<amph_type_A> <skos:narrowMatch> <amph_type_B> . #comment
```

Users are able to edit the file itself and upload it back the PBWiki site (it is automatically versioned). The editing must currently be done by hand and the system would

certainly benefit from a bespoke editing tool. Nevertheless, as with the data in Freebase, contributions made by each user are then freely available to all.

6.5 Evaluation

Introducing Semantics was developed in a wiki format as a living document written at a specific time and for the specific goals of Roman Port Networks. This makes it difficult to evaluate from a general perspective for several reasons. First, the rapid development of tools in this sector means that sections of it may be out of date very quickly. Indeed, some of the tools it makes use of were released only within the last two years and there is no reason not to believe that more sophisticated or user-friendly ones will arrive in due course. Naturally, however, it is only possible to work with the tools that are presently available, for better or worse. Second, it is heavily dependent on the skills of the researcher to write clear instructions targeted at achieving concrete objectives. These factors alone are entirely sufficient to cause the failure of a project, if not enough to guarantee its success. Moreover, any archaeological project has its own socio-technical complexities which may impact on progress. These things being said, how successful was *Introducing Semantics*, in contrast to TRANSLATION, for converting legacy data to RDF?

Superficially at least, success appears to have been extremely limited, based on the simple fact that very few Roman Port Networks partners were either able, or found it worthwhile, to convert their datasets. Of a possible 25 institutional partners engaged with ceramics, only three made known concerted efforts to convert their data to RDF using this method. Of these, one successfully completed only Semantic Level 1, a second also completed Semantic Level 2 and the third partially completed Semantic Level 3 as well. In summary, none of the partners managed to reach the point of producing full RDF descriptions. Should the cookbook approach therefore be written off as a failure? We cannot direct compare the differing levels of engagement between the TRANSLATION and *Introducing Semantics* approaches, as the former was always undertaken during pre-arranged visits by the author which naturally led to higher short-term levels of direct motivation. There are, moreover, some additional facts which suggest that the cookbook approach may still have some merit.

All partners were asked to complete a one-page questionnaire after completing the conversion process (Appendix H). Although the evidence base is naturally limited, it does provide some insight into the strengths and weaknesses of *Introducing Semantics*. First it is notable is that the Semantic Levels are seen to get consecutively more difficult. Whereas no partner had difficulties completing Semantic Level 1, Semantic Level 2 was deemed to be harder and the final Semantic Level was considered extremely difficult by the one partner known to have attempted it. Yet, excepting Semantic Level 3, data

was always converted correctly and accurately the first time round, without the use of bespoke tools or intervention by the author.

Based on further comments in the questionnaire, at least two reasons seem to be behind the usability barrier. The first is the use of unfamiliar technologies, and especially macros. In retrospect it may well have been better to have stayed with a more familiar spreadsheet package, such as Microsoft Excel, despite the potential restriction it would impose on a minority of users. All the same, there is no guarantee that custom-written macros would have proven successful in this case either. The other factor seems to have been a lack of good examples to work with. It was originally hoped that data posted by the partners could provide such material but a positive feedback loop was never attained. Two datasets from Carthage were pre-uploaded to the PBWiki site for each Semantic Level, which were considered very helpful. Again however, there were limitations caused by the specificity of such datasets, which may have features that do not apply to the user's data, and vice versa. Providing real examples of the transformed data for each recipe, as well as the examples in the wiki itself, might have helped to facilitate the process. In addition to these possibilities, and while not specifically mentioned in the questionnaires, it should not be forgotten that the majority of partners are not native English speakers and this must inevitably add an additional layer of difficulty to those already presented by the cookbook.

6.6 Conclusions

The case could therefore be argued in both directions: either that the cookbook approach is fundamentally flawed, or that specific problems with its execution — the software utilities used, the lack of concrete examples, the text itself — inhibited its use. There are reasons to suspect, however, that the problem goes deeper than this. The first and most fundamental one is that extremely few project partners attempted to convert data to Semantic Level 1 although this was universally acknowledged to be straightforward (and while time estimates of 1–2 hours seemed a little low, they were not unduly so. 2.5 hours was cited in one case). This is probably not a consequence of having a technical background, as none of the partners who completed it have extensive experience in informatics. The second reason is that none of the partners who converted data found the Visualization and Analysis tools very useful — their utility scored 3 on a Likert scale of 1–5 in two responses, with no response on the third questionnaire form. Finally, those partners who did convert data seemed comfortable to send it to the author via email, but not to upload it to the restricted-access PBWiki site.

Together with the overwhelming lack of engagement, this seems to suggest not that the cookbook is too difficult *per se*, at least as regards Semantic Levels 1 and 2, but rather that once again there is not enough perceived benefit to those doing the work to justify

what is clearly still a challenging task. Furthermore, while it is unquestionably the case that the wiki and workflow could be improved in a variety of ways based on user feedback, this returns us to the difficulty faced by the TRANSLATION approach. In a closed collaborative environment in which tasks need only be carried out once (or rarely) it is hard to reach an efficient process before the work has already been undertaken. Even worse, there are no additional advantages for those who participate early on to justify having to overcome the inevitable teething difficulties. In the final chapter we will explore this issue in more detail in order to consider the third question that we posed in the Introduction: to what extent is contemporary archaeological practice able and willing to meet the social and technical requirements of semantic technologies?

Chapter 7

Discussion, Conclusions and Future Work

7.1 Review

In the previous six chapters we have addressed a number of specific issues related to the three questions posed in the Introduction. In this final chapter we will review the line of argument so far before turning to a more general discussion that arrives at some clearer, if inevitably provisional, conclusions. This will be followed by a number of concrete suggestions for future work. Let us then begin by restating the three questions:

1. What are the benefits to archaeologists of using semantic technologies to express their data?
2. What are the social and technical requirements for expressing archaeological data with semantic technologies?
3. To what extent is contemporary archaeological practice able and willing to meet the social and technical requirements of expressing data with semantic technologies?

In Chapter 1 we asked whether there was a *prima facie* case that semantic technologies might prove beneficial to archaeologists. By comparing the advantages generally accredited to a Semantic Web approach with the aims and objectives of archaeologists we surmised that such a case could be made and that further investigation was justified. We also noted that the case for semantic technologies was not clear cut, with apparently contrasting views on their levels of utility and historically low levels of adoption in other domains. In Chapter 2 we reviewed both formally and informally published literature on the theory and application of semantic technologies. This review appeared to suggest that, following the seminal *Scientific American* article published in 2001, two

approaches to the use of semantic technologies appear to have arisen which we dubbed ‘Mixed-Source Knowledge Representation’ (MSKR) and ‘Linked Open Data’. It was not clear to what degree either of these perspectives have directly influenced Cultural Heritage practitioners, including archaeologists. We also observed a distinct peak in Semantic Web activity in the years 2006–8, followed by relatively rapid decline, across a range of phenomena including search terms, blog posts and conference proceedings. Whether this apparent reduction in interest is permanent or merely a consolidatory phase of the Gartner Hype Cycle is equally unclear. Surprisingly, archaeological activity in this area does not appear to have been affected as far as can be established from a count of presentations at *Computer Applications in Archaeology* conferences. In Chapter 3 we analysed the results of a survey of forty Cultural Heritage projects that have utilised semantic technologies. It became evident that there was indeed a separation between MSKR and Linked Open Data approaches and that these seemed to be further associated with whether a project was fixed-term or open-ended. Projects seemed equally divided between these two paradigms both in Archaeology and in Cultural Heritage as a whole. It was therefore concluded that the answer to the first question — and thus the second and third question also — could only be answered relative to each of these visions.

Chapter 4 commenced the second part of the thesis by introducing a central case study, Roman Port Networks. The project is fixed-term, distributed across multiple voluntary partners, medium scale, uses legacy digital data in heterogeneous formats and requires restrictions on data access. These made MSKR clearly the more appropriate methodology to follow. A relatively lightweight infrastructural framework was then described which sought to support the project’s goals of merging multiple excavation datasets in order to visualize and analyse them in combination. Chapter 5 described TRANSLATION, a suite of tools designed to facilitate the conversion of traditional data formats to RDF. While the tools were successful on a purely technical level, the evaluation showed that a number of socio-technical problems stood in the way of their successful adoption. In particular, the difficulties associated with using and adapting both the tool and infrastructural components outweighed any direct benefits to the individual project partners undertaking the work. Chapter 6 described a very different ‘cookbook’-based approach, entitled *Introducing Semantics*. This attempted to achieve the same goals by means of a step-by-step guide that used familiar, or off-the-shelf and easily modifiable, technologies. The approach again proved successful technically but not socially. Very little voluntary adoption was observed and again this appeared to be due to an imbalance between the perceived return and the requisite time investment, despite an explicit attempt to provide direct benefits early and regularly. So what can we learn from this experience and, furthermore, what can it tell us about the benefits — and limitations — of semantic technologies more generally?

7.2 Discussion

At the outset we might have suspected that the MSKR vision of semantic technologies would hold greatest promise for archaeologists. The archaeological sector has historically depended on scholars combining disparate but limited resources together in order to produce synthetic works that in turn become the raw material for others. MSKR seems to promise a richer, more sophisticated way of undertaking this labour without disrupting the established order of hard-won academic authority and authorship. The end result should be a cleaner, more tractable set of resources with which scholars can answer meaningful questions. Yet our experience has shown that a number of substantial obstacles still stand in the way to implementing such a vision. Sometimes these challenges are specific to the particular needs of Roman Port Networks but, as will be argued here, the wider picture looks equally problematic.

Scale and Complexity

The ability to deal with large, complex combinations of data is the very *raison d'être* of MSKR. Small or simple combinations of datasets rarely appear to justify the level of investment in time and resources required to apply semantic technologies, as they are readily susceptible to analysis and visualization by more traditional means. This is a trend clearly demonstrated in Figures 3.8–3.10. Unfortunately, it means that MSKR is denied many of the testbed projects in which both technical and domain experts can experiment, learn best practice and refine their tools. It is certainly the case in Roman Port Networks that only the author had any significant prior knowledge of semantic technologies before commencement of the project. Had a significant number of partners had previous experiences with its use, the cognitive barriers would unquestionably be lower. While obviously there are many archaeological problems that can only be solved by processing large quantities of heterogeneous data — and specialist techniques may be required to address them — this restriction on relevant projects necessarily limits the size of the community of expertise and support.

Motivation and Engagement

For MSKR to work, a significant proportion of potential data providers must be convinced that what will inevitably be a challenging technology to use is worth their while. Semantic technologists must not lose sight of the fact that domain experts are rightly more concerned with solving a specific domain problem than experimenting with new technologies or creating abstract over-arching ontologies. In particular, the difficulties associated with explaining the end-to-end process and its eventual outcomes can make even minor problems deeply demoralising to the uninitiated, regardless of whether they are issues specific to semantic technologies or not. The lack of mature and user-friendly tools and training resources is another commonly cited source of exasperation. Without a well thought through strategy

for maintaining high levels of motivation across the project team, semantic technologies are unlikely prove viable.

Alternatives

Technical alternatives which cost less, are more intuitive, or simply more familiar to the archaeologist, have already been mentioned as an important consideration in the cost-benefit equation. The obvious contenders are relational databases and spreadsheets (upon which specialist technologies, such as GIS, are also based). It is important to ask whether the same amount of resource investment required to convert data into a semantic format would deliver comparable results using a more mainstream solution. No serious research has been done on this question within Archaeology and costs are inevitably a product of both the technology and operator. In the author's experience, however, it seems unlikely that semantic technologies can answer questions at a medium scale that could not be asked of relational data, or that they are significantly more efficient. Considered in the light of the cost-benefit graph in Figure 5.10, it appears that we are at the point where both formality front-loading and formality deferring approaches roughly coincide for projects combining a moderate number of datasets.

This of course returns us to the nature of the insight that graph was intended to demonstrate: data management and analysis is a dynamic process in which we must balance the sophistication of a technology with the scale of the task. If a project were to become significantly larger or more complex — through the continual addition of new data, say — then the investment in semantic technology may well begin to pay off. But this is simply to beg the question. To what extent are the typically fixed-term projects that apply MSKR well-placed to take on such large, and presumably long-term, objectives?

Inferencing

In addition to handling scale and complexity, MSKR ostensibly offers the additional benefit of inferencing, i.e. the ability to explicitly deduce information that is only implicit in the source data. Despite this claim, there are few if any clear cut examples in which machine reasoning has delivered meaningful and unexpected results from combined archaeological datasets. This is not to say that inferencing has not been used or proven useful. Projects like STAR have demonstrated that inferencing is *possible* but have yet to publish archaeological results based specifically on this process. Attempts to categorise archaeological finds based on combinations of attributes have also been made (Karmacharya et al., 2010) but as the process uses a Rule Language to define each category solely in terms of these attributes it is hard to argue that it leads to the creation of new archaeological information.

There may be a simple, and practical, explanation for this lack of results. There has been a historic tendency to use complex ontologies — such as the CIDOC

CRM — in order to capture as much of the complexity of a domain as possible. This has the effect of lengthening the chain of reasoning, i.e. the list of statements that are required to be true, for an inference to hold. Unfortunately, if any one link in the chain should be false, the entire inference is invalid (regardless of whether the conclusion itself is true). If a proportion of our data is wrong, valid inferences become less and less probable as our reasoning chain extends and the odds of encountering a false assertion rise. Long chains of reasoning therefore have a greatly reduced level of validity in comparison to shorter ones and so complex ontologies, although theoretically able to answer more sophisticated questions, are inherently less robust for reasoning over imperfect datasets. The upshot is something like the **Uncertainty Principle**: it is possible to optimise for precision (long chains) or accuracy (valid inferences) but not both. In an archaeological context, where a considerable level of error is to be expected, the ramifications for inferencing are serious.

7.3 Conclusions

This list of challenges is potentially disheartening. Must we conclude that semantic technologies are, despite their early promise, simply unsuited to archaeological practice? It is almost certainly too early to make definitive proclamations, but there are at least some reasons to believe that a fully negative appraisal is unwarranted. In order to envisage a more optimistic future for its adoption in Archaeology, however, we need to return to its roots and in particular the twin themes that Berners-Lee raised in his early writings: openness and decentralization.

7.3.1 Openness

The issue of openness plagues many discussions of Web content in Cultural Heritage, laden as it is with the emotive baggage of academic ethics. Few of us oppose it in principle, most of us do so in practice. The reasons, as we know, vary from the legitimate, to the illegitimate, to a simple acceptance of the *status quo*. Yet openness is fundamental to this discussion because it is the wellspring of the Web. The Web's primary function is to facilitate access to information and thus anything that which restricts that capacity undermines the central source of its power. That is certainly not to say that data privacy has no place, but we must remain aware that the Web as a technological and social paradigm is only powerful *because* so much content is made freely and easily available. **Amazon**, **Facebook**, **Data.gov**, **Wikipedia** and other organisations have not revolutionised their respective sectors through the innovacy of their software applications — many of the principles for which predate the Web — but because they are

ubiquitous and cost nothing to the user (at least financially). The same services in a closed environment would be so limited as to seem almost absurd.

Yet neither Archaeology, GLAMs or the Humanities in general seem to have experienced an ‘online revolution’ similar to those experienced in the political, commercial, media and social spheres. Could one take place? We must bear in mind that none of these online revolutions were easy to predict and that other domains may not prove a reliable guide, thanks to their diverse array of concerns around this topic. There are certainly those pressing for greater accessibility of cultural resources online, but they are a distinct minority and there is currently no Internet juggernaut on the horizon. By and large, the humanities, heritage and archaeological communities worldwide still seem to see data publication overwhelmingly as a zero-sum game in which making it available online is more likely to lead to work being scooped or plagiarised than to collaboration or citation. Thus, for the foreseeable future at least, most archaeologists will remain reluctant to provide open access to it.

This ‘presumption of closedness’ has particularly important consequences for the adoption of semantic technologies, for many of the obstacles we have just been discussing would be at least partially ameliorated in an more open data environment. First, the chances of producing large combinations of source data would be much greater, thereby helping to justify the amount of work required to draw new results from them. Second, larger user communities would mean that tools and services could benefit from longer feedback and evaluation cycles, helping to improve productivity and efficiency in the medium-to-long term. The effect on motivation is much harder gauge. It is true that many archaeologists may remain sceptical of contributing data for which traditional modes of attribution are poorly suited. On the other hand, the rich source of information with which to contextualize their own data could prove sufficient motivation to participate in its own right. In short, a broadly open data policy across the archaeological community would clearly improve the odds of successfully utilizing semantic technologies, but once again we are begging the question. Why should archaeologists care?

One trend that may start forcing them to take openness more seriously in the UK are changes in legislation and funding criteria. Driven both by ideological and financial considerations, impact and public engagement have become increasingly important in recent years. Citizen scholarship and co-creation are themes well in tune with the UK Government’s ‘Big Society’ agenda. Of course, this very fact may not always endear them to academics facing other government pressures, but there is no doubting its determination to extract as much public perception of value as possible when considering expenditure from the state purse. If levels of Humanities research funding decrease more significantly, UK Archaeologists will almost certainly need to adapt to this agenda in order to remain competitive. Clear evidence for this shift can be seen in the Joint Information Systems

Committee's (JISC) **Discovery**¹ initiative in which “*there will... be a shift in ethos, the most crucial being the embracement of ‘Openness’*” (Resource Discovery Taskforce, 2011), and the **House of Commons Science and Technology Committee**'s recent report on *Peer Review in Scientific Publications* which concluded that:

Access to data is fundamental if researchers are to reproduce, verify and build on results that are reported in the literature. We welcome the Government's recognition of the importance of openness and transparency. The presumption must be that, unless there is a strong reason otherwise, data should be fully disclosed and made publicly available. In line with this principle, where possible, data associated with all publicly funded research should be made widely and freely available. Funders of research must coordinate with publishers to ensure that researchers disclose their data in a timely manner. The work of researchers who expend time and effort adding value to their data, to make it usable by others, should be acknowledged as a valuable part of their role. Research funders and publishers should explore how researchers could be encouraged to add this value. (House of Commons Science and Technology Committee, 2011, para. 203)

Yet even on the assumption that such changes in research practice do take place, they may still stop short of the level of openness in which semantic technologies really come into their own. For a start, UK archaeologists will still need to work in an international arena in which openness is the exception, rather than the rule. Secondly, the emphasis has so far remained on post-publication access. While preferable to the current state of affairs it could potentially create a ‘twin cycle’ research environment in which smaller, closed projects wait until publication to release information, and are followed by larger digital synthesis projects which work solely with this published data.

One additional social trend is starting to operate to the advantage of openness: the ubiquitous use of mainstream search engines for academic resource discovery. Enhanced by services such as **Google Scholar**² and **Academia.edu**,³ earlier concerns about the nature of this process, while still prescient, seem to be increasingly ignored or simply accepted by academics. As just one example, Humanists across Europe interviewed by the **Preparing DARIAH**⁴ project concurred that Google was the primary search tool within their academic practice (Benardou, 2011, min. 39:55–42:00). This may or may not benefit academic discourse but it seems inevitable that scholars are now more likely to discover content that is openly available on the Web. This may in turn bias the the academic merit system in favour of those who self-archive or deposit their data in

¹<http://discovery.ac.uk/>

²<http://scholar.google.com/>

³<http://www.academia.edu/>

⁴<http://www.dariah.eu/>

open institutional repository systems such as ePrints.⁵ Whether it forces a tipping point across archaeological culture more generally remains an open question.

7.3.2 Decentralization

Decentralization is crucial to the growth of the Web because it creates both the inter-connections and economies of scale that allow it to expand. It requires an open network of loosely connected services, tools and datasets. While Humanities projects have historically had to provide or create most of their data, infrastructure, tools, methods and questions themselves, the Digital Humanities environment is now starting to mature. A growing number of freely available specialised resource services — such as the ADS, Arachne and Pleiades — mean that an increasing proportion of a project's time and money can be targeted at specific research questions. By making such 'low-hanging fruit' available, these tools and services may in turn start to influence which questions are asked, once again driving the research agenda away from more closely-held proprietary datasets. Central to our discussion of semantic technologies is the use of common HTTP URI frameworks as a means of classifying and identifying concepts. As we have seen, services such as GeoNames and Freebase allow us to collectively build digital thesauri and gazetteers that not only provide canonical identifiers for topics of mutual interest, but also supplement them with additional data. Unfortunately it is precisely these fundamental building blocks of the Web which are so sorely lacking in Archaeology.

The problem seems to have arisen due to a conflation in understanding between what constitutes an open and closed system on the Web. Just as with natural language terms, a URI cannot be held privately or its use proscribed. While it *is* possible to restrict the ability to dereference it, doing so entirely eliminates the value of the URI for, unlike natural language, a URI is opaque. A phrase like 'ring fort' can, more or less, be interpreted within its linguistic context. In contrast, a URI should be interpreted only through the information retrieved by dereferencing it. Obscuring this information not only greatly discourages its use, but also its interpretation by third parties and thus its utility on the Web of Data. In contrast, making it available creates a conceptual 'spine' around which relevant datasets can cluster, as we have seen from the Linked Open Data cloud. Regrettably, those who have traditionally maintained Humanities thesauri have long perceived their immediate interests to lie in the sale or licensing of thesauri, rather than establishing themselves as a hub in the Web of Data. Sometimes, the introduction of a new URI-based service (such as GeoNames as a substitute for the Getty Thesaurus of Geographic Names) all but renders the original service obsolete. Sadly, many widely-used archaeological thesauri are still restricted to sale-only printed volumes or cannot be referred to by URI. Until the funding and resolve are found to dismantle what has effectively become a cultural monopoly, any attempt to introduce semantic technologies

⁵<http://www.eprints.org/>

into Archaeology will remain hamstrung. It should be emphasised however that this issue remains separate from the question of Open Data discussed earlier. The use of public URIs in no way requires that data making use of them is publicly accessible as well.

So while it remains unclear whether semantic technologies will offer serious benefits to archaeological practice in the future, it seems that the adoption of URI-based thesauri would be a pre-requisite. Once more we find ourselves in a chicken-and-egg situation however. If there is no guarantee that semantic technologies *will* prove viable, why should funding agencies and memory institutions invest in URI thesauri? One argument is that those advocating a Linked Open Data approach may create a gradual transition that only later escalates into a more transformative revolution.

7.3.3 Linked Open Data

What are we able to say about the Linked Open Data approach in Archaeology? We must return to this topic with caution as we have not had the benefit of concrete experimentation. Nevertheless, it seems fair to observe that it does not face quite the same battery of obstacles as MSKR. To begin with, Linked Open Data is in its essence much more concerned with openness and discoverability than collaboration and inference. It avoids many of the complex ontological challenges faced by those who require inferencing. Value from this perspective comes not from exclusive possession or logical deduction, but by being situated centrally in a network of other resources. As such, quality matters but utility matters more. Yet as we have seen, it may currently be too difficult for archaeological practitioners to create enough RDF *data* to justify the name. Can we get beyond this impasse?

It may be that we can begin the process by moving our focus from data to **annotations**. Rather than creating a Web of Data adjacent to the Web of Documents, we can kickstart the process by enhancing the Web of Documents with RDF snippets that provide data in computationally digestible form. This in turn can create relationships between documents unforeseen by their original authors. Such annotations can be equally applied to pages, tables and images, and are applicable at the level of content rather than merely a document summary. Rather than thinking of this as an impoverished form of Knowledge Representation, it instead should be considered as super-charged metadata. It can be held within the actual document itself, such as with embedded RDFa, but it might also be by means of adjacent files, content negotiation or a SPARQL endpoint. This allows for a much greater range of information to be held about the document — not only the author, subject, date and so on, but the people, places and periods or categories used within it. While it may not be possible to deduce facts from this data, the ability to relate it to other content, whether directly or statistically, is vastly enhanced. The second improvement over traditional methods is of course that URIs make such annotations

language and term-independent. This cannot entirely solve the Vocabulary Problem of course, but the arbitrary quantity of them means that many more keywords can be created for a document, while vocabularies such as SKOS allow for limited reasoning in order to identify related URIs. Such semantic annotations promise something well beyond what we are capable of today — vastly improved discovery and contextulization. The promise of these benefits may well be sufficient to justify the investment in the URI thesauri necessary for a mature Semantic Web.

Semantic annotations are by no means a new idea and automated services for extracting them, such as **OpenCalais**,⁶ have been around for some time. While the algorithms employed by these generic solutions are not usually capable of dealing with the specialist nature of archaeological documents, targetted applications have been more successful. Archaeological grey literature was classified under ‘What’, ‘Where’, ‘When’ categories in 2009 by the **ArchaeoTools**⁷ project (Jeffrey et al., 2009), and **Google Ancient Places**⁸ has extracted place references in classical texts from the Google Books corpus using a combination of the **Edinburgh Geoparser**,⁹ **OpenAnnotation**¹⁰ and the Pleiades gazetteer (Isaksen et al., 2011). The power of this approach does not simply lie in the ability to visualize data in new ways — although that is part of its appeal — but that it can be automatically associated with relevant resources that transcend document formats (Tudhope et al., 2011b) and even domain boundaries (Simon, 2011). Automated processes can be only part of the solution, and semi-automated and manual techniques will also need to be established, depending on the nature of the source data. Public crowd-sourcing (Brusuelas, 2011) is an important development that has yet to be put to use for semantic technologies in this domain, and standardised documentation systems such as **ExeGesIS**,¹¹ **ARK**¹² and **Intrasis**¹³ may also be able to facilitate the creation of ubiquitous lightweight annotations of Web content.

7.4 Contributions and Future Work

In this final section let us look at some of the concrete ways in which the work we have discussed might be carried forward.

Technical and Social Considerations

Despite some of the socio-technical challenges raised by this research, it is important to remember that we have demonstrated that archaeological data *can* be

⁶<http://www.opencalais.com/>

⁷<http://wit.shef.ac.uk/archaeotools/index.html>

⁸<http://googleancientplaces.wordpress.com/>

⁹http://www.ltg.ed.ac.uk/clusters/Edinburgh_Geoparser

¹⁰<http://www.openannotation.org/>

¹¹<http://www.esdm.co.uk/hbsmr.asp>

¹²<http://ark.lparchaeology.com/>

¹³http://www.intrasis.com/engelska/index_eng.htm

integrated and queried by means of semantic technologies. This was much less apparent at the commencement of the work in 2007 and, although important work has also been done by others to establish this, Roman Port Networks has certainly been at the pioneering edge of that research. That it was achieved on a moderate budget — comparable with that which would need to have been invested in a relational database or GIS-driven approach for the same scale of data — is an equally important discovery. Perhaps the most important contribution however, is to have more clearly problematized the process at two levels.

The first is in identifying and describing the various subtasks which need to be undertaken in order to convert conventional archaeological data into a semantic form. This documentation was drawn partially from the development of the TRANSLATION tools, but made explicit in *Introducing Semantics*. Laying out each requirement in this way makes it not only easier for those wishing to undertake the process themselves but potentially permits the development of more sophisticated — perhaps Web-based — tools that can facilitate and expedite the process.

At a secondary level however, it became clear over the course of the project that by far the most challenging obstacles are social ones to do with skills-balancing and motivation. It is now established fact that a computer scientist can use semantic technologies to merge together diverse archaeological datasets. We are still a very long way from establishing whether a sufficient number of domain experts, who must create and make use of them, can or would wish to do so. If there is one particular area that the author feels future investigation should focus on then it is this. Even using conventional technology, it is likely that large integrative projects of the sort that would most benefit from semantic technologies must be multidisciplinary, involving both domain experts and computer scientists. Establishing research processes that balance the varying needs, motivations and costs between these two groups will be essential.

Even a mid-term goal of semantically enriching the Web of Documents will require a vast amount of human effort. The process of creating a book, article or database is quite time-intensive enough, and it is wishful thinking to imagine that many authors will be able, let alone willing, to semantically mark up materials. We will inevitably have to rely heavily on automated processes run by repositories which in turn raises additional issues of annotation authorship and veracity. Any automated process will unquestionably create erroneous annotations and authors may be uncomfortable with their publications being described in ways they have no control over (despite this being an inevitable process of the library system). Two partial solutions suggest themselves, although neither resolves these issues entirely.

The first is that automatically generated annotations are most useful when used in their wider context. An annotation specifying that a document refers to an

individual place may or may not be correct, for example, but the ability to visualize all of those places at once can give a strong indication as to which annotations are outliers. The second solution returns us to the Semantic Catch-22 we discussed in Section 3.3.2. There is a semantic difference between truth claims made by humans and those which are merely the product of an algorithm. The former is an attempt to understand and interpret the world, the latter is essentially a digital conversion process. We need to establish methods of a) informing users whether they are looking at automated or humanly generated annotations, and b) convert automated annotations into human ones. The magic of (human) semantics is that this need not always require more than ticking a checkbox. It is no less the important for it, however. Such computational tools need to be developed by the repository community that hosts archaeological materials, whether written or tabular, as fixed-term projects will likely have neither the funding nor motivation to create them. Their contribution must be the cumulative process of interpreting, and where necessary correcting, the content and its annotations.

Infrastructure

A more complex output is that of the vocabulary infrastructure created for the project, in particular ArchVocab and the amphora typologies in Freebase. Both of these have been created with only medium-term sustainability in mind, which is adequate for the duration of the project but can no make guarantees about its persistence thereafter. This cannot be ignored, as an important consequence of creating the resources was the interest shown by others working in related fields. The ceramics typology is the least problematic. There are no (current) reasons to suppose that Google will discontinue the Freebase service and its open architecture means that interest and maintenance can go hand in hand. There is clearly great potential for expanding the current set of amphorae to include missing types and even to encompass other kinds of artefact, such as finewares, coursewares, stone or metals. Of particular value to such an approach is the fact that it would help remove the research focus from the better documented types (such as Dressel 20, African Red Slip, or Porphyry) to the full gamut of materials transported in Antiquity. Visualizing the spatio-temporal distribution of this range of products could clearly grant great insight to researchers working in this field as well as identify new questions. While a first phase of such visualisation is more likely to highlight issues with the data than answer questions, it would still be of immense value to the research process.

The ArchVocab excavation ontology is currently hosted on a Virtual Machine server at the University of Southampton. Although permission has been granted to extend the service beyond the lifetime of the PhD, it is clearly not a long-term solution. Should it receive significant interest from researchers beyond Roman Port Networks it will need to be moved to a more persistent platform, such as

the **Open Knowledge Foundation**.¹⁴ Such interest notwithstanding, there is a desperate need for a public, stable, URI-based domain ontology for archaeological excavation. English Heritage, in particular, are in an excellent position to build upon the achievements of the STAR project by publishing persistent URIs for the CIDOC CRM-EH. This single action would not only open up their extensive achievements to the possibility of integration with external datasets, but could provide the basis for similar initiatives in other countries.

Beyond these URI vocabularies there is also need for more work in cross-domain concepts such places, events and things. Fortunately URI place gazetteers are maturing rapidly, with GeoNames including historic place names (Wick, 2011) and benefitting from recent alignment work with Pleiades (Isaksen, 2011). This in turn may help the development of temporal gazetteers — URI sets for historic events — although work in this field, such as **CommonEras**¹⁵ or the STAR.TIMELINE application (Binding, 2010), are at an early stage of development and a great deal of work in this space remains to be done. For the classical era, URI-based prosopographies could also open up extremely important connections between finds¹⁶ and it is hoped that work currently being carried out by the **Lexicon of Greek Personal Names (LGPN)**¹⁷ will break new ground in this area (LGPN, 2011).

Other Issues

One area that was considerably under-explored by this thesis — focussing as it has on publication rather than consumption — are the user interfaces for archaeologists to make the most of semantically formatted data. This is a formidable subject and well worthy of (at least) an equivalent body of work. The central challenge will be to identify and/or develop tools which can account for the flexibility of RDF while directly facilitating the specific needs of the archaeologist. A number of off-the-shelf graph visualization tools exist, such as Protégé, but these are rarely if ever satisfactory for archaeological use, displaying information at a level of granularity that is of little help. This problem only increases with the sophistication of the ontology. Bespoke tools, such as those developed by the author for the purposes of this research, can help to answer specific questions, but the development of them is comparatively research intensive, compared to the level of output, for small or medium sized datasets and they are highly ontology dependent. The author is already involved in work to visualize relationships between archaeological data in separate repositories (Simon, 2011) but the work of Tudhope et al. (2011a) and Huvila (2008) are of particular interest for understanding and analysing intra-site data.

¹⁴<http://okfn.org/>

¹⁵<http://commoneras.ecs.soton.ac.uk/about.html>

¹⁶For the results of a network analysis approach without URIs, see Graham and Ruffini (2007).

¹⁷<http://www.lgpn.ox.ac.uk/>

Another issue that needs to be addressed by the community of developers, archaeologists and heritage specialists working with semantic technologies is a level of vagueness and ambiguity that tends to plague communication between projects. In some ways this is through no fault of their own, as the problem is rooted in the competing visions of the Semantic Web that have been around since at least 2001. It is hoped, however, that this thesis has gone some way to separating out these perspectives — hopefully not at the expense of introducing more unnecessary jargon — and clearer statements of what projects are trying to achieve (and not trying to achieve) may facilitate more fruitful collaboration. Similarly, there is an important need to publish in forums with much faster time-to-press. While conferences such as *CAA* provide excellent venues for discussion, it has become ever more apparent that it is far too slow as a publication mechanism in such a fast-moving field. It is evident that papers discussing semantic technologies from *CAA* 2007 — the most recent year for which a printed volume is currently available — are now woefully out of date. Even a more rapid turnaround of, say, two years, is insufficient to keep digital archaeologists up-to-date. Online journals and blog-posts will need to be the preferred forum of publication if Archaeology is to stay apace with wider digital developments.

The observant reader may have noticed a curious irony running throughout this text. Despite its title, the phrase ‘Semantic Web’ has been rarely used, with a preference for the somewhat more pedestrian ‘semantic technologies’. Indeed, these days the Semantic Web is frequently accompanied by scare-quotes, “so-called”s, or substituted altogether by euphemisms like ‘Linked Data’. One does not need to be an ontologist to appreciate the powerful nuance behind these terminological shifts. The decision to use these phrases is deliberate, for in truth the Web often plays a limited role in the field this thesis has discussed. Yet if its conclusions are correct then they mark a profound issue for everybody working in it, for the power of semantic technologies ultimately derives from the Web and the Web in turn is driven by the principles of openness and decentralization. That is not to say that semantic technologies cannot be used in a closed or centralized environment, but in doing so we restrict ourselves — by definition — to a Knowledge Representation paradigm of yesteryear which “*did not... shake the world to the extent that some of its proponents hoped*” (Berners-Lee, 1998c). Until we try, it remains a moot point as to whether Berners-Lee’s vision of an open and decentralized Knowledge Representation is possible. The question left for archaeologists to consider is:

Could an open and decentralized Archaeology be possible?

Appendix A

Semantic Web Conference Papers

Museums and the Web. Proceedings 2002–2010

2002

1. Adding Value to Large Multimedia Collections Through Annotation Technologies and Tools: Serving Communities of Interest
2. Cyberspace in Our Space
3. Designing With Web Standards
4. Today's Authoring Tools for Tomorrow's Semantic Web

2003

2004

1. Designing With Web Standards
2. Museums and the Web: Maturation, Consolidation and Evaluation

2005

1. New Ways to Search, Navigate and Use Multimedia Museum Collections Over the Web

2006

1. The Inside Out Web Museum

2007

1. Personalized Museum Experience: The Rijksmuseum Use Case
2. Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques

3. When is a Terracotta Hut Urn Like a Sailor's Deck-Log?: Meaning Instantiated Across Virtual Boundaries

2008

1. Be Your Own Curator with the CHIP Tour Wizard
2. Contexta/SR: A Multi-Institutional Semantic Integration Platform
3. The Delphi toolkit: Enabling Semantic Search for Museum Collections
4. Semantic Dissonance: Do We Need (And Do We Understand) The Semantic Web?
5. Technologies, Like Museums, are Social

2009

1. CultureSampo - Finnish Culture on the Semantic Web 2.0: Thematic Perspectives for the End-user
2. hoard.it: Aggregating, Displaying and Mining Object-Data Without Consent

2010

1. The CollectionWeb Digital Ecosystems: A Semantic Web and Web 2.0 Framework for Generating Museum Web sites
2. Recollection: Building Communities for Distributed Curation and Data Sharing

Computer Applications and Quantitative Methods in Archaeology (CAA). Papers and Proceedings 2001–2011

2001 (Proceedings)**2002 (Proceedings)**

1. CRM and Archaeological Research Using Remote Sensing and GIS: Zhouyuan (China) and Lasithi (Greece)

2003 (Proceedings)

1. CIPHER (Communities of Interest to Promote Heritage of European Regions) — Internet Cultural Portals and the Development of an Irish Cultural Heritage Forum
2. Documenting Two Histories at Once: Digging into Archaeology
3. Ontologies as a Reference Framework for the Management of Knowledge in the Archaeological Domain

4. Peer-to-Peer Ways to Cultural Heritage

2004 (Proceedings)

1. Digital Paths to Medieval Naantali From Mobile Information Technology to Mobile Archaeological Information
2. Integration of Complementary Archaeological Sources
3. From XML-tagged Acquisition Catalogues to an Event-based Relational Database
4. To OO or Not to OO? Revelations from Ontological Modelling of an Archaeological Information System
5. Which Period is it? A Methodology to Create Thesauri of Historical Periods

2005 (Papers)

1. BRICKS: Moving Towards Cultural Heritage Integration
2. BRICKS — Building Resources for Integrated Cultural Knowledge Services
3. Encoding Model for the Integration of Greek and Latin Sources and Archaeological Evidences
4. Historical Memory Preservation on the Semantic Web: The Case of the Historical Archive of the Aegean — Ergani

2006 (Papers)

1. Data Integration with ArchaeoML and Tagging
2. The Dutch Knowledge Infrastructure for the Cultural Heritage
3. ECAI — Semantic Topographies and Spatial Markup: The Digital Roman Forum Project
4. ETANA-DL: Leveraging Digital Library Technologies to Support Archaeology
5. Integrating Data from Disparate Projects Using Taxonomies and Thesauri
6. The National Reference Collection Online
7. Semantic P2P Network for Virtual Reality Archeological Resources
8. Towards a Core Ontology for Sharing Knowledge about Cultural Heritage
9. Triple O (Ontology of Ontology) — A Digital Data Survey for Japanese Archaeology

2007 (Papers)

1. CHIMERA: A Service Oriented Computing Model for Archaeological Research
2. The CIDOC-CRM Encoding of Fontes ad Topographiam Veteris Urbis Romae Pertinentes

3. CIDOC Current Research and Standards
4. Filtering RDF Metadata in Java
5. How Open We Really Are? — Archaeological Primary Data Sharing in Cyberspace
6. The National Reference Collection: Bringing Typologies Alive
7. An Ontology Server for Archaeological Information Retrieval and Processing (OSA)
8. An Open Approach to Contextualising Heterogeneous Cultural Heritage Datasets
9. OSCAR — A Web-based Multimedia Communication System for Interdisciplinary Settlement Research
10. Raising PaMeLa
11. Shared Typology and Iconographical Representations with Ontological Models
12. Weaving the Archeological Web. Enabling Semantic Integration and Interoperability of Information Systems Holding Data from Ancient History with Metadata Harvesting and Content Reference Models

2008 (Papers)

1. A CIDOC-CRM Model for Egyptian Middle Kingdom pottery: Archaeological Theory and Ontological Practice
2. Concept for an Ontology Based WebGIS Information System for HiMAT
3. Cross-border Data Sharing: A Case Study in Interoperability and Web Services
4. Data Sharing Between Different Databases by Using an ISO 21127 Related Data Format
5. A Semantic Based Approach to GIS: the PO-BASyN Project
6. Standardization or Mapping? Some Considerations on CIDOC-CRM Extensions and Alignment
7. A STAR is Born: Some Emerging Semantic Technologies for Archaeological Resources
8. The Use of Network Analysis and the Semantic Web in Archaeology: Current Practice and Future Trends
9. When Ontology and Reality Collide: The Archaeotools Project, Facetted Classification and Natural Language Processing in an Archaeological Context

2009 (Papers)

1. ArchaeoKM: Toward a Better Archaeological Spatial Data Sets Management

2. An Archaeologist's Reflections on Semantics and the Web
3. ArcheoInf — Allocation of Archaeological Primary Data
4. Automatic Extraction of Archaeological Events from Text
5. Complex Networks in Archaeology
6. Deducing Event Chronology in an Archaeological Documentation System
7. Extending and Enriching the CIDOC-CRM Ontology for Task-Ontological Domain Model
8. Following a STAR? Shedding More Light on Semantic Technologies for Archaeological Resources
9. iDAI.field and More — Documenting Field Projects at the German Archaeological Institute (DAI)
10. Implementing RDFa in the Publication of Ceramic Data from Troy (Turkey)
11. Implementing Semantic Web Software in the Field of Cultural Heritage Using the CIDOC CRM — Prospects and Challenge
12. Linking Archaeological Data
13. Natural Language Processing within the Archaeotools Project
14. A Prototype for Managing Archaeological Excavation Data in a Digital Library for the American School for Classical Studies in Athens
15. Web GIS-Supported Implementation of the CIDOC CRM

2010 (Papers)

1. ArchaeoKM: Managing Archaeological Data Through Archaeological Knowledge
2. Atom Feeds and Incremental Semantic Annotation of Archaeological Collections
3. A Framework for Transforming Archaeological Databases to Ontological Datasets
4. Interoperate with Whom? Archaeology, Formality and the Semantic Web
5. Making Vector-Based Images Ready for the Semantic Web
6. Methodology for CIDOC CRM Based Data Integration With Spatial Data
7. Ontologies and Semantic Tools for the Management of Full-text Archaeological Documentation. Assessments from the Hala Sultan Tekke Case-study
8. Recent Developments in the ArcheoInf Project — Towards an Ontology of Archaeology
9. Semantic Technologies for Archaeology Resources: Results from the STAR Project
10. Sustaining Database Semantics

11. A Very Short Introduction to the Semantic Web
12. WikiBridge: a Semantic Wiki for Archaeological Applications

2011 (Papers)

1. ACAML — A Markup Language for Ancient Chinese Architecture
2. Being Formal and Flexible: Semantic Wiki as an Archaeological e-Science Infrastructure
3. Exploring Semantic Web Data from Visual Sources in Archaeology
4. Facilitating Database Content Re-use for Semantic Feeds and Mobile Applications
5. Introducing Semantics — Pathways to Data Integration in Archaeology
6. The JISC Data Management Blueprint Project: Archaeological Implications
7. Metamorphoses: Co-referencing Classical Place Data for the CLAROS Project
8. Relational Database Implementation of CIDOC CRM to Model Interdisciplinary Research
9. SEA: A Framework for Interactive Querying, Visualization and Statistical Analysis of Linked Archaeological Datasets
10. Semantic Technologies Enhancing Links and Linked Data for Archaeological Resources
11. Towards a System for Semantic Image-Based 3D Documentation in Archaeological Trenches
12. Was it Worth It? Experiences with a CIDOC CRM-based Database

Appendix B

STCH Survey: Participant Information Sheet

Study Title: Evaluating semantic technologies for data publication in Cultural Heritage

Researcher: Leif Isaksen

Ethics number: E/10/05/002

Please read this information carefully before deciding to take part in this research. If you are happy to participate you will be asked to sign a consent form.

What is the research about?

My name is Leif Isaksen and I am conducting research for a PhD thesis on the use of Semantic technologies to publish archaeological data. This survey is intended to establish how and why semantic technologies have been used in previous cultural heritage projects and compare the perceptions of those who have used them. The results will also be evaluated in comparison with Semantic Web ‘success stories’ outside the Cultural Heritage sector. It is intended that the results will provide a useful baseline from which myself and others can develop semantic systems that specifically address the needs of the sector. This research is being paid for by the UK Engineering and Physical Sciences Research Council (EPSRC) and is sponsored by the University of Southampton.

Why have I been chosen?

As a member of a previous or ongoing project that has utilised semantic technologies to publish data, you are able to provide important perspectives on the merits of semantic technologies. I would like you, and potentially a colleague on the same project but with a different specialism, to describe the approach taken and assess its value to the project.

What will happen to me if I take part?

If you consent to take part you will be asked to provide your name and that of a project you have participated in. This information will be held securely on a Southampton university server and separately from the survey data under the terms of the Data Protection Act. You will then be given an identifying code and directed to the online questionnaire at surveymonkey.com (details of the questions you will be asked can be found on a PDF file accompanying this information sheet).

Subject to your additional consent, I may contact you by email with some follow-up questions or ask to identify the project as an illustrative case study for my PhD thesis.

All personal details and project details (i.e. those provided on the consent form) will be deleted on completion of the PhD)

Are there any benefits in my taking part?

There are no personal advantages (inducements) to be gained from taking part, but the overall results of the survey will be made available to the wider cultural heritage community and this is likely to be of help to those considering the adoption of such technologies in future.

Are there any risks involved?

There are no personal risks involved but you will be asked to candidly assess the costs and benefits of the approach taken by your project. Differences of opinion between people reporting on the same project are possible and while project and personal details will be anonymised it is impossible to rule out any possibility of identification. If, for any reason, you do not feel that an honest appraisal is possible under these circumstances, please **do not consent to the survey**. Do feel free, however, to contact me separately as it may be possible to incorporate information in the final report under stricter terms of confidentiality. You are also free to leave any question unanswered that you feel either unwilling or unqualified to respond to.

Will my participation be confidential?

Any personally identifying information provided by you will be held in accordance with the Data Protection Act and University policy on a password-protected computer at the University of Southampton. It will be deleted on completion of my PhD (October 2011 at the latest). Association of the details with the survey data will be by means of an identifying codename. All possible measures will be taken to ensure anonymity in publication, but as a colleague who worked/works on the same project may also be taking part, it is impossible to guarantee complete anonymity. Under the terms of the Data Protection Act you can request a full copy of all information held about you as a result of this survey by contacting the ECS School Office and citing the above ethics number: Electronics and Computer Science, University of Southampton, SO17 1BJ, United Kingdom. Email: school@ecs.soton.ac.uk.

What happens if I change my mind?

You are able to withdraw or amend your data at any time and for any reason prior to submission of the thesis without your legal rights being affected in any way. Should you request this, both personal and survey data provided by you will be deleted.

What happens if something goes wrong?

In the unlikely case of concern or complaint, you are welcome to contact my supervisor, Kirk Martinez, at km@ecs.soton.ac.uk.

Should he not be able to resolve the issue directly, please contact:

Lester Gilbert, Chair of Southampton University ECS Ethics Committee at l.h.gilbert@soton.ac.uk

Where can I get more information?

Please feel free to contact me with any queries about this survey at:

l.isaksen@soton.ac.uk

Appendix C

STCH Survey: Consent Form

Study title: Evaluating semantic technologies for data publication in Cultural Heritage

Researcher name: Leif Isaksen, Kirk Martinez (supervisor), Graeme Earl (supervisor)

Study reference: **[PARTICIPANT CODE HERE]**

Ethics reference: E/10/05/002

Please fill in the form below and email from a personally identifiable email address (such as work or university) to:

l.isaksen@soton.ac.uk

PARTICIPANT AND PROJECT DETAILS

These will be held securely on a Southampton University password-protected server and deleted on completion of the PhD Research (October 2011 at the latest). They will be kept separate from survey data.

1. Your name:
2. Project name:

Please initial the box(es) if you agree with the statement(s) (this is required for survey participation):

I have read and understood the information sheet (v. 3) and have ☐ had the opportunity to ask questions about the study.

I agree to take part in this research project and agree for my data ☐ to be used for the purpose of this study.

I understand my participation is voluntary and that I may with- ☐ draw or amend any information provided by me at any time prior to submission of the thesis without my legal rights being affected.

I understand that I can leave blank any question which I am un- ☐ willing or unable to answer.

Additional consents (not required for survey participation):

I am willing to be contacted by email with follow-up questions. ☐

I am willing for the information provided by me to used as an ☐ illustrative case study within a PhD thesis. I confirm that I am legally entitled under British law and that of my own country (where different) to provide said information for such purposes. I am aware that I am entitled to withdraw this consent at any time prior to submission of the thesis without my legal rights being affected.

Name of participant (print name):

Signature of participant (this can be typed as form is to be emailed):

Date:

Appendix D

STCH Survey: Questionnaire

This is a transcript of the online survey form disseminated via the SurveyMonkey¹ website. The following conventions are used to represent :

○ = Single choice answers (radio buttons)

□ = multiple choice answers (check boxes)

Evaluating semantic technologies for data publication in Cultural Heritage survey

If you are unable (or unwilling) to answer a question,

- If it is a text field, enter 'NO ANSWER'
- If it is a number field, enter '0'
- If it is a radio button, click on the 'NO ANSWER' button
- If it is a check box, click the 'NO ANSWER' checkbox and leave the rest blank

The Project

Please add additional information about the goals of the project.

1. Project code (as provided):
2. Brief project description (1 sentence):

¹<http://www.surveymonkey.com/>

3. Your role in the project

- ☐ Director
- ☐ Developer
- ☐ Content curator
- ☐ Other (please specify)
- ☐ NO ANSWER

4. Do you consider yourself to be a specialist in:

- ☐ A Computing discipline(s)
- ☐ A Humanities discipline(s)
- ☐ Both
- ☐ Neither
- ☐ NO ANSWER

5. Which term best describes the nature of the data?

- ☐ Archaeology
- ☐ Museums and Archives
- ☐ Culture
- ☐ Other
- ☐ NO ANSWER

6. The project is/was:

- ☐ Fixed term (finished)
- ☐ Fixed term (ongoing)
- ☐ Open-ended
- ☐ NO ANSWER

7. Project start year:

8. Duration of the project was/has been (months):

9. Approximately how many people were involved?

10. Approximately how many institutions were involved?

11. At least one member of the team had a background in:

- ☐ Computer Science
- ☐ Library and Information Science

- ☐ Digital Humanities
- ☐ Humanities
- ☐ NO ANSWER

12. Project aimed to:

- ☐ Link data internally
- ☐ Link data to external data
- ☐ Allow external data to link to project data
- ☐ NO ANSWER

13. Project data was:

- ☐ 'Born semantic'
- ☐ Converted from legacy digital data
- ☐ Converted from legacy paper-based data
- ☐ A thesaurus or controlled vocabulary only
- ☐ NO ANSWER

14. Approximately how many datasets were integrated?

15. Approximately how many datasets with different schemas were integrated?

16. Who are your intended consumers?

- ☐ Restricted target group(s) (e.g. project partners)
- ☐ Unrestricted target group(s) (e.g. other archaeologists)
- ☐ No restriction, no target group (anyone, including general public)
- ☐ NO ANSWER

17. In your opinion, which of the following two criteria took precedence:

- ☐ Data utility (i.e. that the output could be put to better or different use than with other technologies)
- ☐ Data integrity (i.e. that the output was an accurate representation of the semantics of the input)
- ☐ NO ANSWER

18. In your opinion, which of the following two criteria took precedence:

- ☐ Quantity over complexity
- ☐ Complexity over quantity
- ☐ NO ANSWER

Methods

Please provide information about the methods and technologies used.

1. Semantic technologies used:

- ☐ URIs
- ☐ RDF/RDFS
- ☐ OWL
- ☐ Other Vocabulary/Ontology (e.g. SKOS, CIDOC CRM)
- ☐ SPARQL end point
- ☐ Visualization tools
- ☐ Other (please specify)
- ☐ NO ANSWER

2. Other Technologies used:

- ☐ Relational database
- ☐ XML
- ☐ Text-mining
- ☐ Scripting language
- ☐ Bespoke software
- ☐ Other (please specify)
- ☐ NO ANSWER

Controlled Vocabularies

Please provide information about any controlled vocabularies used.

1. The project used controlled vocabularies (such as thesauri) that were:

- ☐ Project specific
- ☐ Defined elsewhere (please specify)
- ☐ NO ANSWER

2. Local terms were normalized before processing

- ☐ Yes
- ☐ No
- ☐ NO ANSWER

URIs

Please provide information about the project's use of URIs (where applicable).

1. If you created (minted) URIs, were they:
 - ☐ HTTP based (they begin with http://)
 - ☐ Resolvable (dereferencable) (they are accessible over the Web)
 - ☐ Persistent (they are intended to be permanent)
 - ☐ Used content negotiation/303 Redirect (they are capable of returning different representations of the concept, e.g. RDF and HTML).
 - ☐ NO ANSWER
2. If resolvable, are they hosted by:
 - ☐ An organization associated with the project
 - ☐ An external organization
 - ☐ NO ANSWER
3. Did you use a template for URI generation?
 - ☐ Yes
 - ☐ No
 - ☐ NO ANSWER
4. Are the project's URIs susceptible to decomposition? (i.e. Can the meaning of the URI be inferred without resolving it?)
 - ☐ Yes
 - ☐ No
 - ☐ NO ANSWER
5. Did the project map data to URIs defined elsewhere?
 - ☐ Yes
 - ☐ No
 - ☐ NO ANSWER
6. If Yes, how are such URIs identified?
 - ☐ By hand
 - ☐ Semi-automated process
 - ☐ Automated process
 - ☐ NO ANSWER

RDF/RDFS

Please provide information about the project's use of RDF/S (where applicable).

1. RDF generation is/was:
 - ☐ Dynamic
 - ☐ One-off Export
 - ☐ NO ANSWER
2. RDF generation was done by:
 - ☐ XSLT
 - ☐ Generic software (please specify)
 - ☐ Bespoke software
 - ☐ Other (specify)
 - ☐ NO ANSWER
3. Conversion was undertaken by:
 - ☐ Humanities specialist
 - ☐ IT specialist
 - ☐ Other (please specify)
 - ☐ NO ANSWER

Ontologies

Please provide information about the project's use of ontologies (where applicable).

1. Did the project use the SKOS (Simple Knowledge Organization System) vocabulary?
 - ☐ Yes
 - ☐ No
 - ☐ NO ANSWER
2. If you used OWL, which sublanguage did you use:
 - ☐ OWL Lite
 - ☐ OWL DL
 - ☐ OWL Full

☐ OWL2

☐ NO ANSWER

3. Did you use the CIDOC CRM?

☐ Yes

☐ No

☐ NO ANSWER

4. If there was a specific reason for this decision, please state:

5. If yes, which version did you use:

☐ CIDOC CRM

☐ CIDOC CRM Core

☐ NO ANSWER

6. Did you use any other URI-based ontologies?

Consumption

Please provide information about project data consumption (in the opinion of the interviewee).

1. Is your data open to the public in its semantic form?

☐ Yes

☐ No

☐ NO ANSWER

2. If yes, is this as:

☐ File - RDF/XML

☐ File - Turtle

☐ File - N3

☐ SPARQL

☐ RDFa

☐ Relational database with URI values

☐ Other (Please specify)

☐ NO ANSWER

3. Do you provide an alternative API or computer-readable interface (JSON, etc)?

- ☐ Yes
- ☐ No
- ☐ NO ANSWER

4. If yes, please describe:

5. Do you provide tools for visualization or an alternative human readable interface?

- ☐ Yes
- ☐ No
- ☐ NO ANSWER

6. If yes, please describe:

7. To the best of your knowledge, is your data being consumed (in its semantic format) by:

- ☐ Project partners
- ☐ Target groups
- ☐ Anyone else
- ☐ NO ANSWER

Assessment

Please provide your thoughts on the success of the approach adopted in the context of the project.

1. How much effort did the adoption of semantic technologies require in relation to the size of the project?

- ☐ Minor. Costs relative to the size of the project were low.
- ☐ Medium. Producing semantic data required a significant proportion of the project's resources.
- ☐ Major. Most of the project's resources were required for producing semantic data.
- ☐ NO ANSWER

2. Did semantic technologies live up to your expectations?

- ☐ Not at all
- ☐ Less than expected
- ☐ As expected

- ☐ Better than expected
- ☐ Much better than expected
- ☐ NO ANSWER

3. Would you advocate the use of semantic technologies for similar tasks in future?

- ☐ Definitely not
- ☐ Probably not
- ☐ Possibly
- ☐ Probably
- ☐ Definitely
- ☐ NO ANSWER

4. What was the greatest advantage(s) of using semantic technologies?

5. What was the greatest disadvantage(s) of using semantic technologies?

6. Would you do anything differently in future?

7. Any other comments?

THANK YOU FOR TAKING THE TIME TO COMPLETE THIS SURVEY.

Appendix E

STCH Survey: Participants

1. 3D COFORM
2. ADS/STELLAR
3. APELLO
4. ARACHNE
5. ArcheoInf
6. ArcheoServer
7. Area Told As Story
8. Athena
9. Athenian Agora
10. BBC
11. The British Museum
12. CIDOC CRM
13. CLAROS
14. COINS
15. Contexta/SR
16. Cultural Heritage Imaging
17. Culture Grid
18. DAI (German Archaeological Institute)
19. DART

20. Delphi
21. DERI
22. Digital Antiquity
23. EPOCH
24. Erlangen OWL CIDOC-CRM
25. Fine Rolls of Henry III
26. Geonames
27. Heurist
28. HiMAT
29. In Patrimonium
30. LGPN
31. Little House on the Hill
32. LODE
33. The Louvre
34. Meditteranean Ceramics
35. The Metropolitan Museum of Art
36. MusuemFinland
37. National Reference Collection, Netherlands
38. Nomisma
39. Open Context
40. Papyri.info
41. PhD Thesis A
42. PhD Thesis B
43. PhD Thesis C
44. Pleiades
45. Portable Antiquity Scheme
46. RCAHMS

- 47. Roman Port Networks
- 48. Science Museum
- 49. SIIS (Sagalassos)
- 50. STAR
- 51. STAR Thesauri
- 52. STITCH
- 53. TEI/CIDOC-CRM Mapping
- 54. Thucydides
- 55. Tracing Networks
- 56. The Victoria and Albert Museum
- 57. VLMA

Appendix F

STCH Survey: Data

The following tables summarize the data collected by the Semantic Technologies in Cultural Heritage (STCH) Survey. A full description of the data collection method and an analysis of the results is provided in Chapter 3. The data presented here has been filtered to only include those projects that make use of URIs, following the criterion set by Berners-Lee (2006) for Semantic Web activities. Furthermore, responses from a number of projects with multiple respondents (nos. 1, 2, 11, 17, 18, 19, 27 and 33) have been merged using a ‘common sense’ approach. Where responses are in agreement only the one answer is given. Where they differ both answers are presented unless one response clearly implies greater knowledge of the project’s activities (e.g. the adoption of a specific technology). All data has been anonymized and the numbering of projects is unrelated to that in Appendix E. Column headings are for reference only — for the full question text please refer to the questionnaire in Appendix D.

A digital version of the full (anonymized) dataset is available with the electronic version of this thesis at the University of Southampton ePrints repository.¹

¹<http://eprints.soton.ac.uk/>

TABLE F.1: STCH Survey Data 1A: Project Information

ID	Role: Developer Content curator	Specialism: Computing/ Humanities	Term: Open/ Fixed (Ongoing or Ended)	Start:	Duration: (months)	People	Institutions:	Team member from: Comp. Sci.	LIS	Dig. Hums.	Hums.
1*	x	C/H	F(E)	2005	60/48	6	2	x			
2*	x	C/H	O/F(O)	2006	54/36	30/3	7/1	x	x	x	x
3	x	C/H	O	2008	24	3	2	x	x	x	
4		C	O	2002	30	5	7	x			x
5	x	C	F(E)	2003	36	20	6	x	x	x	x
6		C	O	2003	7	20	30	x	x	x	x
7	x	C/H	F(O)	2005	62	15	3	x	x	x	x
8		C/H	F(E)	2006	24	3	10	x	x	x	x
9	x	C/H	F(E)	2006	24	30	9	x	x	x	x
10		C	F(O)	2006	47	16	4	x	x	x	x
11*	x	C	F(E)	2007	24	10/7	2/5	x	x		x
12		C/H	O	2008	30	4	3		x		
13	x	C/H	O	2008	30	5	0		x	x	x
14	x	C	F(O)	2008	20	60	19	x		x	x
15	x	C	O	2008	20	6	4	x		x	x
16	x	C/H	O	2009	6	1	1	x		x	x
17*	x	C	F(E)	2009	7	9	2	x	x	x	
18*	x	C	F(E)	2005	48	9/8	3/2	x	x	x	x
19*	x	C/H	O	2006	48/45	6/8	3/1	x	x	x	x
20	x	C	O	1972	456	100	3	x	x	x	x
21	x	C/H	F(E)	2007	36	14	5	x	x	x	
22		H	O	2009	18	6	50	x	x	x	
23	x	C/H	O	2009	18	10	5	x		x	x
24		C	O	2009	10	5	2	x			
25	x	H	F(O)	1997	100	90	63	x		x	
26	x	C/H	O	1998	144	15	1	x		x	x
27*	x	C	O	2004	70/48	25/10	6/5	x	x	x	x
28		H	O	2004	24	3	4	x	x	x	x
29	x	C	F(E)	2004	48	4	2	x	x	x	x
30	x	C/H	O	2006	48	6	2	x	x	x	x
31	x	C	O	2006	9	1	3		x	x	x
32		C/H	O	2007	48	2	2		x	x	x
33*	x	C/H	F(E)/F(O)	2007	46/36/36	6/4/10	3/2/3	x	x	x	x
34	x	C/H	F(O)	2008	20	4	3	x	x	x	
35	x	H	F(O)	2008	60	8	4	x	x	x	x
36		C	F(O)	2009	60	25	9	x			
37		H	O	2009	24	2	1			x	
38	x	C/H	FF(E)	2009	12	5	3	x		x	x
39	x	C/H	F(O)	2010	6	8	5	x		x	x
40		C/H	F(O)	2010	3	30	20	x	x	x	x

* = multiple respondents

TABLE F.2: STCH Survey Data 1B: Project Goals

ID	Data type: M.&A./Arch./ Culture/Other	Aim: Internal links	Outgoing links	Incoming links	Source: Born semantic	Legacy digital	Legacy paper	Vocab only	Datasets:	Schemas:	Consumers: Anyone/Targetted/ Restricted	Priority: Utility vs. Integrity	Complexity vs. Quantity
1*	Arch/O	x	x	x	x	x	x	x	0	0	A	U	-
2*	O	x	x	x	x	x	x		1	1	A	U/I	-
3	O		x	x	x	x			3	3	A	U	Q
4	M&A	x	x	x	x	x			4	4	A	I	C
5	M&A	x	x	x	x				4	4	A	I	C
6	M&A	x	x	x	x	x			25	20	A	U	C
7	M&A	x	x	x					1	2	TG	I	C
8	M&A		x	x	x		x		10	3	RG	I	Q
9	M&A		x	x	x				3	3	RG	I	Q
10	M&A	x	x	x					1	0	TG	U	-
11*	M&A	x	x			x	x	x	3/1	5/3	A	U/I	C
12	M&A	x	x			x	x		4	3	A	U	C
13	M&A	x	x	x		x	x		2	0	A	U	-
14	M&A	x	x	x	x				0	0	TG	I	C
15	M&A		x	x		x			4	4	A	U	Q
16	M&A			x		x			2	2	A	U	Q
17*	M&A		x			x			3	3	TG	U	Q
18*	C/O	x	x			x		x	6/4	6/4	TG/RG	U	C/Q
19*	M&A/C	x	x		x	x	x		4/1	6/3	A	U	Q
20	C		x	x		x			1	1	A	U	-
21	C	x	x	x		x			3	3	A	I	-
22	C		x	x		x			50	40	A	U	Q
23	C	x	x	x	x				5	5	A	U	Q
24	C		x	x					2	2	TG	U	C
25	Arch	x	x	x		x		x	10	3	A	U	-
26	Arch	x	x	x	x	x	x		10	10	A	U	Q
27*	Arch	x	x	x	x	x			12	4	A	I	C
28	Arch		x	x		x			3	3	A	U	C
29	Arch	x	x	x		x			3	3	A	U	Q
30	Arch		x	x		x			15	15	A	U	C
31	Arch	x	x	x		x			2	2	TG	I	C
32	Arch	x	x	x		x		x	1	1	A	U	Q
33*	Arch	x	x	x	x	x			4/4/3	4/4/3	TG	U	C
34	Arch	x	x	x		x			20	15	A	U	Q
35	Arch		x	x		x			8	8	TG	U	C
36	Arch	x	x	x		x			2	2	TG	U	Q
37	Arch	x				x			5	5	RG	I	C
38	Arch	x							3	3	A	I	Q
39	Arch	x	x	x		x			10	10	A	U	C
40	Arch		x	x	x			x	0	0	A	U	-

* = multiple respondents

TABLE F.3: STCH Survey Data 2 & 3: Methods and Controlled Vocabularies

ID	Semantic technologies:				Other technologies:				Other technologies:				Controlled vocabcs:		Normalized
	URIs	RDF/ RDFS	OWL	Other Vocab.	SPARQL	Visual- ization	Other	Relational database	Text- mining	Scripting language	Bespoke software	Project specific	Defined externally		
1*	x	x	x	x	x	x		x	x	x	x	x	x	No	
2*	x			x		x		x			x	x	x	-	
3	x	x	x	x					x	x	x		x	-	
4	x	x						x					x	No	
5	x	x		x		x		x		x	x	x	x	No	
6	x	x	x	x	x	x		x	x	x	x			No	
7	x	x	x	x				x		x	x	x		No	
8	x							x		x	x	x	x	Yes	
9	implicit	x						x		x		x		Yes	
10	x			x				x		x		x		-	
11*	x	x	x	x	x	x		x	x	x		x		Yes	
12	x	x	x		x	x	x	x	x	x		x		Yes	
13	x							x		x			x	-	
14	x	x		x				x	x	x	x	x	x	Yes	
15	x	x		x		x		x		x			x	-	
16	x						x	x		x			x	-	
17*	x	x	x		x		x	x		x	x	x	x	No	
18*	x	x	x	x	x			x		x	x	x	x	Yes/No	
19*	x	x		x				x	x	x	x	x	x	No	
20	implicit	x		x		x		x		x	x	x		Yes	
21	x	x			x			x		x	x			No	
22	x			x			x	x				x		Yes	
23	x	implicit		x	x	x		x	x	x	x			Yes	
24	x	x	x	x	x	x			x	x	x			Yes	
25	x	x	x	x	x			x	x	x	x			Yes	
26	x	x			x	x		x	x	x	x			Yes	
27*	implicit	x	x		x		x	x		x			x	-	
28	implicit	x		x				x		x	x	x		No	
29		x	x	x	x			x	x	x	x	x	x	Yes	
30	x						x	x		x		x	x	Yes	
31	x	x	x	x	x	x		x		x				No	
32	x	x		x	x	x		x		x		x	x	Yes	
33*	x	x	x	x	x	x	x	x	x	x	x		x	Yes	
34	x	x	x	x				x	x	x			x	Yes	
35	x	x	x	x	x	x		x	x	x				No	
36	x	x	x	x	x	x		x						No	
37	implicit	implicit	x					x				x	x	Yes	
38	implicit	x	x	x		x			x	x				Yes	
39	x	x	x	x	x			x		x			x	-	
40	x	x	x	x	x	x	x	x		x	x		x	-	

* = multiple respondents; 'implicit' = inferred from use of dependent technology

TABLE F.4: STCH Survey Data 4: URIs

ID	URIs were:		Hosted by:		Template:	Decomposable:	External mapping:	Mapping process:
	HTTP based	Dereferencable	Persistent	Content Negotiation	Partner/External			By hand/Automated/Semi-automated
1*	x	x	x		Partner	-	No	
2*	x	x	x	x	Partner	Yes	No	By hand/Automated
3	x	x	x	x	Partner	No	Yes	Semi-automated
4	x	x	x		-	No	Yes	
5	x	x			Partner	No	No	
6	x	x	x	x	-	Yes	No	Semi-automated
7	x	x	x		Partner	Yes	Yes	
8	x	x			External	Yes	No	
9	x				Partner	Yes	Yes	Automated
10					-	-	No	Automated
11*	x		x		Partner	Yes	Yes/No	
12	x	x			Partner	No	Yes	By hand
13	x	x			Partner	Yes	Yes	
14	x	x	x		Partner	No	No	
15	x	x	x		-	No	Yes	By hand
16	x	x	x		Partner	-	Yes	Semi-automated
17*	x				External	Yes	-	
18*	x	x		x	Partner	Yes/No	Yes	
19*	x	x	x		Partner	Yes/No	No	Semi-automated
20	x	x	x		Partner	Yes	Yes	
21	x				Partner	Yes	Yes	Semi-automated
22	x	x	x		Partner	Yes	Yes	
23	x	x	x		Partner	No	Yes	
24	x	x	x	x	Partner	Yes	Yes	Semi-automated
25	x	x	x	x	-	No	-	Semi-automated
26	x	x	x		Partner	Yes	Yes	
27*	x	x	x		Partner	-	No	
28			x		Partner	-	Yes	Semi-automated
29	x			x	Partner	Yes	Yes	By hand
30	x	x	x	x	Partner	Yes	Yes	Semi-automated
31	x			x	Partner	Yes	Yes	
32	x	x	x		Partner	No	Yes	By hand
33*	x	x	x	x	Partner	No	Yes	Semi-automated
34	x	x	x	x	Partner	Yes	Yes	
35	x				Partner	Yes	No	
36	x	x		x	Partner	Yes	Yes	Semi-automated
37					-	No	No	
38	x				-	Yes	-	
39	x	x			External	Yes	-	
40					External	-	-	

* = multiple respondents

TABLE F.5: STCH Survey Data 5 & 6: RDF & Ontologies

ID	Export process:	Export tool:	Conversion by specialist in:			Ontologies used:			Version	Other
			Humanities	IT	Other	SKOS	OWL	CIDOC CRM		
1*	Dynamic	XSLT	x	x		Yes/No	OWL Lite	Yes	-	Yes/No
2*	Dynamic	Bespoke software				No	-	no		Yes/No
3	One-off	Bespoke software	x	x		No	OWL Lite	Yes	CIDOC CRM	Yes
4	Dynamic	Bespoke software	x	x		No	-	No		No
5	One-off	Bespoke software	x	x		No	-	Yes	CIDOC CRM	No
6	Dynamic	Bespoke software	x	x		No	-	No		Yes
7	One-off	XSLT	x	x	x	Yes	OWL DL	Yes	CIDOC CRM	Yes
8	One-off	XSLT	x	x		-	-	Yes	CIDOC CRM	-
9	Dynamic	Generic software	x	x		Yes	OWL DL	Yes	CIDOC CRM	No
10	Dynamic	Generic software	x	x		No	-	Yes	CIDOC CRM	No
11*	Dynamic	XSLT	x	x		Yes	OWL Lite	Yes	CIDOC CRM	Yes/No
12	Dynamic	Bespoke software		x		Yes	OWL DL	Yes	CIDOC CRM	No
13	Dynamic	Bespoke software				No	-	No		No
14	Dynamic	Bespoke software		x		Yes	-	Yes	CIDOC CRM	No
15	One-off	Bespoke software		x		No	OWL2	Yes	CIDOC CRM	No
16	Dynamic	XSLT		x		No	-	No		-
17*	Dynamic/One-off	Bespoke software	x	x		No	-	No		No
18*	One-off	Bespoke software/XSLT	x	x		Yes	-	No		-
19*	One-off	Bespoke software/XSLT		x		-	-	No		No
20	Dynamic	XSLT		x		No	-	Yes	CRM Core	No
21	Dynamic	XSLT	x	x		No	-	No		Yes
22	Dynamic	XSLT				Yes	-	No		Yes
23	Dynamic	XSLT	x	x	x	Yes	-	No		Yes
24	Dynamic	Bespoke software	x	x		Yes	-	-		-
25	Dynamic	Bespoke software	x	x		Yes	-	-		-
26	Dynamic	Bespoke software				No	-	No		No
27*	Dynamic	Bespoke software				No	OWL Lite	No		No
28	One-off	Bespoke software		x		-	-	No		No
29	Dynamic	XSLT	x	x	x	Yes	OWL DL	No		No
30	Dynamic	XSLT	x	x		Yes	-	No		Yes
31	One-off	XSLT	x	x		No	-	Yes	CIDOC CRM	No
32	One-off	XSLT		x		Yes	OWL Lite	No		-
33*	Dynamic	Bespoke software	x	x		Yes	-	Yes	CIDOC CRM	Yes
34	Dynamic	Other		x		No	OWL Full	Yes	CIDOC CRM	Yes
35	Dynamic	Bespoke software		x		Yes	-	Yes	CIDOC CRM	Yes
36	Dynamic	Generic software		x		Yes	OWL2	Yes	CIDOC CRM	Yes
37	One-off	Bespoke software				No	OWL Lite	Yes	CIDOC CRM	No
38	One-off	Bespoke software	x	x	x	No	OWL DL	Yes	CIDOC CRM	No
39	One-off	Bespoke software		x		Yes	-	Yes	CIDOC CRM	-
40	One-off	Bespoke software		x		-	-	-		-

* = multiple respondents

TABLE F.6: STCH Survey Data 7: Consumption

ID	Semantic formats:				Public	Alt. API	GUI	Consumed by:		
	Available: RDF/XML Turtle							Partners	Target groups	Others
1*	Yes/No	x			Yes	-	Yes			
2*	No				Yes	Yes	Yes		x	x
3	Yes	x			Yes	No	Yes			
4	Yes	x			Yes	No	Yes		x	
5	Yes	x			No	Yes	Yes			x
6	Yes			x	No	Yes	Yes		x	
7	No				-	No	Yes			
8	No				No	No	No			
9	Yes	x			Yes	No	No			
10	No				No	No	No			
11*	Yes	x			Yes	Yes/No	Yes/No			
12	Yes	x			Yes	No	No		x	x
13	Yes	x			Yes	-	No			
14	No				-	Yes	Yes			
15	No				-	Yes	Yes		x	
16	Yes				Yes	Yes	Yes			x
17*	Yes	x			Yes	No	Yes/No			
18*	Yes	x			No	Yes	Yes/No		x	x
19*	Yes/No				No	No	Yes		x	x
20	Yes	x			Yes	Yes	Yes			
21	Yes	x			Yes	Yes	Yes		x	
22	Yes	x			Yes	Yes	No			
23	Yes				Yes	Yes	Yes			
24	Yes	x			Yes	Yes	Yes			
25	Yes	x			Yes	Yes	No			x
26	Yes	x			Yes	Yes	Yes			
27*	Yes/No				Yes	No	No		x	x
28	Yes	x			Yes	-	-			x
29	Yes				No	No	No			x
30	Yes				Yes	Yes	Yes		x	x
31	Yes	x			No	No	Yes			
32	Yes				Yes	No	Yes			
33*	Yes	x			No	Yes/Yes/No	Yes			x
34	Yes	x			Yes	Yes	Yes		x	
35	-				-	No	Yes			
36	Yes	x			No	Yes	Yes			
37	Yes	x			No	No	No		x	
38	No				No	Yes	Yes			
39	-				Yes	No	No			
40	-				Yes	-	-			

* = multiple respondents

TABLE F.7: STCH Survey Data 8: Satisfaction

ID	Effort:		Met. expectations:		Advocate:	
	Major	Medium/Minor	Much better/Better/As expected/Less/Much less	Probably/Possibly	Definitely/Probably/Possibly	Definitely not/Definitely not
1*	Major		As expected		Probably/Possibly	
2*	Major	Minor	As expected		Probably/Possibly	
3	Major		As expected		Possibly	
4	Major		Much better		Definitely	
5	Medium		Less		Probably	
6	Medium		As expected		Definitely	
7	Major		As expected		Probably	
8	Major		Better		Definitely	
9	Medium		Better		Definitely	
10	Medium		As expected		Probably	
11*	Major		Better/As expected		Definitely/Probably	
12	Medium		As expected		Definitely	
13	Major		Better		Definitely	
14	Minor		As expected		Definitely	
15	Medium		As expected		Probably	
16	Minor		-		Definitely	
17*	Medium/Minor		As expected		Definitely	
18*	Medium		As expected/Less		Definitely/Probably	
19*	Medium		Much better/Better		Definitely	
20	Minor		As expected		Definitely	
21	Minor		Much better		Definitely	
22	Minor		As expected		Probably	
23	Medium		Much better		Definitely	
24	Medium		As expected		Definitely	
25	Minor		As expected		Definitely	
26	Minor		Less		Probably not	
27*	Medium		As expected		Definitely/Possibly	
28	Minor		Less		Definitely	
29	Major		As expected		Definitely	
30	Major		Better		Definitely	
31	Major		Less		Probably	
32	Minor		As expected		Definitely	
33*	Medium/Minor/Minor		Better/Better/Less		Definitely/Probably/Possibly	
34	Major		Much better		Definitely	
35	Major		As expected		Definitely	
36	Medium		As expected		Definitely	
37	Major		As expected		Definitely	
38	Medium		As expected		Probably	
39	Major		Less		Possibly	
40	-		-		Definitely	

* = multiple respondents

TABLE F.8: STCH Questions 3.2 & 6.6: Other vocabularies and ontologies used

Vocabulary	URL (where known)
AAT	http://www.getty.edu/research/tools/vocabularies/aat/
Archaeology Basic Register	—
ArchaeoML	http://www.alexandriaarchive.org/archaeoml.php
Brinkman	http://netuit.kb.nl/
Classical Atlas Project	http://www.unc.edu/depts/cl_atlas/
Concordia	http://www.atlantides.org/trac/concordia
DBpedia	http://dbpedia.org
DCTerms	http://dublincore.org/2008/01/14/dcterms.rdf
EH Period List	http://www.fish-forum.info/i_apl.htm
EpiDoc	http://epidoc.sourceforge.net/
GeoNames	http://www.geonames.org/
GeoRelations	http://www.mindswap.org/2003/owl/geo/geoRelations.owl
Glas	http://www.referentiecollectie.nl/richglas/glascollectie.php
GOO	http://goo.kb.nl/
GTAA	http://ems01.mpi.nl/CHOICE/
Iconclass	http://www.iconclass.nl/
LCSH	http://authorities.loc.gov/
Mandragore	http://mandragore.bnf.fr/jsp/classementThema.jsp
MIDAS	http://www.heritage-standards.org.uk/midas/docs/
NBC	http://goo.kb.nl/basisclassificatie.html
NBD/Biblion	—
OWL-Time	http://www.w3.org/TR/owl-time/
Rameau	http://rameau.bnf.fr/
Regiothesaurus	—
SWD	http://www.d-nb.de/standardisierung/normdateien/swd.htm
ULAN	http://www.getty.edu/research/tools/vocabularies/ulan/
Wordnet	http://wordnet.princeton.edu/

TABLE F.9: STCH Questions 2.2 & 5.2: Other technologies used

Technology	URL (where known)
D2R	http://www4.wiwiw.fu-berlin.de/bizer/d2r-server/
Faceted Search	–
Freebase (MQL)	http://wiki.freebase.com/wiki/MQL
Jena	http://jena.sourceforge.net/
Lucene/Solr	http://lucene.apache.org/solr/
Named Entity Recognition	–
Object database	–
RTI imaging software	–
SPARQLite	http://code.google.com/p/sparqlite/

TABLE F.10: STCH Question 7.4: Other APIs provided

API	URL (where applicable)
Atom	http://www.ietf.org/rfc/rfc4287.txt
CSV	–
JSON/JSON-P/GeoJSON	http://www.json.org/ http://json-p.org/ http://geojson.org/
KML	http://code.google.com/apis/kml/
REST	–
RSS/GeoRSS	http://www.rssboard.org/rss-specification http://georss.org/
SRW	http://www.loc.gov/standards/sru/srw/index.html
TEI-XML	http://www.tei-c.org/
Vocabulary web service	–

TABLE F.11: STCH Question 7.6: Other human interfaces provided

Interface	URL (where applicable)
Custom software	–
Facetted browser	–
GIS	–
Google Maps	http://maps.google.com/
Index/Search form	–
Mobile app	–
Open Layers	http://openlayers.org/
Sig.ma	http://sig.ma/
SIMILE Timeline	http://www.simile-widgets.org/timeline/
TimeMap.js	http://code.google.com/p/timemap/

STCH Question 8.4: What was/were the greatest advantage(s) of using semantic technologies?

1. Flexibility.
2. Unique, sharable, dereferenceable http: URIs make our resources part of the web.
3. Simple combination of different vocabularies.
4. Data integration and intelligent user interface.
5. Common semantics across museum collections. Common approach to defining semantics of individual collections.
6. Data integration, intelligent user interfaces, APIs.
7. The urge to make all possible knowledge (even the most trivial) explicit, the advantage of counting on previously created conceptual structures (e.g. CIDOC-CRM) and the potential of sharing data and interpretations.
- 8.
9. Data exchange and data integration.
10. In our case we were developing tools for imaging in CH, and wanted the digital representations built using the tools to carry certain process history information, as well as be able to link to other relevant data. So, the goal was to establish some examples of how to do that, which could be built into future tools.
11. Naturality of models and data (essential for catalog integration and for external data anchoring). / Knowledge management using domain specific ontologies, faceted searches.
12. Making CRM available in machine processable form.
- 13.
14. Primary basis for accessing underlying data resources. Support for multilingual access.
15. Providing a baseline for convergence of diverse data sources. Open-ended nature of data structuring used, allowing an incremental approach to data integration.
16. Hoping for cross-collections links with other institutions with machine-readable data as well as internal re-use of data.
17. The project did not exploit the advantages of semantic technologies and only the project requirements were fulfilled. However, the advantages of SW technologies were greatly highlighted and appreciated and we are now in a position to build more semantically rich applications.

18. Standard ontologies (especially SKOS), ease of representing land consuming links between resource, flexibility of modeling, web-friendliness. / Hard to say. We use the techniques because people with knowledge of it were at our disposition and they wished to and had the possibilities to do the experiments with us.
19. Semantic technologies enabled us to greatly enlarge the amount of usable data from an otherwise often sparsely populated, text-block-heavy data set. This was done through text mining and ontology latching, with concurrent application of hand-coded semantic logic (e.g., all objects from Location X and found by Person Y are part of Collection Z). The ability to automatically, repeatedly, and quickly convert from a museum professional point of view to a public point of view.
20. Comparability with other projects in future.
21. They just worked. Using an RDF triple store solved a problem of relating data from multiple sources that had plagued the project from its outset.
22. The incredible richness of data harvested/aggregated in an uncontrolled environment makes structured classification difficult, whereas semantics provide a more organic approach to information structure, search and retrieval.
23. Simple structures provide powerful mechanisms for user interactions.
24. Integration.
25. Opening up our content to a new audience and learning new technical skills.
- 26.
27. Integrating data with dissimilar schemas & allowing users not to have to know detailed vocabularies specific to a system to obtain good results.
28. Making sense across databases.
29. The ability to link disparate datasets easily is the greatest advantage, in my opinion.
30. Let's be clear here about "semantic". We map to a common XML (but not RDF-based) vocabulary for expressing archaeological data, and we do lots of entity identification to mint URIs. We do lots of schema mapping of several different datasets, so this is a "semantic" project in that sense. But, I'm not a huge fan of RDF-based technologies, since I think Plain-XML is easier to deal with. I nice phrase I heard was that managing RDF is like handling grains of sand. I prefer a bucket (XML documents). I think these give me adequate flexibility for meeting our project goals, and we are slowly adopting some RDFa to reference useful outside entities (GeoNames places, Concordia vocabulary concepts). But I'm very happy using ArchaeoML to represent our structured data and then transforming these documents into XHTML+RDFa as needed.

31. Concepts and tools available, lively community, interesting research topic and future uses, good for sharing data that is already structured.
32. Greater exposure and standardization.
33. Cross-searching made possible on data sets that would otherwise not have happened. / Pulling together and cross searching disparate data / Cross searching of datasets.
34. Radical transparency of previously opaque, specialized data. Ability to reuse/remix the now 'atomic' data in unforeseeable ways.
- 35.
36. Sharing data across partners.
37. It was the understanding about the implicit knowledge underlying our dataset.
38. Integrating heterogeneous and legacy data source using a common semantic schema.
- 39.
40. Too early to answer this. The adaptation costs are likely to be low in the long term: I'm front loading costs to the early, set-up, stage so that they become part of the method and process. Wherever practicable all data will be made accessible via the Talis triple store.

STCH Question 8.5: What was/were the greatest disadvantage(s) of using semantic technologies?

1. Performance.
2. Uncertain uptake, final version. / Lack of mature ontologies designed for the web, especially in bibliography and citation. Good URI hygiene can be difficult when the data is evolving.
3. Complexity of toolset, buggy software.
4. Integration with legacy systems requires effort from the memory organizations.
5. CIDOC CRM is bloody hard to understand and use with zero tool support available at the time. Museum bods are understandably not knowledge engineers, so require lots of support.
6. Integrating legacy content production to the system is hard.
7. Performance issues, immaturity of tools, naivety of practices.
- 8.

9. Not immediately understandable by CH specialists.
10. It's complex. Our goal is to keep the complexity away from our users, but that means a lot of work on our project to figure out how to best use the technologies to meet our needs. The lack of good clear examples, with user appropriate tools, for using the technologies is a disadvantage.
11. Lack of good practices (in our team) and of trained human resources. / Technologically speaking, slow response times for complex queries.
12. None.
- 13.
14. Accessing data to populate semantic metadata. Managing partial and inconsistent resources (thesauri etc.).
15. Query performance.
16. Time/resources to work out what schemes/ontologies should be used and how to map them to our data.
17. No full-text based searching solutions.
18. The SKOS standard was not stable at the time we started, we had to get involved in it which costed us time ;-) From a more technical perspective scalability was sometimes an issue. / Complex, not many applications make use of these techniques; not easy to understand for non-experts.
19. Dealing with museum staff expectations for accuracy of semantically produced data, yet not being able to offer them the tools to allow them to improve the ontologies and semantic logic. / Considerable effort was required to build the required ontologies that present the public POV.
20. None.
21. At this point I really see only an upside.
22. There are very few end-user services which make meaningful use of it.
23. Time to mark entities.
24. Tool set limited in functions.
25. Learning how to implement it and finding relevant examples in the archaeological/museum world.
26. Processing overhead involved in extravagant abstraction (both memory and time). Lack of functional benefits.

27. Effort to build community consensus on ontologies (underway).
28. Hard for normal people to grasp; hard to get participants.
29. Grounding local URIs against authoritative ones is hard, but necessary for powerful linking. There is a danger that locally minted URIs actually make linking resources harder, if they are too local, obscure, hard to interpret etc.
30. I actually really dislike RDF(a) syntax and find it less than intuitive for anything beyond trivial assertions, so I'm sticking to trivial assertions for now. It's much more immediately rewarding to focus efforts on generating Atom+GeoRSS feeds (and less so JSON), because this is very well supported, simple, and still powerful, especially for sharing lists of URIs. Also, I still don't see much reason to invest in RDF based technologies, except that "all the cool kids are doing it", and maybe grant review committees would like to see it as evidence of being cutting edge or standards-aware. The main value I see in RDFa is in things like "OWL:sameAs", where I can say some entities in [anonymized] are the same as some places in GeoNames. However, it would be really hard and expensive to do this more generally. We'll do this for some biological taxa because these are pretty easy to identify (but still not trivial, since there are zooarchaeological concepts like "sheep/goat" with no clear analog in widely used controlled vocabularies). However, most useful archaeological concepts (different pottery styles, archaeological periods, etc.) don't have canonical URIs. Besides, making assertions like this is an interpretive decision that I am loath to "hard-code" into [anonymized]. I'd rather just mint a bunch of URIs for our published data, and expose these data via simple RESTful services. I'll let other people make those assertions based on their own interpretive choices.
31. Complexity, verbosity, scalability
32. None.
33. Scalability and difficulty of confirming cost benefits of the resulting large triple store for complex queries. / Limited support for commonly required search functionality (within SPARQL language), query performance issues.
34. Incredible amount of work. We opted to define an architecture that would ultimately require almost no formal mapping of terms (simplest triple-based format for entities, properties and relations). We got there, but it was an epic journey.
- 35.
36. Training no semantic web users.
37. Although the ontology schema is easy to design and create, the mapping of data it is still difficult.

38. Time of development (too much), difficulty to explain the complexity and potentials of the approach to the project partners (especially to archaeologists).
- 39.
- 40.

STCH Question 8.6: Would you do anything differently in future?

1. Yes, semantic technologies need to developed to a more mature state constantly.
2. Trade ‘complete’ for ‘good enough’ earlier. Shipping (as Joel Spolsky says) is pretty much the most important feature.
3. Focus on getting data used before worry about modeling too much.
4. Semantic content creation should be supported at the memory organizations. Solving interoperability problems is much harder afterwards if data quality is lower.
5. Wouldn’t use full blown CRM! Things have moved on a lot since the project with simpler reference models e.g. CRM core probably yielding a better return on investment, i.e. still significant benefits, but with a lot lower investment, and with linked data providing a much better model for linking across collections, including 303 redirection.
6. More focus on creating high quality semantic content. Content is the king.
7. I would insist for the data to be made publicly available; study better how to make the framework in which the ontology was developed and stored lighter.
- 8.
9. Not differently but in a more clever way.
10. We need to establish some examples for imaging, and get buy-in from others so that we and others do not have to figure this out from scratch for every project or new set of imaging tools.
11. Get a semantic specialist technologies to advise on adoption of some in-house decisions.
12. No.
- 13.
14. More project resource on semantic issues.
15. To achieve the initial goal we set ourselves, probably not. Maybe earlier attention given to query performance? It’s possible that knowing what we now know, that the project goals would be formed differently, but I’m not sure how.

- 16.
- 17.
18. Maybe.
19. 1. Have a budget (this was done for \$0 with volunteered time and effort). 2. Build/use tools to give museum staff (and qualified external experts) the ability to contribute to and improve the ontologies and semantic logic. 3. Focus more effort on institutional buy-in. While the project is viewed as an overwhelming success by its intended (external) users, it is frequently viewed as a failure by museum staff who have expectations of 100% data accuracy and of all conceivable functionality. — I would have leveraged ETL tools for initial extraction and pre-processing of data from the source DB.
20. No.
21. Start using RDF sooner.
22. We will be using the semantically-enriched material to generate more visualisations to provide richer mechanisms for interrogating datasets. We would probably define a core ontology to map to , to keep things controlled.
- 23.
- 24.
25. Have a larger team to assist with creating the semantic output.
- 26.
27. Would have been good for me to understand these technologies better before we started (and now for that matter).
28. Try to be born semantic!
29. I would spend more time on normalisation of local URIs, to merge co-referential ones.
30. Lots of things, but these have more to do with marketing and focusing on data quality more earlier in the life of our project. I think that the semantics issue is interesting and am willing to do more should a compelling application/reason present itself. If we have users. / Colleagues demanding more RDFish data, we'll produce it. But so far, the demand has been virtually nil.
31. Metadata creation according to functional requirements.
32. The project is ongoing and I look to take advantage of the growing semantic web.

33. Possibly consider exploring federated triple stores for different types of data sets. / Project used fairly limited .NET component for triplestore functionality — would instead adopt a commercial engine e.g. AllegroGraph
34. We established development sprints late in the game. Formalizing development sprints actually liberates creativity and sets expectations. Money would have been useful, we really needed programming and semantic resources.
- 35.
36. Assess the need of the whole project, instead of dealing with individual subprojects.
37. We need tool to map dataset on the ontology.
38. Include more partners having knowledge and skills in Semantic Web models and technologies in order to better manage and deploy new projects.
- 39.
- 40.

STCH Q.uestion 8.7: Any other comments?

- 1.
- 2.
- 3.
- 4.
5. It was hard work being at the bleeding edge of semantic web R&D as we were at the time — few standards, shifting sands, little tool support, lack of awareness of the benefits in end users etc.
- 6.
7. It was very exciting to work with semantic technologies: it opens up a world that goes much beyond simply syntactic technologies such as XML. I think to express at best its potential, a case study involving data useful to more than one project would have made the all endeavour more interesting.
- 8.
- 9.

10. I'm not sure how well our project fits the survey, since we weren't actually working with CH data directly, but rather trying to build in semantic knowledge management for a new set of imaging tools. We explored a number of ideas, some seem more promising than others, but the specific project funding didn't allow us to go as far with this as we would have liked.
11. The project was lots of fun, and we are continuing now with a digital libraries-oriented semantic indexing project.
- 12.
- 13.
- 14.
15. This is an ongoing project with a shoestring budget (for the data integration aspects) — there are many things we would like to have done, and hopefully will do, so the response above is really in relation to whether we could have done more with the available resources.
16. This is an on-going project, largely fit in around available resources and funded projects, so some of the questions were in the wrong tense for me!
- 17.
- 18.
19. From my perspective, this project has been an overwhelming success, having provided full public access to previously unavailable (and often unusable) data. I would definitely recommend semantic technologies to anyone looking to improve access to specialized, obscure, jargonized, and/or sparse data. / We are still actively developing new features allowing humanities domain experts to maintain the ontologies, and to create inferencing rules that enrich the metadata.
- 20.
21. At this point, the data in our triple store drives the generation of static views, search indexing, and the importing of data into the editing tool. It is the linchpin of the whole project. We have not used OWL, SKOS, or CIDOC-CRM yet, but we may in the future.
- 22.
- 23.
- 24.
- 25.

26. There might be an advantage in terms of abstracting multiple complex datasets (such as mapping [anonymized]'s archaeological data and the [anonymized]'s into something like Dublin Core) but such mappings seem only to produce a more generalized dataset. Most data consumers will typically prefer the richer set if it's available. In terms of using semantic technologies at a local level, most needs can be more efficiently met by using simpler data structures and relationships. Or such has been our experience.
27. Some semantic work on data integration is operational, other is under development. I think some of this is relatively lightweight (but intended for large scale) compared to what others are doing. / Currently, we are using OWL ontologies with an eye to the future but not to their fullest potential. As far as I understand it, the main features we want from ontologies are: (1) structure (2) relationships but we currently only utilize (1). A lot of our project focus thus far has been developing a maintainable web application to extract semantic metadata at the appropriate granularity. As a result, our level of integration with semantic technologies is still quite the moving target — as we learn more about them, we need to evaluate which approach is the most suitable for our requirements taking into account maintainability and ease of development.
- 28.
- 29.
- 30.
- 31.
- 32.
33. For the neophyte semantic modelling techniques can seem impractical, rigid and unrealistic. Implementation issues often descend into ethereal discussions about meaning itself, with no clear usable outcome.
- 34.
- 35.
- 36.
- 37.
- 38.
- 39.
- 40.

Appendix G

Roman Port Networks: Partners

Institutional Partners

Ausonius (Université Bordeaux 3 — CNRS), Bordeaux

Corinne Sanchez (Ceramics: Narbonne)

British School at Rome

Simon Keay (also University of Southampton)

Roberta Cascino (Administration) and Steve Kay (Computing)

Centre Camille Jullian (Aix-Marseille Université — CNRS), Aix-en-Provence

Giulia Boetto (coordination and shipwrecks)

Michel Bonifay (coordination and ceramics)

Marie-Brigitte Carre, Antoinette Hesnard and Céline Huguet (Ceramics: Marseille)

Lucien Rivet and Sylvie Saulnier (Ceramics : Fréjus)

André Tchernia and Catherine Virlouvet (Trade History)

DRASSM, Marseille

Hélène Bernard (marble and transport)

Marie-Pierre Jézégou (shipwrecks, Narbonne)

Frédéric Leroy (GIS)

Luc Long (shipwrecks, Arles)

Florence Richez (Advisor, documentation archive)

Institut Catalá d'Arqueologia Clàssica

Isabel Rodà

Anna Gutirrez (Marble)

Marta Prevosti (Ceramics)

Arnau Fernández Trullen (Ceramics)

Institut National de Recherches Archéologiques Préventives (INRAP)

Stéphane Bien (Ceramics: Marseille)

Susanne Lang (Ceramics: Marseille)

Musée Archéologique Départemental d'Arles

Jean Piton and David Djaoui (Ceramics: Arles)

Musée Archéologique d'Istres

Frédéric Marty (Ceramics : Fos-sur-Mer)

Museu Nacional Arqueològic de Tarragona

Josep Anton Remolà (Ceramics : Tarragona)

National Museum of Denmark

John Lund (Ceramics)

Parsifal Cooperativa (Rome)

Fabrizio Felici (Ceramics from hinterland of Leptis Magna)

Sabrina Zampini (Ceramics from Portus, Ostia and Rome)

Sergio Fontana (Ceramics from north Africa)

Pôle archéologique départemental du Var

Chérine Gebara (Ceramics: Fréjus)

Patrik Digelmann (Marbles: France and Fréjus)

Service du Patrimoine, Fréjus

Michel Pasqualini (Ceramics: Fréjus)

Soprintendenza Speciale per i Beni Archeologici di Napoli e Pompei

Daniela Giampaola (ceramics: Naples)

Universidad de Barcelona (CEIPAC)

Jose Remesal Rodríguez (Amphora stamps) and others

Università di Bologna

Andrea Augenti and Enrico Cirelli (Ceramics: Ravenna/Classe)

Universidad de Cadiz

Dario Bernal Casasola and others (Ceramics: Cádiz and Baelo)

Università di Catania/CNR

Daniele Malfitana and others (Ceramics: Sicily)

University of Ghent

Patrick Monsieur (Ceramics)

Universiteit de Leuven (ICRATES)

Jeroen Poblome and Michel Bes (Ceramics: East Mediterranean)

Université de Nice-Cimiez

Pascal Arnaud (Ancient Navigation)

University of Oxford

Andrew Wilson (Ceramics: Berenike Euhesperides)

University of Pisa

Marinella Pasquinucci (Portus Pisanus and Vada Volterrana)

Universidad de Sevilla

Jose Beltrán Fortes (Marble: Seville, etc.)

Enrique Garcia Vargas (Ceramics: Seville, etc.)

University of Southampton

Graeme Earl (GIS and Computing)

Lucy Blue (Ancient harbours, navigation, Alexandria)

David Wheatley (GIS and Computing)

Leif Isaksen (Databases and Semantic Technologies)

David Potts (GIS)

Università di Urbino

Fede Berti (Marble)

Individual Collaborators

Tamas Bezecky (Vienna: Ceramics from Ephesus etc)

Matthias Bruno (Rome: Marble from Portus and Mediterranean sites)

Cesar Carreras (Barcelona: Ceramics from Barcelona and Catalan sites)

Vittoria Carsana and collaborators (Ceramics: Naples)

Timmy Gambin (University of Malta/Aurora Trust: Ceramics from Malta)

Roberta Tomber (British Museum: Ceramics from Egypt)

Appendix H

Introducing Semantics Questionnaire

Thanks for taking the time to use the *Introducing Semantics* guide. Please provide some feedback so that I can evaluate it and continue to improve the contents.

1. Which Semantic Level(s) did you apply to your dataset (please delete as appropriate):

Semantic Level 1: Literal standardization

Semantic Level 2: Introducing URIs

Semantic Level 3: Introducing RDF

2. Did you feel the guide was clear in explaining the concepts behind the semantic technologies used?

Very unclear 1 2 3 4 5 Very clear

3. Did you feel the recipes in the guide were easy to follow?

Very difficult 1 2 3 4 5 Very easy

4. Did you feel the visualization techniques at the end of the guide were useful?

Useful 1 2 3 4 5 Not useful

5. Were there any sections that you think need changing/improving at all? Please give details:

6. Any other comments?

Glossary

Annotation (Linked Data)

One or more **RDF triples** describing part of the content or nature of a document on the Web.

API

An Application Programming Interface (API) is a set of software functions provided for use by another computer program. This is generally in contrast to human-computer interfaces (typically either a Command Line or **GUI**).

Canonical URI

A persistent and resolvable **URI** that represents a shared concept.

CIDOC CRM

The CIDOC Conceptual Reference Model (CRM) is a core **ontology** for Cultural Heritage. It is an event-based ontological schema by which the interactions between people, places, periods and things can be described in domain-neutral terms.

Co-reference Problem (URI)

The problem of establishing whether two different **URIs** refer to the same **Resource**.

Decomposition (URI)

Inferring the referent of a **URI** based on the string of characters from which it is composed. This is generally considered poor practice as the concept to which it refers can only be formally established by **dereferencing** it.

Dereference (URI)

Inferring the meaning of a **URI** by requesting information about it from the server which hosts it.

DISTINCT (SQL)

A **SQL** keyword that ensures that no duplicate results are returned from a query.

Domain

The field or topic to which a particular technology is applied.

EVE

Estimated Vessel Equivalent (EVE) is a method of approximately calculating the number of amphorae by dividing a metric for the entire assemblage (such as weight or percentage of rims) with those for a single vessel. See **MNI** for an alternative method of quantification.

ETL

Extract-Transform-Load (ETL) is a non-dynamic method of data conversion in which the data of interest is extracted from the original source, transformed as appropriate and then loaded into a separate data management system.

Formalization deferring

Technologies that defer the process of data formalization only until it is needed (such as spreadsheets) can often provide sufficient computational power for simpler tasks at low cost, but their efficiency tends to degrade as the task gets more complex. See also **Formalization front-loading**.

Formalization front-loading

Technologies that front-load the process of data formalization (such as RDF) are more efficient at processing complex data, but the initial overhead may be too great for some users to bear. See also **Formalization deferring**.

Fragment identifier (URI)

A **URI** that is defined in relation to the Web document in which it is expressed. Fragment identifiers generally begin with a # symbol.

Gartner Hype Cycle

A theory which states that the use of emerging technologies tends to follow a predictable adoption curve. An initial burst of enthusiasm leads to a *Peak of Inflated Expectations* with a high public profile. This causes a *Trough of Disillusionment* when these expectations fail to be met and the press cycle moves elsewhere. Slowly, a more realistic appraisal emerges along the *Slope of Enlightenment* before the technology becomes fully established on the *Plateau of Productivity*.

GUI (Graphical User Interface)

A visual interface by which a human can interact with a software system, frequently composed of ‘widgets’ such as buttons, menus or interactive images (icons).

HTML

Hypertext Markup Language (HTML) is a set of interpreted tags used for marking up text in Web documents. It is most notable for introducing the concept of links to other Internet documents on the World Wide Web.

HTTP

Hypertext Transfer Protocol (HTTP) is a series of Request-Response protocols for serving information (typically **HTML** documents) over the World Wide Web.

HTTP 303 Redirect

An **HTTP** Response used by **Semantic Web** applications to state that a requested Non-Information **Resource** could not be found and providing the **URI** of an Information Resource about it.

Inferencing (Knowledge Representation)

The automatic application of logical reasoning over a set of formally represented knowledge statements in order to derive new knowledge.

InfoViz

Any of wide range of methods for presenting data in a manner more amenable to human interpretation.

IDE

An Interactive Development Environment (IDE) is an integrated package of tools for software development.

JDOM

A **Java** software library for processing **XML** documents.

JSON

A concise digital data format frequently used for Web applications.

JUNG

A **Java** software library for network analysis and visualization.

Knowledge Representation

A field of Computer Science that represents knowledge in symbols so as to facilitate inferencing from those knowledge elements to new knowledge.

Label (RDF)

A human readable label associated with a (machine-readable) **URI**. Usually assigned with the `rdfs:label` property.

Linked Open Data

An interpretation of the **Semantic Web** that emphasises openness, decentralization and interconnectivity between independent resources on the Web. See also **MSKR**.

Literal (RDF)

A piece of information on the **Semantic Web** that is not a **URI** and thus not **dereferencable**. Typically a string of characters (such as a **label**), numeric value, date or other primitive data type.

Local Term

A dataset-specific term used to identify a concept.

Mapping

The process of connecting data held within a relational database to the concepts within an **ontology**.

Microdata (HTML)

An **HTML** specification for embedding data into Web documents. It performs a similar function to **RDFa** but does not use **triples** or require **URIs**.

Microprovision

The aggregation of many small datasets on the **Semantic Web**.

MNI

Minimum Number of Individuals (MNI) is a method of calculating the minimum number of objects in an archaeological assemblage based on a smaller subset of members (such as bases) in which the number associated with a single vessel is known. See **EVE** for an alternative method of quantification.

MSKR

Mixed-Source Knowledge Representation (MSKR) is a vision of the **Semantic Web** which emphasizes **inferencing** over a set of heterogenous source datasets, often (although not necessarily) in a closed data environment. See also **Linked Open Data**.

N3 (RDF)

Notation 3 (N3) is a simple notation format for **RDF**.

Namespace

The Web context of an identifier. A **URI** is frequently composed of a namespace and a **fragment identifier**.

Nonunique Naming Assumption

The assumption that multiple **URIs** may exist that refer to the same concept.

Ontology (Computer Science)

An “*explicit specification of a conceptualisation*” (Gruber, 1995). It provides a definition (i.e. the semantics) of entity types and their possible properties and relationships in a theoretical model. A domain ontology models the relationships between concepts used in a particular field (the **domain**).

OWL

A family of **RDF** knowledge representation languages used to describe and combine **ontologies**.

Python

A high level programming language frequently used for Web applications.

RDF

Resource Description Framework (RDF) is a technique for modelling semantically meaningful statements on the **Semantic Web** via Subject-Predicate-Object statements known as **triples**. These are composed of either three **URIs** or two **URIs** and a **Literal**. Combinations of these statements form a graph over which more complex meanings can be inferred. RDF is notation-independent and maybe expressed in a variety of notations including N3, **XML** or as a visual diagram.

RDFa

RDF in Attributes (RDFa) is an **RDF** notation that can be embedded in **HTML** Web documents.

RDFS

RDF Schema (RDFS) provides a small **RDF** meta-vocabulary for concepts, such as `rdfs:Class` and `rdfs:Property`, that permit the construction of simple **ontologies**.

RDF/XML

A notation format for **RDF** based on **XML**.

Reconciliation (Google Refine)

The process of mapping **local terms** to **canonical URIs** defined in Freebase.

Resource

The referent of a **URI**. A Non-Information Resource is typically an object or concept that exists independently of the Internet. An Information Resource is an object on the Internet, typically a Web document, that contains information about other Resources.

Semantic Web

The use of the Web technologies to encode information in machine-readable form. This is achieved by creating **RDF** statements (**triples**) out of **URIs**, each of which asserts an individual piece of information. The aggregation of these statements forms a graph. Graphs are constrained by the use of **ontologies** that define what types of **Resource** can exist (the nodes of the graph) and the types of relation that can exist between them (the arcs).

SQL

Structured Query Language (SQL) is a standardized family of languages for querying relational databases.

SKOS

Simple Knowledge Organization System (SKOS) is a **URI** meta-vocabulary for expressing Knowledge Organisation Systems (KOS), such as thesauri, taxonomies, glossaries and other classification schemes in **RDF**, so as to make them available on the **Semantic Web**.

SPARQL (RDF)

A formal language for querying **RDF** graphs.

Term Contextualization

Retrieving additional information about a **Resource** by **dereferencing** its **URI**.

Term Standardisation

The creation of a *de facto* common vocabulary among different datasets.

Terminus ante quem

‘Limit before which’ — The latest date at which an archaeological context could end its deposition.

Terminus post quem

‘Limit after which’ — The earliest date at which an archaeological context could begin its deposition.

Topic (Freebase)

An entry in the Freebase online encyclopaedia associated with its own **URI**.

Triple (RDF)

A single unit of information in **RDF**, composed of a **URI** Subject, URI Predicate and an Object which may be either a URI or a **Literal**.

Triplestore

A data repository for storing **RDF** triples.

Turtle (RDF)

A simple notation format for **RDF**.

Type (Freebase)

A category of **Topic** in Freebase. Topics may have multiple Types.

URI

A Uniform Resource Identifier (URI) is a globally unique identifier referring to either an Information **Resource** (on the Internet) or a Non-Information Resource (independent of the Internet). It is composed of a schema identifier (such as ‘http:’) and schema-dependent identifying string of characters. It may or may not be **dereferencable**, i.e. capable of returning a representation of the Resource over the Internet. URIs also form the atomic words of **RDF** triples.

VIEW (SQL)

A virtual table in a relation database.

Vocabulary Problem

The problem that arises when people use different lexical terms to refer to the same concepts in a document. Related to the **Co-reference Problem**.

VGI

Volunteered Geographic Information (VGI) is spatial data contributed by those without specialist geographic training, typically via the Web. This process is sometimes referred to as ‘neogeography’.

W3C

The World Wide Web Consortium (W3C) is the Web’s official standards body.

Web 2.0

A phrase promoted by O’Reilly Media that refers to a shift in Web publishing patterns that began in or around 2003. In particular it emphasises Web technologies that permit greater dynamic interaction, such as social media or user-generated content.

Web of Documents

A phrase used to denote the traditional pattern of Web usage which largely involves the hosting of ‘pages’ which are human readable but not easily interpretable by machine.

Web Science

The study of the social and technical phenomena which have given rise to, and been caused by, the World Wide Web.

XML

Extensible Markup Language (XML) is a general purpose specification for creating custom markup languages.

XSLT

Extensible Stylesheet Language — Transformations (XSLT) is a language for transforming **XML** documents into other formats (which may or may not be XML).

Bibliography

Matthew Addis, Paul Lewis, and Kirk Martinez. ‘ARTISTE image retrieval system puts European galleries in the picture’. *Cultivate Interactive*, (7), June 2002.

Matthew Addis, Kirk Martinez, Paul Lewis, James Stevenson, and Fabrizio Giorgini. ‘New Ways to Search, Navigate and Use Multimedia Museum Collections over the Web’. In *Museums and the Web 2005*, Vancouver, Canada, May 2005. Archives & Museum Informatics.

Ben Adida, Mark Birbeck, Shane McCarron, and Steven Pemberton. *RDFa in XHTML: Syntax and Processing*. Technical report, W3C, 2008.
<http://www.w3.org/TR/2008/REC-rdfa-syntax-20081014/>.

Harith Alani. *Spatial and Thematic Ontology in Cultural Heritage Information Systems*. PhD thesis, University of Glamorgan, July 2001.

Harith Alani, Wendy Hall, Kieron O’Hara, Nigel Shadbolt, Peter Chandler, and Martin Szomszor. ‘Building a Pragmatic Semantic Web’. *IEEE Intelligent Systems*, 23(3): 61–68, May 2008.

Dean Allemang and James Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Morgan Kaufmann, Burlington, MA, May 2008.

Chris Anderson and Michael Wolff. ‘The Web is Dead. Long Live the Internet’. *Wired*, September 2010.

Grigoris Antoniou and Frank Harmelen. *A Semantic Web Primer*. Cooperative Information Systems. MIT Press, 2004.

ArcheoInf. ‘Der Klick in die Antike’. *Rubin*, (Autumn), 2008.

Lora Aroyo, Natalia Stash, Yiwen Wang, Peter Gorgels, and Lloyd Rutledge. CHIP Demonstrator: Semantics-Driven Recommendations and Museum Tour Generation. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web*, pages 879–886. Springer, Berlin, 2007.

- Hernan Astudillo, Pablo Inostroza, and Claudia Andrea López Moncada. ‘Contexta/SR: A Multi-Institutional Semantic Integration Platform’. In *Museums and the Web 2008*, Montreal, Canada, 2008. Archives & Museum Informatics.
- Sören Auer, Christian Bizer, Jens Lehmann, Georgi Kobilarov, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web*, Lecture Notes in Computer Science, pages 722–735. Springer, Berlin, November 2007.
- Sören Auer, Sebastian Dietzold, Jens Lehmann, Sebastian Hellmann, and David Aumueller. ‘Triplify - Lightweight Linked Data Publication from Relational Databases’. In *Proceedings of WWW 2009*, Madrid, Spain, April 2009. Universidad Politécnica de Madrid.
- Charles Babbage. *Passages From the Life of a Philosopher*. Longman, Green, Longman, Roberts & Green, London, 1864.
- Elton Barker. ‘Welcome to PELAGIOS’. *PELAGIOS*, February 2011.
<http://pelagios-project.blogspot.com/2011/02/welcome-to-pelagios.html>.
- Jeffrey Beall. ‘The Semantic Web is Dead’. *Metadata*, May 2010.
<http://metadata.posterous.com/the-semantic-web-is-dead-0>.
- Zach Beauvais. ‘Editor’s notes’. *Nodalities*, (14), 2011.
- David Beckett. *RDF/XML Syntax Specification (Revised)*. Technical report, W3C, February 2004.
<http://www.w3.org/TR/REC-rdf-syntax/>.
- David Beckett and Tim Berners-Lee. *Turtle - Terse RDF Triple Language*. Technical report, W3C, March 2011.
<http://www.w3.org/TeamSubmission/turtle/>.
- Emily Bell. ‘Why Facebook’s Open Graph Idea Must be Taken Seriously’. *PDA - The Digital Content Blog*, April 2010.
<http://www.guardian.co.uk/media/pda/2010/apr/26/facebook-f8-emily-bell>.
- Tyler Bell and Harrison Eiteljorg. ‘Still More on XML — Finding a Common Ground’. *CSA Newsletter*, 18(3), 2006.
- Agiatis Benardou. ‘Classical Studies Facing Digital Research Infrastructures: From Practice to Requirements (MP3)’. *Digital Classicist Work in Progress Seminar*, July 2011.
<http://www.digitalclassicist.org/wip/wip2011-05ab.mp3>.

- Mike Bergman. ‘Advantages and Myths of RDF’. *AI3*, April 2009.
<http://www.mkbergman.com/483/advantages-and-myths-of-rdf/>.
- Mike Bergman. ‘Structured Web Gets Massive Boost’. *AI3*, June 2011.
<http://www.mkbergman.com/962/structured-web-gets-massive-boost/>.
- Tim Berners-Lee. *Information Management: A Proposal*. Technical report, CERN, March 1989.
<http://www.w3.org/History/1989/proposal.html>.
- Tim Berners-Lee. ‘Notation3 (N3) A readable RDF syntax’. *Design Issues*, 1998a.
<http://www.w3.org/DesignIssues/Notation3>.
- Tim Berners-Lee. ‘Semantic Web Road Map’. *Design Issues*, 1998b.
<http://www.w3.org/DesignIssues/Semantic.html>.
- Tim Berners-Lee. ‘What the Semantic Web is Not’. *Design Issues*, 1998c.
<http://www.w3.org/DesignIssues/RDFnot.html>.
- Tim Berners-Lee. ‘Web Architecture from 50,000 feet’. *Design Issues*, 1999.
<http://www.w3.org/DesignIssues/Architecture.html>.
- Tim Berners-Lee. ‘Linked Data’. *Design Issues*, 2006.
<http://www.w3.org/DesignIssues/LinkedData.html>.
- Tim Berners-Lee, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, and David Sheets. ‘Tabulator: Exploring and analysing linked data on the Semantic Web’. In *Proceedings of the 3rd International Semantic Web User Interaction Workshop (2006)*, Athens, GA, November 2006.
- Tim Berners-Lee, R. Fielding, U.C. Irvine, and L. Masinter. *RFC 2396: Uniform Resource Identifiers (URI): Generic Syntax*. Technical report, The Internet Society, 1998.
<http://tools.ietf.org/pdf/rfc2396.pdf>.
- Tim Berners-Lee and Mark Fischetti. *Weaving the Web : The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. Harper, San Francisco, 1999.
- Tim Berners-Lee, James Hendler, and Ora Lassila. ‘The Semantic Web’. *Scientific American*, May 2001.
- Ceri Binding. ‘Implementing Archaeological Time Periods Using CIDOC CRM and SKOS’. In Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors, *Proceedings of the 7th Extended Semantic Web Conference, Heraklion*, pages 1–15, Heraklion, 2010. Springer-Verlag.

- Christian Bizer and Richard Cyganiak. ‘D2R Server — Publishing Relational Databases on the Semantic Web’. In *Proceedings of the 5th International Semantic Web Conference*, Athens, Georgia, USA, November 2006.
- Christian Bizer, Tom Heath, and Tim Berners-Lee. ‘Linked Data - The Story So Far’. *International Journal on Semantic Web and Information Systems*, (Special Issue on Linked Data), 2009.
- Andrea Bonomi, Glauco Mantegari, Alessandro Mosca, Matteo Palmonari, and Giuseppe Vizzari. A System for Supporting Users of Cultural Resource Management Semantic Portals. In Roberto Basili and Maria Teresa Pazienza, editors, *AI*IA 2007: Artificial Intelligence and Human-Oriented Computing*, Lecture Notes in Computer Science, pages 757–764. Springer, Berlin, 2007.
- Chris Bourg. ‘Linked Data = Rationalized Serendipity’. *Feral Librarian*, June 2010.
<http://chrisbourg.wordpress.com/2010/06/24/linked-data-rationalized-serendipity/>.
- Greg Boutin. ‘Linked Data, a Brand with Big Problems and no Brand Management’. *Growth Times*, July 2009.
<http://www.growthtimes.com/2009/07/linked-data-a-brand-with-big-problems-and-no-brand-management/>.
- Richard Bradley. ‘Bridging the Two Cultures — Commercial Archaeology and the Study of Prehistoric Britain’. *The Antiquaries Journal*, 86, 2006.
- Dan Brickley and R.V. Guha. *RDF Vocabulary Description Language 1.0: RDF Schema*. Technical report, W3C, February 2004.
<http://www.w3.org/TR/rdf-schema/>.
- Frederick P. Brooks. ‘No Silver Bullet: Essence and Accidents of Software Engineering’. *Computer*, 20(4):10–19, April 1987.
- James Brusuelas. ‘Welcome to Ancient Lives’. *Ancient Lives*, July 2011.
<http://blogs.zooniverse.org/ancientlives/2011/07/26/welcome-to-ancient-lives/>.
- Vannevar Bush. ‘As We May Think’. *The Atlantic Monthly*, July 1945.
- Kate Byrne. *Populating the Semantic Web — Combining Text and Relational Databases as RDF Graphs*. PhD thesis, University of Edinburgh, 2009.
- Kate Byrne and Ewan Klein. ‘Automatic Extraction of Archaeological Events from Text’. In Bernard Frischer and Lisa Fischer, editors, *Proceedings of the 37th Computer Applications and Quantitative Methods in Archaeology Conference (CAA 2009)*, Williamsburg, USA, 2009.

- Paul Cripps, Anne Greenhalgh, Dave Fellows, Keith May, and David Robinson. *Ontological Modelling of the Work of the Centre for Archaeology*. Technical report, English Heritage, September 2004.
http://soton.academia.edu/PaulCripps/Papers/325372/Ontological_Modelling_of_the_Work_of_the_Centre_for_Archaeology.
- Nicholas Crofts. *Combining data sources – prototype applications developed for Geneva’s department of historical sites and monuments based on the CIDOC CRM*. Technical report, Direction du Patrimoine et des Sites, Geneva, 2004.
http://www.cidoc-crm.org/docs/st-petersburg-combining-data-sources_.doc.
- Nicholas Crofts, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff. *Definition of the CIDOC Conceptual Reference Model. Version 5*. Technical report, ICOM, January 2010.
http://www.cidoc-crm.org/docs/cidoc_crm_version.5.0.2.pdf.
- Nicholas Crofts, Martin Doerr, and Mika Nyman. *Call for Comments — Linked Open Data Recommendation for Museums*. Technical report, ICOM, March 2011.
http://www.cidoc-crm.org/URIs_and_Linked_Open_Data.html.
- Richard Cyganiak and Anja Jentzsch. ‘Linking Open Data Cloud Diagram 2007’. May 2007.
http://richard.cyganiak.de/2007/10/lod/lod-datasets_2007-05-01.png.
- Richard Cyganiak and Anja Jentzsch. ‘Linked Open Data Cloud Diagram 2011’. September 2011.
http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.pdf.
- Michael C. Daconta, Leo J. Obrst, and Kevin T. Smith. *The Semantic Web: A Guide to the Future of XML, Web Services and Knowledge Management*. John Wiley & Sons, 2003.
- Souripriya Das, Seema Sundara, and Richard Cyganiak. *R2RML: RDB to RDF Mapping Language*. Technical report, W3C, 2011.
<http://www.w3.org/TR/2011/WD-r2rml-20110920/>.
- Ian Davis. ‘The Linked Data Brand’. *Internet Alchemy*, July 2009.
<http://iandavis.com/blog/2009/07/the-linked-data-brand>.
- Leigh Dodds. ‘Announcing the Talis Connected Commons’. *Nodalities Blog*, March 2009.
<http://blogs.talis.com/nodalities/2009/03/announcing-the-talis-connected-commons.php>.
- Martin Doerr. ‘The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata’. *AI Magazine*, 24(3):75–92, 2003.

- Martin Doerr and Dolores Iorizzo. 'The Dream of a Global Knowledge Network — A New Approach'. *Journal on Computing and Cultural Heritage*, 1(1):1–23, 2008.
- Martin Doerr, Kurt Schaller, and Maria Theodoridou. 'Integration of Complementary Archaeological Sources'. In Franco Niccolucci and Sorin Hermon, editors, *Beyond the artefact — Digital Interpretation of the Past — Proceedings of CAA2004*, Prato, Italy, 2004. Archaeolingua.
- Heinrich Dressel. *Inscriptiones Urbis Romae Latinae. Instrumentum Domesticum*, volume 15 of *Corpus Inscriptionum Latinarum*. De Gruyter, Walter, Inc., Berlin, 1899.
- Tom Elliott. 'Digital Geography and Classics'. *Digital Humanities Quarterly*, 3(1), January 2009.
- English Heritage. 'English Heritage Timelines Thesaurus'. 2000.
<http://www.fish-forum.info/i.time.htm>.
- Orri Erling. *Requirements for Relational to RDF Mapping*. Technical report, W3C, September 2008.
<http://esw.w3.org/topic/Rdb2RdfXG/ReqForMappingBy0Erling>.
- Stuart Eve and Guy Hunt. 'ARK : A Development Framework for Archaeological Recording'. In Axel Poluschny, Karsten Lambers, and Irmela Herzog, editors, *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), 2007*, Berlin, 2008. Dr. Rudolf Habelt GmbH.
- Jackie Fenn. *When to Leap on the Hype Cycle*. Technical report, Gartner, 1995.
<http://www.citeulike.org/user/AdamHazdra/article/6662751>.
- FISH. 'FISH - the Forum on Information Standards in Heritage'. 2008.
<http://www.fish-forum.info/>.
- Martin Fowler, Kent Beck, John Brant, William Opdyke, and Don Roberts. *Refactoring: Improving the Design of Existing Code*. Addison-Wesley, 1999.
- Brian Fuchs, Leif Isaksen, and Amy C. Smith. 'The Virtual Lightbox for Museums and Archives: A Portlet Solution for Structured Data Reuse Across Distributed Visual Resources'. In *Museums and the Web 2005*, Victoria, 2005. Archives & Museum Informatics.
- George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. 'The Vocabulary Problem in Human-System Communication'. *Communications of the ACM*, 30(11):964–971, 1987.
- Peter Gärdenfors. 'How to Make the Semantic Web More Semantic'. *Formal Ontology in Information Systems: Proceedings of the third international conference (FOIS-2004)*,

- November 2004.
<http://yaxu.org/tmp/Gardenfors04.pdf>.
- Guntram Geser, Joost van Kasteren, Seamus Ross, and Michael Steemson, editors. *Towards a Semantic Web for Heritage Resources*. DigiCULT. Koninklijke Bibliotheek, The Hague, May 2003.
- Alejandro Giacometti. ‘Understanding Image-based Evidence’. *alejandrogiacometti.com*, November 2009.
<http://giacometti.tumblr.com/day/2009/11/17>.
- Sean Gillies. ‘What’s an Un-GIS?’. *Sean Gillies Blog*, 2011.
<http://sgillies.net/blog/1055/whats-an-un-gis/>.
- Hugh Glaser, Afraz Jaffri, and Ian Millard. ‘Managing Co-reference on the Semantic Web’. In *WWW2009 Workshop: Linked Data on the Web (LDOW2009)*, Madrid, April 2009.
- Shawn Graham and Giovanni Ruffini. Network Analysis and Greco-Roman Prosopography. In Katherine S. B. Keats-Rohan, editor, *Prosopography Approaches and Applications. A Handbook*, Occasional Publications of the Unit for Prosopographical Research, Linacre College, Oxford. University of Oxford, Oxford, 2007.
- Thomas R. Gruber. ‘A Translation Approach to Portable Ontology Specification’. *Knowledge Acquisition*, 5(2):199–220, 1993.
- Thomas R. Gruber. ‘Towards Principles for the Design of Ontologies Used for Knowledge Sharing’. *International Journal of Human-Computer Studies — Special issue: the role of formal ontology in the information technolog*, 43(5-6):907–928, 1995.
- Catherine Hardman. ‘OASIS: Sharing Information Across the Profession’. *Conservation Bulletin*, 51(Spring), April 2006.
- Stephen Harris and Nicholas Gibbins. ‘3store: Efficient Bulk RDF Storage’. In Raphael Volz, Stefan Decker, and Isabel Cruz, editors, *Proceedings 1st International Workshop on Practical and Scalable Semantic Web Systems*, Sanibel Island, Florida, USA, June 2003.
- Trevor M. Harris, L. Jesse Rouse, and Susan Bergeron. The Geospatial Semantic Web, Pareto GIS, and the Humanities. In David J. Bodenhammer, John Corrigan, and Trevor M. Harris, editors, *The Spatial Humanities*, pages 124–142. Indiana University Press, Indiana, 2010.
- Sebastian Heath. ‘Nomisma.org Annotations’. *PELAGIOS*, 2011.
<http://pelagios-project.blogspot.com/2011/05/nomismaorg-annotations.html>.

- Sebastian Heath and Billur Tekk  k. ‘Implementing RDFa in the Publication of Ceramic Data from Troy (Turkey)’. In Bernard Frischer and Lisa Fischer, editors, *Proceedings of the 37th Computer Applications and Quantitative Methods in Archaeology Conference (CAA 2009)*, Williamsburg, Virginia, March 2009.
- Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 2011.
- James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel Weitzner. ‘Web Science: An Interdisciplinary Approach to Understanding the Web’. *Communications of the ACM*, 51(7):60–69, 2008.
- Ian Hickson. *HTML Microdata*. Technical report, W3C, May 2011.
<http://www.w3.org/TR/microdata/>.
- Gerald Hiebel and Klaus Hanke. ‘Concept for an Ontology Based Web GIS Information System for HiMAT’. In Elizabeth Jerem, Ferenc Red, and Vajk Szever  nyi, editors, *On the Road to Reconstructing the Past; Proceedings of the 36th Annual Conference on Computer Applications and Quantitative Methods in Archaeology (CAA 2008)*, Budapest, Hungary, 2008.
- Peter Hinton. *Introduction to Standards and Guidance*. Technical report, Institute for Archaeologists, October 2008.
http://www.archaeologists.net/sites/default/files/node-files/ifa_standards_intro.pdf.
- Pascal Hitzler and Krzysztof Janowicz. ‘Semantic Web — Interoperability, Usability, Applicability’. *Semantic Web*, 1(1-2):1–2, 2010.
- House of Commons Science and Technology Committee. *Peer Review in Scientific Publications*. Technical report, House of Commons, London, July 2011.
<http://www.publications.parliament.uk/pa/cm201012/cmselect/cmsctech/856/85602.htm>.
- Isto Huvila. ‘Participatory Archive: Towards Decentralised Curation, Radical User Orientation, and Broader Contextualisation of Records Management’. *Archival Science*, 8(1):15–35, 2008.
- Eero Hyv  nen, Eetu M  kel  , Mirva Salminen, Arttu Valo, Kim Viljanen, Samppa Saarela, Miikka Junnila, and Suvi Kettula. ‘MuseumFinland — Finnish Museums on the Semantic Web’. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3):224–241, October 2005.
- Kingsley Uyi Idehen. ‘State of the Linked Data Web’. *Kingsley Idehen’s Blog Data Space*, October 2008.
<http://www.openlinksw.com/blog/kidehen@openlinksw.com/blog/?id=1455>.

- Leif Isaksen. 'Pleiades+'. *Google Ancient Places*, January 2011.
<http://googleancientplaces.wordpress.com/pleiades/>.
- Leif Isaksen, Elton Barker, Eric C. Kansa, and Kate Byrne. 'GAP: A NeoGeo Approach to Classical Resources'. *Leonardo Transactions*, May 2011.
- Mike Jackson. 'SPQR's Linked Data Online'. *SPQR Blog*, 2011.
<http://spqr.cerch.kcl.ac.uk/?p=187>.
- Stuart Jeffrey, Julian Richards, Fabio Ciravegna, Stewart Waller, Sam Chapman, and Ziqi Zhang. 'The Archaeotools Project: Faceted Classification and Natural Language Processing in an Archaeological Context'. *Philosophical Transactions of the Royal Society A.*, 367:2507–19, 2009.
- Ian Johnson. 'Reinventing the ECAI Clearinghouse — A Web 2.0 Approach to Research Data'. In Bernard Frischer and Lisa Fischer, editors, *Making History Interactive: 37th Annual International Conference on Computer Applications and Quantitative Methods in Archeology (CAA 2009)*, Williamsburg, Virginia, 2009.
- Ashish Karmacharya, Cristoph Cruz, Frank Boochs, and Franck Marzani. 'Spatial Rules through Spatial Rule Built-ins in SWRL'. *Journal of Global Research in Computer Science*, 1(2), 2010.
- Simon Keay. *Late Roman Amphorae in the Western Mediterranean. A Typology and Economic Study: The Catalan Evidence*. BAR International Series. BAR, Oxford, 1984.
- Simon Keay. 'Project Introduction'. *Roman Port Networks*, 2009.
<http://www.romanportnetworks.org/introduction/index.html>.
- Simon Keay and David F. Williams. 'Roman Amphorae: A Digital Resource'. 2005.
http://ads.ahds.ac.uk/catalogue/archive/amphora_ahrb.2005/.
- Alexander Korth. 'The Web of Data: Creating Machine-Accessible Information'. *ReadWriteWeb*, April 2009.
http://www.readwriteweb.com/archives/web_of_data_machine_accessible_information.php.
- Rasmus Krempel. 'Creating the Pleiades to Arachne Annotation'. *PELAGIOS*, July 2011.
<http://pelagios-project.blogspot.com/2011/07/creating-pleiades-to-arachne-annotation.html>.
- Amit Kumar. 'The Yahoo! Search Open Ecosystem'. *Yahoo! Search Blog*, March 2008.
<http://www.ysearchblog.com/2008/03/13/the-yahoo-search-open-ecosystem/>.

- Donna Kurtz, David Shotton, Florian Schroff, Yorick Wilks, Greg Parker, Graham Klyne, and Andrew Zisserman. 'CLAROS — Bringing Classical Art to a Global Public'. In *5th IEEE International Conference on e-Science*, Oxford, 2009.
- Ora Lassila and Ralph R. Swick. *Resource Description Framework (RDF) Model and Syntax Specification*. Technical report, W3C, February 1999.
<http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>.
- LGPN. 'LGPN Data — a Re-evaluation'. *LGPN and Information Technology*, 2011.
<http://www.lgpn.ox.ac.uk/online/documents/LGPNdataXML.htm>.
- Richard MacManus. 'Linked Data is Blooming: Why You Should Care'. *ReadWriteWeb*, May 2009a.
http://www.readwriteweb.com/archives/linked_data_is_blooming_why_you_should_care.php.
- Richard MacManus. 'Understanding the New Web Era: Web 3.0, Linked Data, Semantic Web'. *ReadWriteWeb*, May 2009b.
http://www.readwriteweb.com/archives/understanding_the_new_web_era_web_30_linked_data_s.php.
- Eetu Mäkelä and Eero Hyvönen. 'How to Deal with Massively Heterogeneous Cultural Heritage Data – Lessons Learned in CultureSampo'. *Semantic Web Journal*, (in press), 2012.
- Ashok Malhotra. *W3C RDB2RDF Incubator Group Report*. Technical report, W3C, January 2009.
<http://www.w3.org/2005/Incubator/rdb2rdf/XGR-rdb2rdf/>.
- Nick Malik. 'Having a High Bus Factor'. *Inside Architecture*, June 2005.
<http://blogs.msdn.com/b/nickmalik/archive/2005/06/28/highbusfactor.aspx>.
- Glauco Mantegari, Maurizio Cattani, Raffaele C. De Marinis, and Giuseppe Vizzari. 'Towards a Web-based Environment for Italian Prehistory and Protohistory'. In Jeffrey T. Clark and Emily M. Hagemester, editors, *Digital Discovery. Exploring New Frontiers in Human Heritage. CAA2006. Computer Applications and Quantitative Methods in Archaeology*, Fargo, USA, 2007. Archaeolingua.
- John Markoff. 'Start-Up Aims for Database to Automate Web Searching'. *New York Times*, March 2007.
<http://www.nytimes.com/2007/03/09/technology/09data.html>.
- Catherine C. Marshall and Frank M. Shipman III. 'Which Semantic Web?'. In Helen Ashman, Tim Brailsford, Les Carr, and Lynda Hardman, editors, *HYPERTEXT '03: Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, pages 57–66, Nottingham, UK, 2003. ACM.

- Kirk Martinez and Leif Isaksen. The Semantic Web Approach to Increasing Access to Cultural Heritage. In Chris Bailey and Hazel Gardiner, editors, *Revisualizing Visual Culture*, Digital Research in the Arts and Humanities, pages 29–44. Ashgate, London, 2010.
- Keith May, Ceri Binding, and Douglas Tudhope. ‘Following a STAR? Shedding More Light on Semantic Technologies for Archaeological Resources’. In Bernard Frischer and Lisa Fischer, editors, *Proceedings of the 37th Computer Applications and Quantitative Methods in Archaeology Conference (CAA 2009)*, Williamsburg, VA, 2009.
- Deborah L. McGuinness and Frank van Harmelan. *OWL Web Ontology Language: Overview*. Technical report, W3C, February 2004.
<http://www.w3.org/TR/owl-features/>.
- Jack Menzel. ‘Deeper Understanding with Metaweb’. *Official Google Blog*, July 2010.
<http://googleblog.blogspot.com/2010/07/deeper-understanding-with-metaweb.html>.
- Alistair Miles and Sean Bechhofer. *SKOS Simple Knowledge Organization System Reference*. Technical report, W3C, August 2009.
<http://www.w3.org/TR/skos-reference/>.
- Paul Miller. ‘XTech Day 3 — Rufus Pollock and Jo Walsh Talk About ‘Atomisation and Open Data’’. *Nodalities Blog*, May 2007.
http://blogs.talis.com/nodalities/2007/05/xtech_day_3_rufus_pollock_and_.php.
- Paul Miller. ‘Editor’s Notes’. *Nodalities*, (1), April 2008.
- Paul Miller. ‘Does Linked Data Need RDF ?’. *The Cloud of Data*, July 2009.
<http://cloudofdata.com/2009/07/does-linked-data-need-rdf/>.
- Paul Miller. ‘A Tale of Two Conferences’. *The Cloud of Data*, July 2010.
<http://cloudofdata.com/2010/07/a-tale-of-two-conferences/>.
- Knud Möller, Michael Hausenblas, Richard Cyganiak, Gunnar Grimnes, and Siegfried Handschuh. ‘Learning from Linked Open Data Usage: Patterns & Metrics’. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, Raleigh, NC, 2010. Web Science Trust.
- Tom Murphy and Nova Spivack. ‘Interview (Part 2): Nova Spivack On The Fragmentation Of The Semantic Web’. *New Tech Post*, October 2010.
<http://newtechpost.com/2010/08/10/interview-part-2-nova-spivack-on-the-fragmentation-of-the-semantic-web>.
- Theodor Nelson. *Computer Lib/Dream Machines*. Theodor Nelson, 1974.

- Kevin Newcomb. 'Yahoo Researcher Declares Semantic Web Dead'. *Search Engine Watch*, May 2007.
<http://searchenginewatch.com/article/2056255/Yahoo-Researcher-Declares-Semantic-Web-Dead>.
- Philipp Nussbaumer and Bernhard Haslhofer. CIDOC CRM in Action – Experiences and Challenges. In László Kovács, Norbert Fuhr, and Carlo Meghini, editors, *Research and Advanced Technologies for Digital Libraries*, Lecture Notes in Computer Science, pages 532–533. Springer, Berlin, 2007.
- Martin Oischinger, Bernhard Schiemann, and Guenther Goerz. *Short Documentation of the CIDOC CRM (4.2.4) Implementation in OWL-DL*. Technical report, University of Erlangen-Nurember, May 2008.
http://erlangen-crm.org/docs/documentation_crm_owl-dl_4.2.4.pdf.
- Dominic Oldman. 'British Museum Looks at Linked Data'. *Museums and Technology*, July 2009.
<http://www.oldman.me.uk/wordpress/2009/07/19/british-museum-goes-semantic/>.
- Tope Omitola, Nicholas Gibbins, and Shadbolt. 'Provenance in Linked Data Integration'. In *Linked Data in the Future Internet (LDFI-2010)*, Ghent, Belgium, 2010.
- Ross Parry, Nick Poole, and Jon Pratty. 'Semantic Dissonance: Do We Need (And Do We Understand) The Semantic Web?'. In *Museums and the Web 2008*, Montreal, Canada, April 2008. Archives & Museum Informatics.
- David P. S. Peacock and David F. Williams. *Amphorae and the Roman Economy: An Introductory Guide*. Longman Archaeology. Longman, London, 1986.
- Frank Price and Kate Geary. *Benchmarking Archaeological Salaries*. Technical Report April, Institute for Archaeologists, April 2008.
http://www.archaeologists.net/sites/default/files/node-files/ifa_salary_benchmarking.pdf.
- Eric Prud'hommeaux and Andy Seaborne. *SPARQL Query Language for RDF*. Technical report, W3C, January 2008.
<http://www.w3.org/TR/rdf-sparql-query/>.
- Jonathan Purday. 'Intellectual Property Issues and Europeana, Europe's Digital Library, Museum and Archive'. *Legal Information Management*, 10(3):174–180, September 2010.
- W. Boyd Rayward. 'Visions of Xanadu: Paul Otlet (1868-1944) and Hypertext'. *Journal of the American Society for Information Science*, 45:235–250, 1994.

- Resource Discovery Taskforce. ‘Approach’. *Discovery — A metadata ecology for UK education*, 2011.
<http://discovery.ac.uk/vision/approach/>.
- Julian D. Richards. ‘Archaeology, e-Publication and the Semantic Web’. *Antiquity*, 80 (310):970–979, 2006.
- Julian D. Richards. ‘From Anarchy to Good Practice: the Evolution of Standards in Archaeological Computing’. *Archeologia e Calcolatori*, (20):27–35, March 2009.
- Ric Roberts. ‘What is Linked Data?’. *Learn Linked Data*, March 2011.
<http://learnlinkeddata.com/articles/what-is-linked-data>.
- Bruce Robertson. ‘Exploring Historical RDF with Heml’. *Digital Humanities Quarterly*, 3(1), 2009.
- Simon Rogers. ‘What Kind of Data is on the Guardian’s Datastore?’. *Nodalities*, (7): 1–4, August 2009.
- Seamus Ross. Position Paper. In Guntram Geser, Joost van Kasteren, Seamus Ross, and Michael Steemson, editors, *Towards a Semantic Web for Heritage Resources*, DigiCULT. Koninklijke Bibliotheek, The Hague, 2003.
- Bertrand Russell. ‘On Denoting’. *Mind*, 14(56):479 – 493, 1905.
- Satya S. Sahoo, Wolfgang Halb, Sebastian Hellmann, Kingsley Uyi Idehen, Ted Thibodeau Jr., Sören Auer, Juan Sequeda, and Ahmed Ezzat. *A Survey of Current Approaches for Mapping of Relational Databases to RDF*. Technical report, W3C, January 2009.
http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf.
- m. c. Schraefel, Daniel A. Smith, Alisdair Owens, Alistair Russell, Craig Harris, and Max Wilson. ‘The Evolving mSpace Platform: Leveraging the Semantic Web on the Trail of the Memex’. In *HYPERTEXT ’05 Proceedings of the Sixteenth ACM conference on Hypertext and Hypermedia*, pages 174–183, Salzburg, Austria, 2005. ACM.
- Tom Scott. ‘Traversing the Giant Global Graph’. *Technology Forecast*, (Spring), 2009.
- Tom Scott and Michael Smethurst. ‘Building coherence at bbc.co.uk’. *Nodalities*, (5), January 2009.
- Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. ‘The Semantic Web Revisited’. *IEEE Intelligent Systems*, 21(3):96–101, 2006.
- Fred R. Shapiro. *The Yale Book of Quotations*. Yale University Press, New Haven, 2006.
- Ryan Shaw, Raphaël Troncy, and Lynda Hardman. *LODE: Linking Open Descriptions of Events*. Technical report, School of Information, U.C. Berkeley, August 2009.
<http://escholarship.org/uc/item/4pd6b5mh>.

- Frank M. Shipman and Catherine C. Marshall. ‘Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems’. *Computer-Supported Cooperative Work*, 8(4):333–352, 1999.
- Rainer Simon. ‘The PELAGIOS Graph Explorer: A First Look’. *PELAGIOS*, August 2011.
<http://pelagios-project.blogspot.com/2011/08/pelagios-graph-explorer-first-look.html>.
- Patrick A. S. Sinclair, Paul Lewis, Kirk Martinez, Matthew Addis, Adrian Pillinger, and Daniel Prideaux. ‘eCHASE: Exploiting Cultural Heritage using the Semantic Web’. In *4th International Semantic Web Conference, ISWC 2005*, Galway, Ireland, November 2005.
- Koven Smith and Paul Miller. ‘Koven Smith talks about the Semantic Web and Museums (Podcast)’. *Nodalities Blog*, January 2009.
<http://blogs.talis.com/nodalities/2009/01/koven-smith-talks-about-the-semantic-web-and-museums.php>.
- Paul Spence and Arianna Ciula. ‘Technical Introduction’. *Henry III Fine Rolls Project Commentary*, 2009.
<http://www.finerollshenry3.org.uk/content/commentary/technical.html>.
- Manu Sporny. ‘The False Choice of Schema.org’. *The Beautiful, Tormented Machine*, June 2011.
<http://manu.sporny.org/2011/false-choice/>.
- Michael Steemson. DigiCULT’s Expert 13 Tangle with the Semantic Web. In Guntram Geser, Joost van Kasteren, Seamus Ross, and Michael Steemson, editors, *Towards a Semantic Web for Heritage Resources*, DigiCULT. Koninklijke Bibliotheek, The Hague, 2003.
- Regine Stein, Jürgen Gottschewski, Regine Heuchert, Axel Ermert, Monika Hagedorn-Saupe, Hans-Jürgen Hansen, Carlos Saro, Regine Scheffel, and Gisela Schulte-Dornberg. *Das CIDOC Conceptual Reference Model: Eine Hilfe für den Datenaustausch?* Technical report, Staatliche Museen zu Berlin, Berlin, 2005.
http://www.museumbund.de/cms/fileadmin/fg_doku/publikationen/CIDOC_CRM-Datenaustausch.pdf.
- Michael Stutz. *The Linux Cookbook: Tips and Techniques for Everyday Use*. No Starch Press, 2nd edition, 2004.
- Ordnance Survey. *A Guide to Coordinate Systems in Great Britain*. Ordnance Survey, 2010.
- Richard J. A. Talbert. *Barrington Atlas of the Greek and Roman World*. Princeton University Press, Princeton, 2000.

- Jiao Tao. ‘Adding Integrity Constraints to the Semantic Web for Instance Data Evaluation’. In *9th International Semantic Web Conference (ISWC2010)*, Shanghai, 2010.
- Jeni Tennison. ‘Why Linked Data for data.gov.uk?’. *Jeni’s Musings*, January 2010.
<http://www.jenitennison.com/blog/node/140>.
- Douglas Tudhope. ‘Semantic Interoperability in Archaeological Collections’. In *SIEDL 2008: Semantic Interoperability in the European Digital Library*, pages 88–99, Tenerife, Spain, June 2008.
- Douglas Tudhope, Ceri Binding, Stuart Jeffrey, Keith May, and Andreas Vlachidis. ‘A STELLAR Role for Knowledge Organization Systems in Digital Archaeology’. *ASIST Bulletin*, 37(4):15–18, 2011a.
- Douglas Tudhope, Keith May, Ceri Binding, and Andreas Vlachidis. ‘Connecting Archaeological Data and Grey Literature via Semantic Cross Search’. *Internet Archaeology*, (30), July 2011b.
- Jacco van Ossenbruggen, Alia Amin, Lynda Hardman, Michiel Hildebrand, Mark van Assem, Borys Omelayenko, Guus Schreiber, Anna Tordai, Victor de Boer, Bob Wielinga, Jan Wielemaker, Marco de Niet, Jos Taekema, Marie-France van Orsouw, and Anne-miek Teesing. ‘Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques’. In *Museums and the Web 2007*, San Francisco, CA, 2007. Archives & Museum Informatics.
- W3C OWL Working Group. *OWL 2 Web Ontology Language Document Overview*. Technical report, W3C, October 2009.
<http://www.w3.org/TR/owl2-overview/>.
- Marc Wick. ‘Semantic Web : Concept vs Document’. *GeoNames Blog*, October 2006.
<http://geonames.wordpress.com/2006/10/21/semantic-web-concept-vs-document/>.
- Marc Wick. ‘Historical place names’. *GeoNames Blog*, April 2011.
<http://geonames.wordpress.com/2011/04/29/historical-place-names/>.
- Erik Wilde. ‘The Linked Data Police’. *dretblog*, November 2009.
<http://dret.typepad.com/dretblog/2009/11/the-linked-data-police.html>.
- Yahoo! ‘Yahoo! PlaceFinder Guide’. *Yahoo Developer Network*, June 2010.
<http://developer.yahoo.com/geo/placefinder/guide/index.html>.
- Valentin Zacharias. ‘Ban the Semantic Web Layer Cake!’. *Valentin’s Blog*, April 2007.
<http://www.valentinzacharias.de/blog/2007/04/ban-semantic-web-layer-cake.html>.