Seeing Triple
Archaeology, Field Drawing and the Semantic Web

Holly Ellen Wright

Submitted for the degree of PhD
The University of York
Department of Archaeology

September 2011

H E Wright

Seeing Triple: Archaeology, Field Drawing and the Semantic Web

Submitted for the degree of PhD

Abstract

This thesis explores the Semantic Web with relation to archaeology, and whether it is yet possible for non-specialist archaeologists to create, use and share their data using Semantic Web technologies and principles. It also considers whether spatial data derived from field drawings can be incorporated alongside textual data, to ensure a more complete archaeological record is represented on the Semantic Web. To determine if these two related questions can be answered, a practical application was undertaken, followed by a discussion of the results, and recommendations for future work.

Two archaeological datasets were chosen for the practical application. The first was an Anglian and Anglo-Scandinavian site in the Yorkshire Wolds located near Burrow House Farm, Cottam, excavated by the Department of Archaeology at the University of York. The second was from the Anglo-Scandinavian area of the multi-period Hungate site in the York city centre, excavated by the York Archaeological Trust. One of the primary tenets of the Semantic Web is interoperability of data, and the sites were chosen because they were related archaeologically, but differed technologically. Both datasets included field drawings from which data could be extracted, along with augmentory databases to enhance the demonstration. The data was carried through a complete workflow, from extraction, alignment to an ontology, translation into RDF, querying and visualisation within an RDF store, and through to publication as Linked Data.

This practical application was completed primarily using newly available generic tools, which required a minimal amount of specialist knowledge during most phases of the process. It demonstrated it is currently possible for non-specialist archaeologists to work with their data using Semantic Web technologies, including some data derived from field drawings. It showed how the Semantic Web allows archaeologists to use their data in new ways, and that it is a fruitful area for further work.

Contents

List of Figures 5
List of Accompanying Material 9
Acknowledgements 10
Author's Declaration 10

Chapter 1: Introduction 11

Chapter 2: The Semantic Web is like Archaeology: It's All About Context 27

2.1 Introduction 27
2.2 The development of the World Wide Web 32
2.2.1 Ted Nelson and the advent of Hypertext 36
2.2.2 Berners-Lee at CERN 37
2.2.3 The advent of the Hypertext Markup Language (HTML) 40
2.2.4 The advent of the World Wide Web Consortium (W3C) 44
2.3 The Semantic Web 46
2.4 XML 49
2.4.1 The structure of XML 50
2.4.2 XML as the foundation of the Semantic Web 52
2.5 RDF 54
2.5.1 RDF triples 55
2.5.2 RDF/XML 57
2.5.3 RDF and relational databases 59
2.5.4 RDF Schema 61
2.6 Ontology 62
2.6.1 Types of ontologies 64
2.6.2 An archaeological ontology 65
2.7 Logic, Proof and Trust 66
2.7.1 Logic 70
2.7.2 Proof 73
2.7.3 Trust 77
2.8 Beyond the 'layer cake' 80
2.8.1 The 'layer cake' 10 years on 81
2.8.2 The rise of Linked Data and SPARQL 83
2.9 Conclusion 85

Chapter 3: Archaeological Field Drawing: The Significance and Evolution of the Visual Archaeological Record 90

3.1 Introduction 90
3.2 A brief history of archaeological field drawing 95
3.2.1 Field drawing in the 17th and 18th centuries 95
3.2.2 Field drawing in the late 18th and 19th centuries 97
3.2.3 Field drawing in the 20th century 101
3.3 Modern field drawing 107
3.3.1 Drawing conventions 108
3.3.2 Plan drawing 113
3.3.3 Section drawing 115
3.4 Field drawing goes digital 117
3.4.1 Digital data capture 119
3.4.2 Retrospective conversion and 'heads-up' digitising 131
3.5 Conclusion 137

Chapter 4: A Practical Application of Archaeological Field Drawing Data Using Semantic Web Principles 141

4.1 Introduction 141
4.2 The sites 145
4.2.1 The excavations at Cottam 148
4.2.2 The excavation at Hungate 153
4.3 The data 158
4.3.1 The data From Cottam 161
4.3.2 The data from Hungate 168
4.4 The domain ontology 172
4.4.1 The CIDOC-CRM 174
4.4.2 The CRM-EH 176
4.4.3 Using STELLAR 179
4.4.4 Assigning URIs 182
4.5 Working with the data in RDF 184
4.5.1 Creating and populating the RDF store 187
4.5.2 Querying the data 192
4.5.3 Visualising the data with Gruff 200
4.5.4 Publishing the data with D2R and Pubby 207
4.6 Spatial approaches 213
4.7 Future work 227
4.8 Conclusion 241

Chapter 5: Conclusion 248

Appendices 262
Appendix A - List of files on CD 262
Appendix B - Thesis workflow 265
Appendix C - Selected Glossary of Acronyms 266

Bibliography 270

List of Figures

1	*Graphic of the Semantic Web technology stack as originally designed by Tim Berners-Lee in 2001 (Hendler 2009)*	32
2	*Screenshot of the HyperMedia browser/editor created by Tim Berners-Lee to read HTML. From the W3C website*	42
3	*Further screenshot of the original browser windows Tim Berners-Lee created to read HTML*	43
4	*Image of N3 triples linked together to form an RDF graph*	57
5	*The ontology spectrum (Daconta et al. 2003, 157)*	64
6	*Tim Berners-Lee image of the relationships of logic and proof to the lower layers of the Semantic Web*	75
7	*The Semantic Web 'layer cake' by Jim Hendler (ISWC 2009)*	82
8	*The Semantic Web for Dummies*	87
9	*John Aubrey's drawing of Avebury, c. 1675*	96
10	*Section drawing showing human jaws bones found in proximity with stone tools used for hunting mammoth (Boucher de Perthes 1864, 179)*	98
11	*Plan drawing by General Pitt-Rivers from his Excavations in Cranborne Chase (Piggott 1965, plate XXXIV)*	100
12	*Section drawing by Pitt-Rivers, showing the use of 'average sections' (Bowden 1991, 128)*	101
13	*Plan drawing by Heywood Sumner of Hambledon Hill (Piggott 1965, 173)*	102
14	*Plan drawing of earthworks by Robert Gurd (Goddard 2000, 8)*	104
15	*Plan drawing of the cairn at Cairnpapple Hill (Piggott 1947-8, 82)*	105
16	*Drawing by Mortimer Wheeler of a section across the cellar in Sacellum at Segontium (Adkins and Adkins 1989, 6)*	106
17	*Examples of common drawing conventions (Adkins and Adkins 1989, 76)*	109
18	*Examples of common symbols used to illustrate different materials found within archaeological units (Adkins and Adkins 1989, 74)*	110
19	*The hachure system for illustrating slope in two-dimensions, as commonly used in archaeology (Adkins and Adkins 1989, 67)*	112

20	*Examples of the hachure system (J. M. Hawker 2001, 18)*	112
21	*A TST in use at the site of Burdale in the Wolds of North Yorkshire, UK*	120
22	*Sharp PC-1500 Pocket Computer (Laroche 2010)*	122
23	*Left: Capture of vector-based spatial data within FieldNote on the Apple Newton. Right: Testing the MCFE system in the field (Ryan et al. 1999)*	123
24	*The FieldNote system downloaded from the Newton (Ryan et al. 1999)*	124
25	*Excavation carried out by the Landscape Research Centre in North Yorkshire, UK (Powlesland et al. 2009)*	125
26	*A handheld version of the FieldMap software (Ryan and Ghosh 2005, 20)*	126
27	*Left: Mike Rains holding the ruggedised tablet PC tested in the field at Silchester. Right: Working in the Silchester field office with a tablet PC*	127
28	*Left: The IADB on a handheld PC. Right: The IADB on a tablet PC*	127
29	*Use of the Apple iPad at the Pompeii Archaeological Research Project: Porta Stabia (Ellis and Wallrodt 2010)*	129
30	*3D laser scanning example from Laser Scanning for Archaeology: A Guide to Good Practice (Payne 2011)*	130
31	*Judith Winters with a large-format permatrace drawing*	132
32	*Upper: An original inked plan drawing. Lower: a vector drawing created using 'heads-up' digitising in AutoCAD (Hopkinson and Winters (2003)*	134
33	*A single context plan from an early version of the IADB (Lock 2003, 113)*	135
34	*A vector-based phase plan from the IADB*	136
35	*Graph showing the usefulness of elements within an archaeological publication (Jones et al. 2001)*	137
36	*Graph showing the usefulness of Internet resources to practitioners working in the Historic Environment (Brewer and Kilbride 2006)*	138
37	*Locations of the Cottam and Hungate excavations*	144
38	*Location of the excavation near Burrow House Farm, Cottam (Richards 2001a)*	150
39	*The COT95 excavation trench and contexts (Richards 2001a)*	152

40	*Left: The COT95 excavation trench from the north. Right: The COT95 excavation trench from the south (Richards 2001a)*	152
41	*Block H of the Hungate excavation facing southeast (Connelly 2007, 1)*	154
42	*Left: Location of the Hungate excavation within the York city centre. Right: Location of Block H, and Areas H1 and H2 (Connelly 2008, 1)*	155
43	*Left: Location of the deep trench in Area H2. Right: Plan drawing of the sunken floored building in Area H2 (Hunter-Mann 2009, 4-5)*	156
44	*The deep trench facing southwest, showing the Anglo-Scandinavian features in the lower half (Hunter-Mann 2009, 4)*	157
45	*Plan and section of the boat timbers found in Area H2 (Allen 2009, 10)*	157
46	*Entity relationship diagram for the Cottam context database (Richards 2001c)*	163
47	*Plan drawing from the COT95 excavation trench (Richards 2001c)*	165
48	*Plan drawing from the COT95 excavation trench, georeferenced and projected in ArcMap 9.3.1*	166
49	*Plan drawing from the COT95 excavation trench as created within the GIS*	167
50	*Location of the data from the deep trench from Hungate*	171
51	*The data from the deep trench from Hungate, as exported from the IADB*	171
52	*Screenshot of the CRM-EH in the Protegé ontology editor*	177
53	*Screenshot of the STAR research demonstrator*	178
54	*Screenshot of the STELLAR.Web browser-based application*	181
55	*The structure and relationships of the STELLAR templates*	181
56	*The underlying architecture of the AllegroGraph 4.3 RDF store*	188
57	*Screenshot of the superuser read/write access view of the Thesis repository within AllegroGraph*	189
58	*Screenshot of the repository namespaces*	191
59	*Screenshot of the 'Select by Context Type' query*	194
60	*Screenshot of the 'Construct by Context Type' query*	196
61	*Screenshot of the 'Construct by Stratigraphic Matrix' query*	198

62	*Screenshot of the 'View Sites on a Map' query*	199
63	*Screenshot of 'graph view' in Gruff*	202
64	*Screenshot of 'graph view' in Gruff showing stratigraphic relationships*	203
65	*Screenshot of the properties and values associated with a single context from Hungate, shown in Gruff 'table view'*	204
66	*Screenshot of 'query view' in Gruff, showing a list of descriptive notes associated with the contexts in Hungate and Cottam*	205
67	*Screenshot of graphical query view in Gruff*	206
68	*Screenshot of query view in Gruff, showing the tabular results and the SPARQL code created in graphical query view*	207
69	*The design of the full D2RQ Platform architecture*	209
70	*The design of the D2R Server architecture*	209
71	*Screenshots of the D2R Server interface*	210
72	*The design of the Pubby Server architecture*	211
73	*Screenshot of the start page of the Linked Data publication demonstrator created with Pubby*	212
74	*The relationship of the five GeoSPARQL components (Open Geospatial Consortium 2011b, 2)*	226
75	*Screenshot of the CLAROS interface*	234
76	*Screenshot of the Herodotus Encoded Space-Text-Imaging Archive (HESTIA) project's Herodotus' Narrative Timeline*	234
77	*Screenshot of the Pelagios: Enable Linked Ancient Geodata In Open Systems (PELAGIOS) project Graph Explorer*	235
78	*Screenshot of the Semantic Explorer for Archaeology (SEA) query interface (Solanki et al. 2011)*	236
79	*Tom Coates proclaiming 'Death to the Semantic Web' at dConstruct in September of 2010*	249

List of Accompanying Material

1 *Digital thesis copy and accompanying files on CD inside back cover.*

2 *A companion website for this thesis has been created at:*

www.diggingitall.co.uk

The website consists of:

1. An online copy of the thesis with hyperlinks

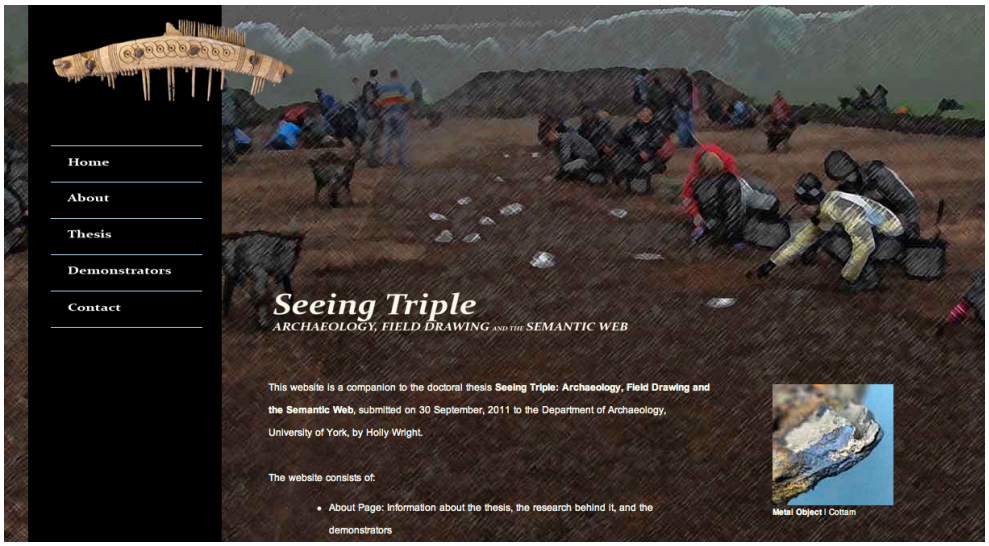
2. The online demonstrators described within the thesis text for:

AGWebView
D2R Server
Pubby Linked Data

As D2R Server and Pubby are both Linked Data servers, they are publicly accessible, but a (case sensitive) login and password is required for full access to AGWebView:

Login: [redacted]
Password: [redacted]

Within AGWebView, users can explore the Cottam and Hungate data sets in the Allegrograph RDF store, and by choosing the 'Thesis' repository, can access and execute saved versions of the SPARQL queries discussed in section 4.5 of the text.



Acknowledgements

So much has happened during the years throughout which this research was undertaken, it is difficult to take it all in, much less sum it up and provide adequate thanks to everyone who deserves it. It is equally difficult to separate the personal from the academic support I have received when it comes to the Department of Archaeology at York, but I particularly need to thank Claire McNamara, Alizon Holland, and Matthew Collins for stepping into the breach at key moments along the way, and to Julian Richards in his capacity as Head of Department, for making it all work. You don't know what a place is made of until things become difficult, and the Department of Archaeology at York is made of good people who understand what is really important. Long may it remain so.

I am very grateful to the members of my TAP committee, Penny Spikins and Steve Roskams, for their help, support and many good questions along the way. I am also very grateful to Mike Rains at the York Archaeological Trust for providing the Hungate data, and for his good advice and inspiring work. Thanks to my sister Carrie, and to Catherine Casson and Cath Neal for being my 'native speakers' and proofreading important chapters.

I give heartfelt thanks to my supervisor, Julian Richards, whose patience, wisdom, friendship and unswerving support has been the foundation of whatever I have accomplished here. Invariably, time spent with other PhD students turns to the topic of problems with their supervision, and I was always forced to sit in silence, lest the fact that I had only good things to say upset the others. I can't think of a bigger compliment. You are the best.

I want to thank all the amazing friends who have picked me up, dusted me off, gotten me drunk, made me laugh and generally put up with me when things were very bleak indeed: Chris and Emily Heagle, Andrea Vermeer, Tess Quadres, Chrissie White, Jennifer Melnick Bar-Nahum, Jody Morris, Brian Rahn, Hannah Koon, Beatrice Demarchi, Ashley Coutu, Enrico Cappellini, Igor Gutiérez Zugasti, Ivan Briz, Caroline Solazzo, John and Jillian Currey, Caner and Ceren Güney, Cath Neal, Catherine Casson, Daniel Löwenborg, Debra Hunt, Kathleen Maloney, Flora Gröning, Gail Falkingham, Eva Fairnell, Ben Gourley, Hannah Brown, Mark Edmonds, Kieron Niven, Lisa Pierce, Donna Page, Léan Ni Chleirigh, Matt McGuire, Matthew Collins, Phillipa Turner, Ramsey Rayyis, Stephen Street, Svetlana Matskevitch and Wendy Romer. I have been so lucky to have you all in my life. Particular thanks need to be said to Dawn Chapman and Michael Charno. Not only are you the best of friends and the best of people, but your advice, technical expertise and willingness to talk through a problem helped me through several critical places in this research. I cannot thank you both enough. Drinks continue to be on me.

I am so grateful to my mother and sister for your unwavering and unqualified support of whatever I attempt, in the many ways you both give it, but am equally saddened that two members of my family who deserve my thanks are no longer here to receive it. I dedicate this thesis to my father and grandmother, whose loss will always be a part of it, though I have to say, you really didn't have to go to such extremes to get out of proofreading. Much love to you all.

Author's Declaration

I declare this thesis is entirely my own work, and the responsibility for any errors is my own. Some of the ideas relating to the publication of field drawings, discussed in the conclusion of my third chapter, were previously published in a book chapter (Wright 2007), and Figure 31 was published in an article for *Internet Archaeology* (Wright 2006), but the rest of the practical work, discussion and conclusions are presented here for the first time. As stated in the text, the STELLAR.Preloader Java application included on the CD was created by Michael Charno.

Chapter One

Introduction

This thesis was begun in the Autumn of 2005, which is now more than half the life of the Semantic Web ago. The original research idea was to explore the Semantic Web with relation to archaeology, refined by a desire to determine whether the data from field drawing could also be represented, as the research carried out up to that point addressed primarily textual data (Karmacharya *et al.* 2009). This topic provided continuity with the author's previous work regarding the use of vector graphics, and specifically the Web standard for vector graphics called Scalable Vector Graphics (SVG) and field drawing in archaeology, and was therefore deemed a good fit. Returning to academia after two years as a full-time commercial field archaeologist, and despite knowledge of traditional Web design, standards and history, the Semantic Web was largely an unknown entity for the author. It is probably not unusual for the process of writing a thesis to feel more like an odyssey than simply a piece of research, but for an archaeologist to pursue this new and complex subject based in a different field, was to go on quite a journey.

The journey had further new twists, as the field drawing data forming the basis of this research was created using a type of archaeological field recording known as single context recording. Since the 1980s, single context recording has become a widely used recording system in the UK (Museum of London Archaeology Service 1994, 5), and differs fundamentally from the top planning or single-level planning tradition used in North America, where the author was trained. The choice of datasets was the author's first foray into the world of Anglo-

Scandinavian archaeology as well. This thesis being a product of study at the University of York, and York being a former Viking capital, it was natural that the search for good datasets to explore Semantic Web technologies would land squarely in the Anglo-Scandinavian north of England.

At a certain point, the challenge of the work itself was seen in an increasingly positive light, and a way to set useful parameters for the research questions. The desire to show a practical rather than purely theoretical result was strong as well, and resulted in a search for how archaeologists could best rise to the challenge and begin using the Semantic Web with archaeological data now. Thus, the research questions became:

- Just how difficult is it for an archaeologist to get started with the Semantic Web?
- Is it possible to use free and generic tools? If so, how much specialist knowledge is required?
- Archaeological field drawing is a fundamental part of field recording. How can the point, line and polygon data comprising the visual archaeological record be included alongside the textual record with regard to the Semantic Web?

So why is the Semantic Web such a challenge? For a start it is a new and complex way of thinking about not only the Web, but of the structure of data on the Web, and what can be done with it. Ask someone familiar with the Semantic Web to define what it is, and the answer can vary greatly. It is not that the answer is incorrect (though the complexity of the Semantic Web is such that there is often confusion about what it is, which does inadvertently lead to incorrect answers), but that it has many facets. People interested in the Semantic Web are often drawn in by a particular concept or feature that appeals to them, or as a new way of working which could be useful, rather than by the vision of the Semantic Web as a whole.

Those who are interested in making heterogeneous data interoperable through the use of controlled vocabularies and relationships are drawn to the Semantic Web because of its use of things called domain ontologies, which provide a way to map data from different sources to the same structure, and allow that data to be used together without losing its original meaning. The proponents of something called Linked Data are drawn to the idea that data held within Web pages and databases is 'siloed', and therefore not accessible because it is trapped within the structure of a Web page, or because it is held in an inaccessible database. The Linked Data movement therefore wishes to dismantle these data silos, and make raw data available for use and reuse. Those working with data mining, or the ability to gather meaningful, machine readable data from text-based information in an automated way will be interested in the way Natural Language Processing (NLP) can be used with Semantic Web concepts and technologies. Those interested in making their data richer by using automated reasoning to infer new relationships and information from existing data will also be interested in the Semantic Web and its use of things called reasoning engines. The aspects of the Semantic Web which include NLP and automated reasoning are both part of the area of Computer Science known as Artificial Intelligence (AI), which is why an archaeologist going to a library to look for Semantic Web books for the first time is bafflingly sent to the AI area of the Computer Science section, rather than the Web section. What strange world is this for an archaeologist to enter, where the answer to what the Semantic Web is depends largely on the area of interest of the person asked, and the library says you have entered the realm of AI?

Probably the simplest way to define the Semantic Web is to say it is a term coined by Tim Berners-Lee, the inventor of the World Wide Web, for his particular vision of the next evolution of the Web. All the concepts and technologies just mentioned form some part of that deep vision, and while Berners-Lee sees everything working together, some things have developed more quickly than others, some have yet to develop because they must be based on other things which are in the process of developing, some can be used all on their own while other things catch

up, and some may not be developed at all. Perhaps the key idea is that Berners-Lee sees the current Web (sometimes referred to as the 'classic Web') as now being hamstrung by its document-based format, and developed the ideas behind the Semantic Web to move from a 'Web of Documents' to a 'Web of Data'. The Web of Data is meant to consist of raw data freed from its proprietary database and document-based structures so that it may be used and combined in useful ways. It is therefore easy to see why archaeologists might be interested in this.

Though still small when compared with the use of other computing concepts and technologies in archaeology, use of the Semantic Web seems to be gaining considerable momentum. During its short history, the parallel development of something called the CIDOC-CRM has been central to much of the thinking and debate, and seems have been the source of some of the initial enthusiasm. The CIDOC-CRM is an ontology for the Cultural Heritage domain, meaning it is a formalised set of terms and relationships between those terms, specific to Cultural Heritage data. It is important to note however, that work on the CIDOC-CRM began well before the Semantic Web was a glimmer in the eye of Tim Berners-Lee. Growing out of the development of the CIDOC Relational Data Model (RDM), in 1996 the CIDOC Documentation Standards Working Group (DSWG) decided to change from a relational data model to an object-oriented model to better take into account the range of data structures in use within Cultural Heritage (CIDOC CRM 2010). This decision meant the resulting CIDOC-CRM, which became an official ISO standard in 2006, was well placed to become a basis for ontological modelling in the archaeological domain.

While not comprehensive, the importance of the Semantic Web to archaeology can be charted by looking at its presence within the discourse at the Computer Applications in Archaeology (CAA) conferences over the past several years. Workshops on how to use the CIDOC-CRM have been presented nearly every year at CAA since the 2002 meeting in Heraklion, Crete (CIDOC CRM 2010), and research incorporating the CRM began to appear shortly thereafter. In 2006,

at the meeting in Fargo, North Dakota, three papers were presented in the *Cultural Heritage Databases and Web-based Resources* session, which incorporated or discussed the CRM, along with a poster (Clark and Hagemeister 2007). In 2007, at the meeting in Berlin, five papers in the *Data Management* session showed incorporation of the CRM into their work (Posluschny *et al.* 2008). The 2008 meeting in Budapest featured the first session of papers dedicated to using the CIDOC-CRM. Entitled *CIDOC-CRM in Data Management and Data Sharing,* the session showed the growing consensus that the CRM could provide a solution for one of the key problems found across the discipline of archaeology; the need to find ways to align heterogeneous datasets and make them more interoperable (Jerem *et al.* 2008, 277). Within this session was also the first paper to explicitly set out that ontologies built on the CRM, used with Semantic Web technology, could provide this much needed interoperability. The paper *CIDOC-CRM in Data Management and Data Sharing* by Andrea D'Andrea presented how this could be approached, and explored some of the positive and negative outcomes (2008).

Important related work, though not strictly to do with either the Semantic Web or the CIDOC-CRM was also presented at the meeting in a session called *Alternative Ontologies and Approaches to Data Management and Data Sharing.* It included projects using Natural Language Processing, user generated content (also commonly referred to as Web 2.0), using standards to help provide virtual access to large amounts of archaeological material, and the usefulness of including standards as part of initial data capture (Jerem *et al.* 2008, 269). While the CIDOC-CRM was a natural place for exploration of Semantic Web technologies to start, it was not the only place. Within the *Data Management Systems for Archaeological Excavations* session, was the paper *The use of network analysis and the semantic web in archaeology: Current practice and future trends* by Leif Isaksen, Kirk Martinez and Graeme Earl, which explored the potential of combining network analysis with the Semantic Web (Jerem *et al.* 2008, 170).

By the 2009 meeting in Williamsburg, Virginia, use of the CIDOC-CRM could be found across a broad range of projects in different areas, and interest expressed in Budapest the year before led to the first dedicated Semantic Web session, chaired by Leif Isaksen and Tom Elliot, called *The Semantic Web: Second Generation Applications;* the reference to the 'second generation' meaning a call to come together, take the exploratory lessons learned, and begin building Semantic Web applications (Fischer *et al.* 2009, 220). The session included 11 papers, and ranged from a discussion of the meaning of semantics and the relationship between content and metadata, further work on ontologies and the CIDOC-CRM, interoperability projects, Natural Language Processing and other types of data extraction, and making information from existing Web pages more machine readable. It also included the first foray into Linked Data.

Based on the strong response at Williamsburg, a further Semantic Web session, which included 10 papers, was held at the 2010 meeting in Granada. The session was called *Semantic Infrastructures in Archaeology* and was chaired by Leif Isaksen and Keith May (Melero *et al.* 2010). The title reflected a more substantial, practical foothold for the Semantic Web in archaeology. The papers included an introduction to this complex subject for those who were interested but uninitiated, along with a discussion of the nature of Semantic Web technology and its compatibility with archaeological practice. The session also consisted of new methodological ideas for the inclusion of spatial, geographical, and temporal data, along with further work on projects presented the year before. There were also new ideas about data aggregation, integration and transformation from relational data sources.

The most recent meeting, held in Beijing in 2011, also had a strong Semantic Web session, with a total of 12 papers. The session was chaired by Leif Isaksen, Keith May and Monica Solanki and was titled, *Semantic Technologies,* the choice of which illustrates that practical implementation is now well underway in a variety of projects (Zhou 2011). The projects in the session explored things like

using Semantic Web technologies for the linking of information to data about places, Semantic Wikis, Semantic Web interactions with relational databases, how Semantic Web principles can be part of institutional data management for archaeological research, visualisation and querying interfaces for interoperable data sets, a survey of use of the Semantic Web within the archaeological domain, the creation of a markup language for ancient architecture, and most relevantly for this research, the development of generic Semantic Web tools for non-specialists. Once again, it included reporting on further work from projects presented the year before, showing that practical work was now ongoing.

There have been several projects presented at CAA over multiple years that seem to have formed the backbone of much of the work combining the Semantic Web with archaeology. These include ArcheoKM, ArcheoInf, The Port Networks Project, Tracing Networks/SEA, and STAR/STELLAR. This list is not meant to be comprehensive, but rather a way to explore some of the specific projects and their development.

ArcheoKM: The ArcheoKM project is a collaboration between technology researchers working at the Mainz University of Applied Sciences and the University of Burgundy. It consists of a platform for use with an industrial archaeology dataset, and uses Semantic Web technologies for data generated primarily by 3D laser scanning, specifically from the site of the Krupp steel production factory in Essen, Germany. The main goal of the project is to create semantic annotation for objects within the point clouds generated by the scanning, and to link those objects to related information within documents, GIS and images. ArcheoKM is based on a traditional relational database structure with a spatial extension, which will then be aligned to a CIDOC-CRM-based domain ontology with archaeological extensions created through the semantic annotation (Karmacharya *et al.* 2009; Karmacharya *et al.* 2008).

The ArcheoKM group is one of the first to address the importance of spatial data in archaeology with regard to the Semantic Web, and see linking vector objects to other data types as a way to make sure the spatial resource is included. Once the data is aligned to their ontology, they plan to create Web-based interfaces allowing the data to be viewed in 2D, 3D, and a GIS view along with a 'spatial facilitator', allowing true spatial operations and querying (Karmacharya *et al.* 2010b). As of the 2010 CAA meeting, the domain ontology was still in development, but an interface has been built to allow archaeologists to add objects with semantic annotation (Karmacharya *et al.* 2010a, 260-2). ArcheoKM takes the approach that semantic annotation will be sufficient to build meaningful relationships within their data, rather than mapping it to match a specific domain ontology. Their work is ongoing, but only being trialed on a single data set. The more archaeologists participate in the annotation, presumably the 'smarter' the relationships will become, but whether this data can ever be made reliably interoperable is another question. In any case, their work with spatial data is interesting and its development will be relevant for anyone working with archaeological field drawing.

ArcheoInf: The ArcheoInf project is under development by a consortium of German research centres, based primarily in Dortmund and Bochum. ArcheoInf is similar to ArcheoKM in that it is meant to allow better use and querying of field data (including spatial data), but their emphasis is on providing interoperability across heterogeneous datasets, including legacy databases. They do this by creating a layer on top of their data called a 'Mediator', which provides translation for their query interface (Lang 2009). ArcheoInf is also using the CIDOC-CRM as a basis for its domain ontology, but they have chosen to combine it with the Functional Requirements for Bibliographic Record (FRBR) ontology from the bibliographic domain. Beyond the CIDOC-CRM and FRBR, they have been mapping their data to a thesaurus specific to archaeology to define their terms and classifications. ArcheoInf is also concerned with preservation, forward migration and ensuring the original datasets remain autonomous (though still held within the

ArcheoInf system), with their Mediator pulling together and translating the data so that it may be queried simultaneously. This has proved much more difficult than expected, as archaeological data was found to be far more heterogeneous than data from other domains. As such, considerable pre-preprocessing was required, but the work has been successful and subsets of the data have been made interoperable and searchable within their user interface (Battenfeld *et al.* 2009, 281).

The focus of the project has been on textual data rather than spatial data thus far, and as of the CAA 2010 meeting the work has been focused on moving beyond the use of thesauri into the development of an extension for the CIDOC-CRM to create a formal ontology for the archaeological domain (Lang and Türk 2010). The approach of the ArcheoInf project contrasts greatly with ArcheoKM. By embracing the Semantic Web promise of interoperability between datasets, and the forward migration and stewardship of legacy datasets, ArcheoInf is tackling some difficult but important needs within archaeology, which is perhaps not as much the case with ArcheoKM. By working with only one dataset, much of which is derived from a technology out of reach by most archaeologists at this time (3D laser scanning), their approach seems less useful than ArcheoInf, but perhaps this may change in the future.

The Port Networks Project: The Port Networks Project (which is part of The Roman Ports in the Western Mediterranean Project) is a collaboration between researchers in the Department of Archaeology and the Department of Electronics and Computer Science at the University of Southampton. Within the Port Networks Project, a case study was undertaken to explore how data might be mapped to an ontology with an interface allowing archaeologists to make the mapping themselves, and then make their data available to others by publishing it as Linked Data. Specifically, the case study was designed to look at the distribution of marble and amphora, and how these networks might allow a better understanding of ancient trade routes. To do this, a domain ontology specifically

designed for the data was developed, along with a tool allowing users to map and work with the data called the Data Inspector Wizard. The Data Inspector Wizard uploads their data and helps them to match the column names within their data to the ontology. The Inspector also allows location mapping to either specific 'spaces' like coordinates associated with the data, or 'places' like the name of the place where an object was found. Once the mapping is complete, a configuration file is generated. The data is then ready to be run through a translation tool called a 'Data Importer' that generates the data aligned to the configuration file. This data can then be included within a Semantic Web database, or served as a static file. In both cases, the data is formatted so that it conforms to the tenets of Linked Data, and can therefore be used by others in Semantic Web applications (Isaksen *et al.* 2009b, 130-6; Isaksen *et al.* 2010).

This project looks at the Semantic Web from the opposite direction of the ArcheoKM and ArcheoInf projects. It intentionally steps away from the idea that to use Semantic Web concepts and technologies requires large, overarching structures and mappings with long development times before data can be made useful, and that most of the work must be done by specialists with input by archaeologists only at key points. The Port Networks Project takes a nimble approach, allowing the data to be quickly mapped only to an ontology specific to the exact data (which can then be mapped to more general ontologies like the CIDOC-CRM, if desired), and provides tools which allow archaeologists to transform their data themselves; keeping them in control of the mapping choices, and allowing them to use and share the data however they wish.

Tracing Networks/SEA: The Tracing Networks/SEA projects are a collaboration between researchers at the Universities of Leicester, Glasgow and Exeter. The *Tracing Networks: Craft Traditions in the Ancient Mediterranean and Beyond* project explores similar archaeological information to the Port Networks Project in that it looks at ancient networks within the Mediterranean, but focuses on crafts-people and craft traditions. Within this project is the Collaborative Working

Environment (CWE) and Ontology sub-project. The two primary goals of this project are to create a domain ontology for the data, and to create tools which researchers can use to map and interact with their data (Fiadeiro *et al.* 2009). As of the 2010 CAA meeting, work on the project included a sophisticated bespoke conversion and mapping tool allowing automated transformation of large amounts of data from traditional relational databases to Semantic Web format, aligned to a custom domain ontology, and based on the CIDOC-CRM (Hong and Solanki 2010, 271-4). As of the 2011 CAA meeting, the CWE project has been expanded into the Semantic Explorer for Archaeology (SEA) project, which provides a bespoke Web interface that allows querying, interaction, visualisation and statistical analysis across the seven Tracing Networks datasets. Of particular interest is the incremental query builder interface, which allows the data to be filtered and grouped, and then visualised using pie charts, bar charts or on a map. The query itself can also be visualised (Solanki *et al.* 2011).

CWE/SEA is similar to the Ports Networks Project in that their primary data source is made up of specific objects, rather than spatial data derived from fieldwork (although both use spatial information). It is similar to the ArcheoKM and ArcheoInf projects as it seeks to create an overarching solution for bringing Semantic Web functionality to archaeological data. It is more similar to ArcheoInf than ArcheoKM, in that it brings interoperability to disparate datasets held within a single framework. How interoperable any of the above projects are outside of their domain ontology remains to be seen, but most at least share coarse-grained interoperability with the CIDOC-CRM. Where the CWE/SEA project really shines is its user interface. It is still under development, and currently only for use with the Tracing Networks data, but there is great potential for it to be adapted for broader use, and the development of generic tools for use by non-specialists is of particular interest for this research.

STAR/STELLAR: The Semantic Technologies for Archaeological Resources (STAR) project is a collaboration between English Heritage and the University

of Glamorgan. The STAR project grew out of a data modelling project at the Centre for Archaeology (CfA) at English Heritage called Revelation, and had several objectives. The first was to use the CfA data modelling to create a domain ontology for archaeology that would be an extension of the CIDOC-CRM. This extension was called the CRM-EH, and once created the decision was made to further test its applicability and potential across a range of archaeological datasets beyond those generated by the CfA (May *et al.* 2008). The project then set out to test whether the domain ontology could be used to make gray literature more accessible to broader research using NLP, and whether several heterogeneous datasets could be mapped to the CRM-EH and made interoperable. Once the data was combined, an online demonstrator was created, allowing simultaneous querying across all the datasets (STAR 2011; May *et al.* 2010). The success of the STAR project, and the potential of the CRM-EH to act as a more universal domain ontology for archaeology, led to the Semantic Technologies Enhancing Links and Linked data for Archaeological Resources (STELLAR) project, which is a collaboration between English Heritage, the University of Glamorgan and the Archaeology Data Service.

The focus of STELLAR was to create generic tools to allow archaeologists to do two things that are quite difficult for non-specialists: to map their data to an appropriate domain ontology and to transform the data into a Semantic Web format. This was done by breaking the very complex CRM-EH into templates corresponding to different archaeological data types, such as finds or contexts, which were then developed into a desktop application and a simplified Web application consisting of an alignment and conversion tool. The tool creates data ready for sharing and use by other Semantic Web applications, and is interoperable at a coarse level with any other dataset mapped to the CIDOC-CRM, and fully interoperable with data from other sources also mapped to the CRM-EH (Tudhope *et al.* 2011a, 15-8; Tudhope *et al.* 2011b). The STAR/STELLAR projects are similar to ArchaeoInf, and CWE/SEA in that both projects are trying to increase interoperability and querying across heterogeneous datasets, and in

the development of a CIDOC-CRM based ontology that can potentially be used by others. It is also similar to the Ports Mapping Project in that it attempts to put simple mapping and conversion tools into the hands of non-specialists so that archaeologists can control and share their own data.

How non-specialist archaeologists might be able to use generic tools to engage with the Semantic Web being one of the main subjects of this thesis; STAR and STELLAR are two projects that have been used within this research. Without their development (or even if the timing had been different), this thesis would have been more theoretical, and much of the work carried out in the fourth chapter would not have been possible. Much work has been done using the Semantic Web within the domain of archaeology since the start of this thesis in 2005, and this work would have been quite different even if it had been completed one or two years earlier. That said, what follows reflects the length of the journey as well. The next chapter, entitled *The Semantic Web is like Archaeology: It's All About Context*, begins with a history of the development of the Semantic Web, both practically and conceptually. The chapter was written very early on in this research, and reflects the fact that the Semantic Web was more theoretical at the time. Editing the chapter to bring it sufficiently up to date for inclusion within a thesis submitted in 2011, was an instructive illustration of just how far the Semantic Web has come. Though its basis remains largely the same, revisiting the chapter the author felt a bit naive and nostalgic. Some updates were made to reflect the current state of the Semantic Web as the thesis neared completion, but it also remains a reflection of the moment in time when it was written. As such, it is hoped the chapter stands as interesting documentation of the research journey, reflecting how much the Semantic Web has developed during that time.

The third chapter *Archaeological Field Drawing: The Significance and Evolution of the Visual Archaeological Record* recounts the history and importance of the drawn field record, both perceptually and practically. In particular, it emphasises the importance of field drawing as carried out as part of an excavation, as

archaeology's destructive and unrepeatable nature results in the products of field recording becoming the primary resource. Thus, it argues that the visual record made during field recording must be included in any attempt to work with archaeological field data and the Semantic Web, and discusses the challenges of incorporating visual data along with the textual. To do this, it traces the development of archaeological field drawing, its current methodological practise, and the various means by which the drawn record becomes a digital record. It also explores the different types of field recording methods, including single context recording, and why it may be particularly useful within the structure of the Semantic Web.

The fourth chapter takes the historical, theoretical and methodological information recounted in the two previous chapters and combines them to form the basis of a practical demonstration. Titled *A Practical Application of Archaeological Field Drawing Data using Semantic Web Principles* the chapter recounts the attempt to take data derived from archaeological field drawings through a complete Semantic Web workflow successfully. It begins by explaining the two sites used in the demonstration from an archaeological standpoint, what their relationship is to each other, and why they were chosen. It then goes on to explain the sites from a technological standpoint. It recounts the origins of the data, how it was gathered, structured, stored and used, and the process by which the data was extracted to make it ready for use with the practical Semantic Web application.

Once the data has been extracted, the importance of aligning it to an appropriate ontology is then discussed. In this instance the ontology chosen is the CRM-EH. The process of aligning the data to the CRM-EH and then translating it into the format used by the Semantic Web is then discussed. The translation and alignment are done using STELLAR, which also makes it possible to assign the naming conventions necessary to make sure the data can be used as Linked Data if desired. The rules and conventions used to create Linked Data are then discussed, showing how the data was named and why. Once the data has been taken through

the workflow where it has been aligned to an ontology, with the correct naming structure and translated into the data format used within the Semantic Web, it is ready for use.

The next section describes the process of importing the data into a shared structure using generic software, where it can be accessed and used. Once the data is ready for use, the means by which it can be browsed and queried is discussed. The practical application was built with live online access so the reader may login and use the data for themselves. All of the queries discussed in the practical chapter are saved within the data interface, so readers may load and execute the queries if they care to do so. This includes queries that return data in table format, but there are also queries that return data where the relationships can be visualised. One of the saved queries returns the location of the archaeological sites, and can be shown on a map. The practical demonstration discusses other means of viewing the data using different kinds of generic software as well, along with the ability to create the queries visually, rather than having to write them by hand. An important part of the lifecycle of Semantic Web data is to make it available to others for re-use (Isaksen *et al.* 2011), so possibilities for the how the data might be published were also explored. This includes a 'quick and dirty' solution allowing data still in a relational database structure to be 'virtually' translated into the format used by the Semantic Web, and true publication using data that has been 'actually' translated.

Throughout the practical chapter, the challenges of working with data derived from archaeological field drawings are discussed, but much of the discussion is centred upon its current limitations. At this point the chapter is forced to become more theoretical, and the discussion turns to ways the data derived from digital field drawings made up of points, lines and polygons, which may also be geolocated, could be incorporated more fully. It includes information about new technologies on the horizon that may make this incorporation possible in the very near future, and how they might be used. The chapter concludes with a

discussion of how the work might be carried forward. It examines some the most recent work within the domain of archaeology using Semantic Web technologies, showing exemplars of how that future work might be done, and the areas within the workflow that could be expanded and improved along the way.

The speed at which the Semantic Web has grown and changed since this thesis was begun feels exponential, and its use within the archaeology domain has moved along with it apace. It is hoped that this thesis will make a modest contribution to the exploration of how non-specialist archaeologists can make use of the Semantic Web, both now and in the near future. It is also hoped it will open up more discussion about how the complete picture of the archaeological field record, which must include data from the visual record created through field drawing, can be included as well.

Chapter Two

The Semantic Web is Like Archaeology: It's All About Context

> *In an extreme view, the world can be seen as only connections, nothing else. We think of a dictionary as the repository of meaning, but it defines words only in terms of other words. I liked the idea that a piece of information is really defined only by what it's related to, and how it's related. There really is little else to meaning. The structure is everything. There are billions of neurons in our brains, but what are neurons? Just cells. The brain has no knowledge until connections are made between neurons. All that we know, all that we are, comes from the way our neurons are connected.*
>
> –Tim Berners-Lee (2000, 14)

It's all about context!

–The Mighty Boosh (2005)

2.1 Introduction

In May of 2001, Tim Berners-Lee and two co-authors published an article in Scientific American entitled *The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities* (Berners-Lee *et al.* 2001, 186). Publication of this article in a mainstream outlet by the creator of the World Wide Web marked a watershed moment. The well-formed vision already demonstrated by Berners-Lee, both in seeing the potential of hypertext as a tool to link the world in an almost alarmingly non-linear and un-hierarchical way, and his further invention of the World Wide Web Consortium (W3C) to help regulate its rampant expansion, shows the importance of his ideas and opinions in the Web's ongoing development. The publication of this article was the formal announcement of the next major chapter in the vision of Berners-

Lee. True to form, his concept simply leapt over much of the discussion and speculation about how the Web could and should change, to utilise tools and ideas already under development or in use through the work of the W3C.

The Semantic Web is Berners-Lee's particular vision for the future of the Web. Reading his article in 2001 would have felt almost as foreign as today's Web felt to us ten years before. He outlined concepts that would fundamentally change the way we live our lives, do our work and interact with virtually all areas of information. Over the last ten years, much of the world population has gone from viewing the Web as an interesting and (at times) frustrating novelty, to a rich source of interaction and information, and now for many of us it is an integral part of our lives. For the Semantic Web to have a similar level of impact as the advent of the Web itself would be impossible, but Berners-Lee is right to term it a revolution for the current Web. It is as much a revolution in technology as a revolution in how we *feel* about the Web. If the invention of the Web was about unprecedented access to people and ideas, it has also left us feeling very exposed. Ultimately, the Semantic Web will be about *trust.*

The 2001 article begins with a description of a brave new Web where our schedules, preferences, social networks and physical environment are defined and understood in a way that allows automated interaction. The Semantic Web is an environment where a huge amount of mundane and specialised information traditionally processed by humans becomes processable by machines. This is why trust is so important. Like any other technological revolution where humans choose to let go of their control in favour of progress and ease of life, the change is invariably accompanied by a feeling of unease, suspicion or downright fear. It is much easier to sell an individual technology, which solves a particular problem rather than a vision, but this is precisely what Berners-Lee sets out to do—again. He asserts that 'properly designed, the Semantic Web can assist the evolution of knowledge as a whole' (Berners-Lee *et al.* 2001, 43) which has quite serious implications for the role the Web will continue to play in our lives.

A vision can be difficult to explain, and this is certainly the case with The Semantic Web. As with most of Berners-Lee's ideas however, it is formed from his understanding of both an overarching need, and his ability to see the potential for how those needs might be practically met. This need is well explained by Allemang and Hendler (2008, 2):

> An information 'web' is an organic entity that grows from the interest and energy of the community that supports it. As such, it is a hodgepodge of different analyses, presentations, and summaries of any topic that suits the fancy of anyone with the energy to publish a webpage. Even as a hodgepodge, the Web is pretty useful. Anyone with the patience and savvy to dig through it can find support for just about any inquiry that interests them. But the Web often feels like it is 'a mile wide but an inch deep.' How can we build a more integrated, consistent, deep Web experience?

The idea of the Web feeling 'a mile wide and an inch deep' is resonant. In a sense, the vision behind the Semantic Web is for a Web that is just as deep as it is wide, but where you always know where you are, because you know your context. The term 'semantic' is most often used to describe the relationship of words and their meanings, but Berners-Lee has chosen the word 'semantic' to describe his vision for a deeper level of meaning in the Web. While his use of 'semantic' is not strictly the way the term is used in other disciplines, it is likely that the sense of the word to which Berners-Lee is referring has to do with the idea of 'semantics' being the study of the *relationships* between things. That is to say, a Semantic Web is the idea of a Web 'in context'.

Whether this is the correct interpretation of Berners-Lee's use of 'semantic' in Semantic Web is unclear. Most of his writing tends towards the practical, as in this early explanation:

> In communicating between people using the Web, computers and networks have as their job to enable the information space, and otherwise get out of the way. But doesn't it make sense to also bring computers more into the action to put their analytical power to work making sense of the vast content and human discourse of the Web?...The first step is putting data on the Web in a form that machines can naturally understand, or converting it to that form. This creates what I call a *Semantic Web*—a web of data that can be processed directly or indirectly by machines (Berners-Lee 2000, 191).

Berners-Lee expresses the Semantic Web in the form of layers, making up what is commonly known as a 'technology stack' (Antoniou and van Harmelen 2004, 18; Berners-Lee 2004, xii). It has gone through several iterations over the last 10 years, but the earliest version created by Tim Berners-Lee was constructed thus:

- The foundation is the existing Web technologies of Unicode and URIs. Unicode is the international standard for character sets, which allow all languages used by humans to be read and understood by computers. Uniform Resource Identifiers (URIs) allow computers to uniquely identify a resource on the Web.
- The lower layers are made up of the technologies necessary to create the Semantic Web, of which the eXtensible Markup Language (XML) is foundational. The next layer is the Resource Description Framework (RDF). XML is used to hold information, and RDF is used to create the relationships between that information.
- The next layer is ontology, where a subject-based classification system appropriate to the data (in this case, the subject would be something to do with Cultural Heritage),

is used to define the terminology used and the relationships established between those terms. An ontology may refer to a variety of classification systems, which are defined by their level of expressivity (Garshol 2004, 380-4). The expressivity of an ontology can be weak or strong, according to what is appropriate for the data and its application. To fully take advantage of the Semantic Web, the subject-based classification system must be at the level of strength known as a *conceptual model.*

- The upper part of the pyramid is made up of three layers, and these layers are more conceptual. The lowest is logic, which is the application of logical reasoning to the information. This requires setting up overarching rules that allow relationships to be checked so they may be properly understood (Passin 2004, 15). It also allows for the creation of relationships which use inference. Inference is a fundamental concept used in Artificial Intelligence (AI), to allow new relationships to be built in an automated fashion.
- The next layer is proof. This shows how the logic is applied and provides validation (Antoniou and van Harmelen 2004, 18).
- The uppermost layer is trust. Berners-Lee describes it as the 'Oh yeah?' button (Berners-Lee 2004, xviii), where you literally require the Semantic Web to prove the veracity of its information and inference choices before you are willing to accept it.

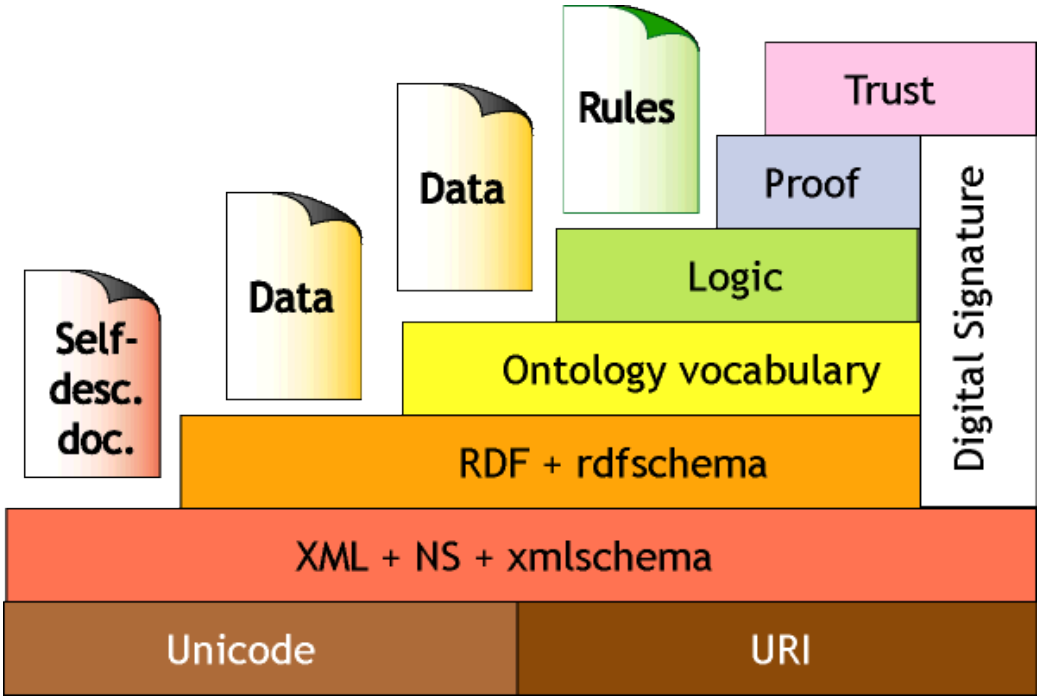


Figure 1: Graphic of the Semantic Web 'layer cake' or 'technology stack' as originally designed by Tim Berners-Lee in 2001. Reproduced from *My Take on the Semantic Web Layercake* by Jim Hendler (2009).

All of these layers are necessary to change the current Web, or Allemang and Hendler's 'mile wide, but an inch deep' Web, also known as the 'Web of Documents', into The Semantic Web, or the 'deeper Web', also known as the 'Web of Data' (Heath 2009). What this change entails is the subject of this chapter.

2.2 The development of the World Wide Web

In order to understand the Semantic Web as Tim Berners-Lee's current vision, it is important to understand the thinking that brought the Web into existence. Berners-Lee cites his upbringing as the son of mathematicians as being integral to his ideas. Both parents were involved in some of the earliest computer programming in the UK, and were part of the team who created the 'Mark I' computer at Manchester University in the 1950s. His father was already starting to see the limitations of the linear structure typical of programming at the time, and was exploring other ways of thinking about how to structure information. Berners-

Lee describes a pivotal moment when, coming home from school as a teenager, he found his father engrossed in books about how the human brain works. They proceeded to have a brief conversation about how much more useful computers could be if they were made to work more like human brains (Berners-Lee 2000, 3). He credits this incident with influencing his subsequent career path. He continued to think about this idea while at Oxford pursuing a degree in physics, and after graduating in 1976, he kept it in his mind while building his own computer and working in the telecommunications industry (Berners-Lee 2000, 4). Berners-Lee was hired by the European Particle Physics Laboratory (known as CERN) in 1980. As all Web scholars and enthusiasts know, Berners-Lee invented the World Wide Web during his time at CERN, in Switzerland. What is less well known is that it was very much work he did under the radar of his employers, and had little or nothing to do with what he was paid to do. The first Web-like program Berners-Lee wrote at CERN was called Enquire. He describes it as a humble project he wrote for his own use to help understand the non-linear relationships between the many people at CERN. CERN was made up of about 10,000 researchers, only 5,000 of which were ever in residence at any one time. Many of them moved back and forth between CERN and a home institution, as academics are wont to do. Trying to understand the structure between the transient researchers, their projects, and the associated equipment could simply not be expressed in a hierarchical or linear way. He observed, 'Informal discussions at CERN would be invariably accompanied by diagrams of circles and arrows scribbled on napkins and envelopes, because it was a natural way to show relationships between people and equipment. I wrote a four-page manual for Enquire that talked about circles and arrows, and how useful it was to use their equivalent in a computer program' (Berners-Lee 2000, 9). From this small beginning he began to realise that a specific concept had permeated the way he was thinking about programming at this point. He describes it thus:

> Suppose all the information stored on computers everywhere were linked, I thought. Suppose I could program my computer

> to create a space in which anything could be linked to anything. All the bits of information in every computer at CERN, and on the planet, would be available to me and to anyone else. There would be a single, global information space…Once a bit of information in that space was labelled with an address, I could tell my computer to get it. By being able to reference anything with equal ease, a computer could represent association between things that might seem unrelated but somehow did, in fact, share a relationship. A web of information would form (Berners-Lee 2000, 5).

Now that the Web has taken such a firm hold on the world, it is difficult to remember what a radical idea this was. The Internet had been in existence as ARPANET since the 1970s. Its purpose was to supply a decentralised network connecting strategically important sites across America in the event of a Cold War attack. About the time Berners-Lee had moved to CERN, the Internet was becoming accessible outside of the individuals in academia involved with Cold War science. With the advent of Usenet, anyone with access to a machine running UNIX and a phone line could exchange information. The timing of this development was fortuitous for Berners-Lee. It not only increased the number of machines that could be connected, it fundamentally changed the way the Internet was to develop.

> …Usenet did not enable you to log in as a user on remote machines, or do the other things which were possible on the ARPANET – it merely allowed you to exchange data and information with others in an exceedingly democratic and uncensored way. But in the end, that was to prove more important than anything else…A significant point in its development, however, came when Usenet reached the University of California at Berkeley, which was also an

> ARPANET node. The Berkeley people created a bridge between the two systems which essentially poured the exchanges in ARPANET discussion groups into Usenet News. This facility highlighted the differences between ARPA and Usenet, for ARPA discussion groups were essentially mailing lists in which the 'owner' of the list decided who was entitled to receive it, whereas Usenet was constructed on the opposite basis – that individuals decided which News groups they wished to subscribe to. In that sense, Usenet was distinctly more 'democratic' in spirit – which is why it was the model which eventually triumphed on the Internet (Naughton 2000, 180).

The advent of a publicly available, democratically minded Internet would be critical to Berners-Lee's development of the Web. It would be the communications backbone that would make applications like the Web possible. Upon returning to CERN in 1984 after a project back in England, Berners-Lee began to think more about how to organise information across an extremely diverse group of hardware and software options. Personal computers were still so new that there were no rules about adhering to a particular type of network or operating system. Berners-Lee observed as others tried to come up with solutions to manage the unwieldy information and connections across the organisation, and saw each of them fail. Not only were the solutions frequently proprietary, they simply did not understand the ethos of the work environment. He states how he 'saw one protagonist after the other shot down in flames by indignant researchers because the developers were forcing them to reorganise their work to fit the system. I would have to create a system with common rules that would be acceptable to everyone. This meant as close as possible to no rules at all' (Berners-Lee 2000, 17).

Rather than cursing the non-conformist researchers, or being threatened by their chaotic approach, Berners-Lee began to think about CERN as a microcosm of the world. He felt this was precisely the sort of situation that would lend itself

to the creation of a communication concept that might be acceptable across any shared environment, no matter how diverse. His first attempt (called Enquire) was written in Pascal, but to implement his ideas further he began to think about other possibilities, and quickly settled on hypertext as the most promising candidate (Berners-Lee 2000, 18).

2.2.1 Ted Nelson and the advent of Hypertext

Ted Nelson created the concept of hypertext, while a graduate student at Harvard in the 1960s. His desire to form a non-linear way of organising information also started in childhood, and similarly began as a way to write that was closer to the way humans think. He felt it was unnatural to take information held in our non-linear minds, organise it into a written format which must be linear, only to have it dissembled into pieces again by the reader (Whitehead 1996), and sought to create a 'whole system of literature to replace all systems of writing and publication' (Nelson 2009, 68). In an opinion piece for *Wired* magazine, Gary Wolf describes Nelson's ambitious outlet for hypertext, which he called Xanadu:

> Xanadu was meant to be a universal library, a worldwide hypertext publishing tool, a system to resolve copyright disputes, and a meritocratic forum for discussion and debate. By putting all information within reach of all people, Xanadu was meant to eliminate scientific ignorance and cure political misunderstandings. And, on the very hackerish assumption that global catastrophes are caused by ignorance, stupidity, and communication failures, Xanadu was supposed to save the world (Wolf 1995).

Xanadu's controversy lies in its 30-year development without ever coming into popular use, but for global vision, Berners-Lee was in good company by choosing hypertext. Nelson disagrees with Berners-Lee's interpretation of how hypertext should be used however, and rejects the structure of the Web as being fundamentally flawed:

> Xanadu and the World Wide Web are totally different and incompatible. The Web has one-way links and a fixed rectangular visualization based on the strictly-enforced rules of the browser...Xanadu alumni consider the Web illicit and broken, exactly what they were trying to prevent—for having only one-way links, for conflating a document with a place, for locking it to one view, for having no means of visible connection to points within a document, for imposing hierarchy in a variety of ways (Nelson 2009, 70).

While one became wildly successful and the other not (although the computing industry is notorious for promoting inferior technology at the expense of true innovation), both were trying to use the medium of hypertext as a great human leveller with the potential to change the world. Nelson himself ranges from ambivalence to downright displeasure with the ascendancy of the Web to the detriment of Xanadu (Kahney 1999; Silberman 1998), but hypertext has been fundamental to both. It is interesting to note that issues of current importance to Web development, were part of Xanadu early on. For example, the controversies surrounding the move by corporations to use government censorship of Web content as a way of protecting their copywritten material, might not have arisen if more attention had been paid to Nelson's ideas about copyright disputes. As another example, stabilised content addresses (or 'permadresses') have always been an important part of Xanadu (Nelson 2009, 68-9). As will be explored later in this thesis, the idea of content with stabilised or 'persistent' addresses is becoming more important, specifically for the Semantic Web.

2.2.2 Berners-Lee at CERN

Weaving the Web is Berners-Lee's (2000) personal chronicle of the twists and turns he made while creating HTML and the World Wide Web. The event that started his journey from a very talented computer scientist, to someone questing for a way to bring people and knowledge together in an entirely new way,

occurred in 1980. Two of the particle accelerator control systems at CERN were in need of replacement, and the work was seriously behind schedule. A phone call from a fellow colleague in England suggesting they both apply to work on the project, led to his hiring shortly thereafter. Berners-Lee describes CERN as being 'like a huge, chaotic factory' and the control room as 'an electrical engineer's paradise' (Berners-Lee 2000, 9). Every bit of space was full of custom-built electronics, equipment, and machines that go ping.

Computing was still done through various central facilities throughout the buildings. Berners-Lee became part of the team responsible for replacing the centralised systems with the terminal-based systems we use today. This transition to decentralisation was fundamental to the way Berners-Lee began to think about information sharing. While he mourned the loss of the computer as 'a sort of shrine to which scientists and engineers made pilgrimage' (Berners-Lee 2000, 9), he was equally compelled by the dynamic style of interaction at CERN.

> The big challenge for contract programmers was to try to understand the systems, both human and computer, that ran this fantastic playground. Much of the crucial information existed only in people's heads. We learned the most in conversations at coffee at tables strategically placed at the intersection of two corridors. I would be introduced to people plucked out of the flow of unknown faces, and I would have to remember who they were and which piece of equipment or software they had designed. (Berners-Lee 2000, 10).

For the sake of his own sanity, when he had time away from his primary task working on the Proton Synchrotron Booster, he began creating a program to help keep track of the people working at CERN, the computing equipment they used, the programs they ran and to which project they belonged. The non-linear nature of the work as undertaken at CERN inspired him to create a non-hierarchical way

of organising information. The result was the program he named Enquire.

> In Enquire, I could type in a page of information about a person, a device or a program. Each page was a 'node' in the program, a little like an index card. The only way to create a new node was to make a link from an old node…I liked Enquire and made good use of it because it stored information without using structures like matrices or trees. The human mind uses these organising structures all the time, but can also break out of them and make intuitive leaps across the boundaries – those coveted random associations. Once I discovered such connections, Enquire could at least store them (Berners-Lee 2000, 11).

Enquire was never taken up by anyone else at CERN, and was eventually lost. After a stint back working in the UK, as previouly noted, Berners-Lee returned to CERN with more enthusiasm to explore non-linear ways of connecting information. The next step was a program called Tangle. Tangle had a new twist which made it feel more Weblike:

> Computers store information as sequences of characters, so meaning for them is certainly in the connections among the characters. In Tangle, if a certain sequence of characters recurred, it would create a node that represented the sequence. Whenever the same sequence occurred again, instead of repeating it, Tangle just put a reference to the original node. As more phrases were stored as nodes, and more pointers pointed to them, a series of connections formed (Berners-Lee 2000, 15).

While Tangle predates the Web, and certainly the Semantic Web, this shows the concepts behind organising data using nodes, and then inferring new relationships between those nodes were already part of Berners-Lee's thinking.

2.2.3 The advent of the Hypertext Markup Language (HTML)

The concept of hypertext is simply the idea that one resource can be linked to any other resource in a non-linear way. This, coupled with the notion that resources must be available to any user, regardless of their choice of software or operating system (Castro 2007, 14), guided the way Berners-Lee chose to develop HTML. Armed with a new NeXT desktop computer (NeXT was developed by Steve Jobs after he left Apple in 1985, and while too ahead of its time for commercial success, much of its innovation was incorporated when Jobs returned to save Apple in 1997), with the thinly veiled assignment of assessing it as a potential development environment for CERN, Berners-Lee set out to find a hypertext editor that could be adapted to send information and instructions over the Internet. To his surprise, he found other researchers involved in hypertext development were not interested in, or able to see the potential of, what would become the World Wide Web. By 1990, Berners-Lee began to realise he would have to create a new way to use hypertext on his own (Berners-Lee 2000, 30).

In order to give the concept of hypertext a tangible structure, Berners-Lee turned to the Standard Generalized Markup Language (SGML). SGML is a standardised version of the Generalized Markup Language (GML) developed by a team led by Charles Goldfarb at IBM in the1960s. SGML was part of the pioneering work into ‘generic coding’, which would provide the essential separation between content and formatting, and make possible everything created by Berners-Lee that followed:

> Historically, electronic manuscripts contained control codes or macros that caused the document to be formatted in a particular way (‘specific coding’). In contrast, generic coding, which began in the late 1960s, uses descriptive tags (for example, ‘heading’, rather than ‘format-17’). Many credit the start of the generic coding movement to a presentation made by

> William Tunnicliffe, chairman of the Graphic Communications Association (GCA) Composition Committee, during a meeting at the Canadian Government Printing Office in September 1967: his topic – the separation of the information content of documents from their format (Goldfarb 1996).

HTML is SGML made specifically to control hypertext by separating the content of a document (or what would become a Website), from its formatting, and does so using code which is readable and understandable by humans. Berners-Lee began working with SGML in October of 1990, and within six weeks he had created a browser/editor, which he named WorldWideWeb. Two weeks later he had written the Hypertext Markup Language (HTML) which was the essential formatting tool allowing hypertext navigation between different areas of information (Berners-Lee 2000, 31). Figure 2 is a screenshot of Berners-Lee's NeXT desktop, showing the first browser/editor and some of the various interactions possible with hypertext. For example, the underlined text we recognise as a 'link' to another piece of information, which he created with a graphical shortcut in the form of a floating 'Links' palette, or with a 'styles' palette for designating font formats like 'bold' or for displaying a font in a particular size.

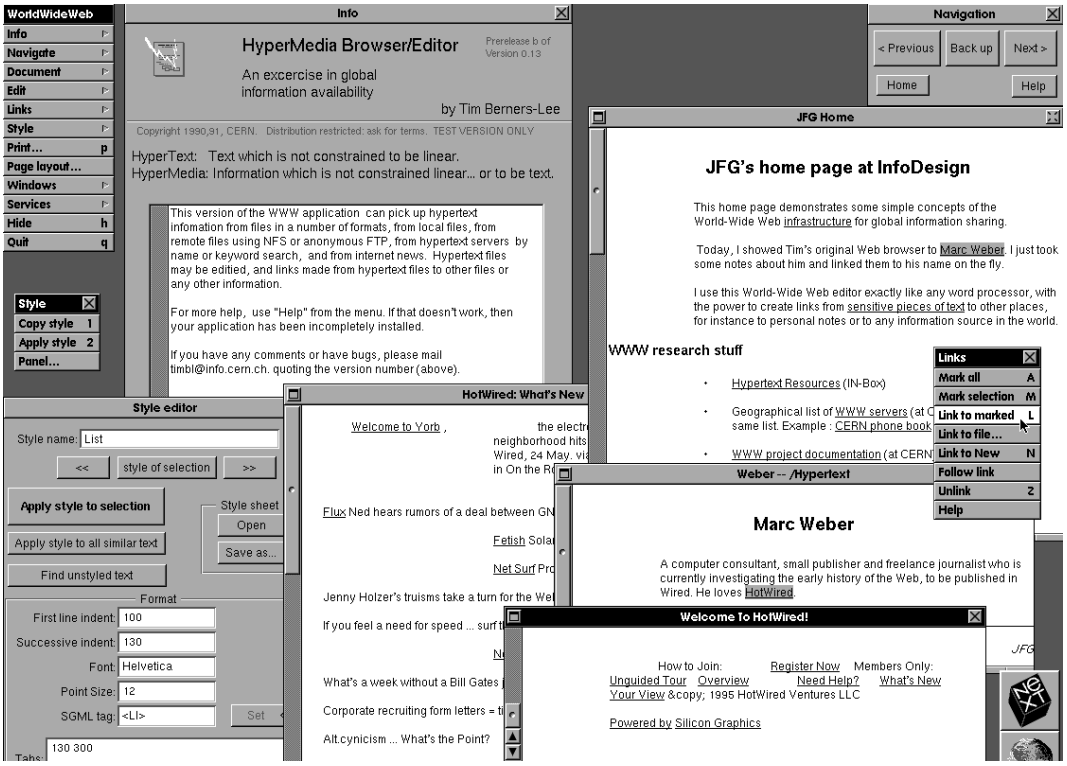


Figure 2: Screenshot of the HyperMedia browser/editor created by Tim Berners-Lee to read HTML. From the W3C website. http://www.w3.org/MarkUp/tims_editor.

This image is very interesting, as it shows Berners-Lee's use of graphical shortcuts to create his HTML formatting. When HTML became a formalised language, early Web designers were always taught to hand-code every part of a Website in a plain text-editor, so that the code was clean and fast. What You See Is What You Get (WYSIWYG) graphical HTML editors like Macromedia's Dreamweaver (and the dreaded Microsoft FrontPage) were initially seen as HTML editors for people who were too lazy to learn HTML. As websites became increasingly complex, WYSIWYG tools have become a necessity, but clearly Berners-Lee was using them from the start.

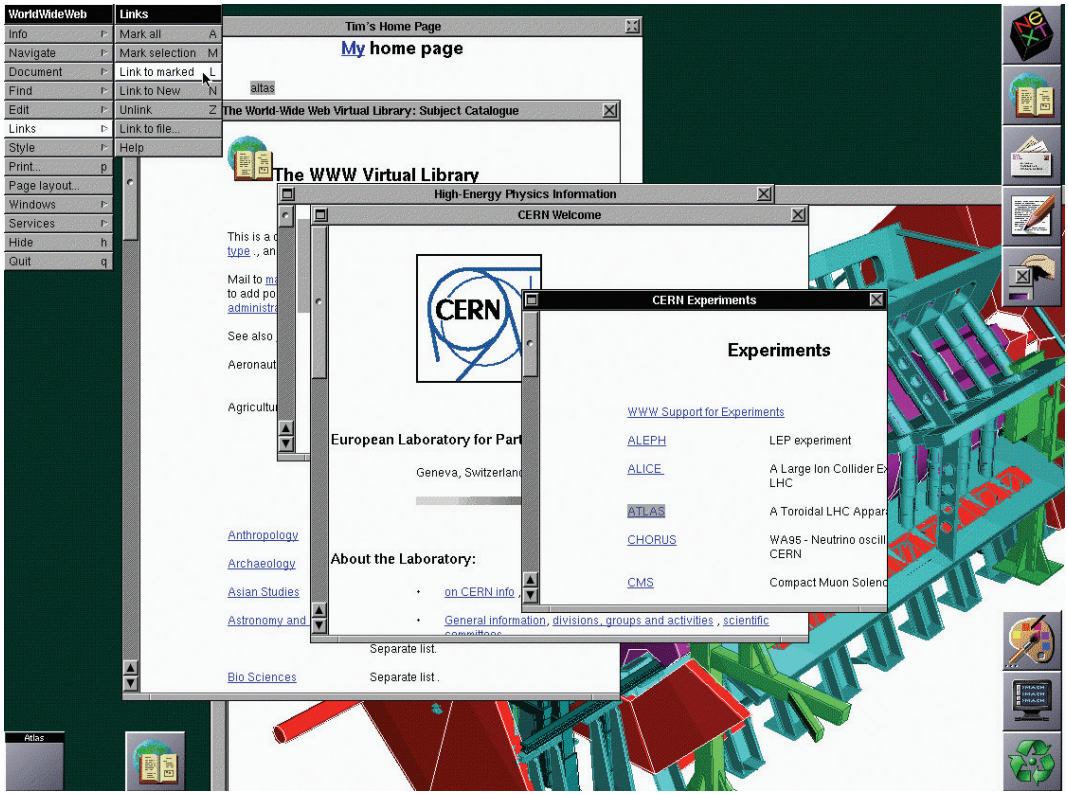


Figure 3: Further screenshot of the original browser windows Tim Berners-Lee created to read HTML. From the W3C website. http://www.w3.org/People/Berners-Lee/WorldWideWeb.html.

The impact of HTML as a markup language for Web content, and the development of the first graphical Web browsers like Mosaic, is well documented. As the Web began to take off, and Website design became more sophisticated, HTML began to show its limitations. It became apparent that while HTML allowed content to be separated from the formatting and structure of information, a further separation was necessary. HTML was relatively simple to use, but that simplicity limited its power (Castro 2007, 18). The solution was Cascading Style Sheets (CSS), which took over the formatting of a Website, and left HTML with the sole task of handling structure. CSS allowed single style choices to apply over an entire site, making universal changes far easier, and allowing Web designers to create much more powerful, sophisticated, scalable and intuitive sites.

2.2.4 The advent of the World Wide Web Consortium (W3C)

With the exception of Berners-Lee, by the 1990s the growth and success of the Web was beyond anyone's wildest dreams. It was a heady time, a bit like a high-tech version of the boom and bust of the California Gold Rush in the mid-nineteenth century. While both events attracted the hardest workers and the biggest swindlers, (and gave a very few dizzying wealth and power, while leaving most with less than they had when they arrived), in the end what was created for the benefit of all was a bit of law and order. As the chaos of the Gold Rush gave California its first laws and government, the feeding frenzy fuelling the development of the Web was also tearing it apart, and Berners-Lee became hugely concerned for its future.

To defuse the proprietary battles known as the 'browser wars', in 1994 Berners-Lee started an organisation to create and promote Web standards, which he named the World Wide Web Consortium (W3C). The two main combatants in the 'browser wars' were Internet Explorer, created by Microsoft, and Navigator, created by Netscape. Netscape began the battle by creating proprietary features that made their browser very popular. Microsoft followed suit in the same way software developers have typically done, by developing their own set of features they hoped users would find more appealing, and lure them away from Netscape. As the two main browsers began to move in different directions, Web designers were forced to create multiple versions of their websites, which accounted for 25% of their work (Castro 2007, 16). The growing problem was largely ignored during the Dotcom Boom, but after the Dotcom Bust, when funds suddenly became scarce, the situation became untenable.

As the inventor of the World Wide Web, Berners-Lee was the logical person to start an organisation like the W3C. In creating the W3C, it was felt the only way forward for the Web was to make sure every stakeholder in the development of a particular area of the Web was part of the decision making process. By making competitors part of the development of the Web as a whole, it forced them to

make compromises, or be left in the cold. After such a fractious beginning, the move to Web standards has been understandably slow, and it would be close to a decade before browsers had real standards compatibility, and Web designers would start to implement them properly. In Jeffrey Zeldman's 2003 standards manifesto *Designing with Web Standards*, he describes this crisis of obsolescence thus:

> Peel the skin of any major 2003-era site, from Amazon to Microsoft.com, from Sony to ZDNet. Examine their torturous non-standard markup, their proprietary ActiveX and JavaScript (often including broken detection scripts), and their ill-conceived use of CSS—when they use CSS at all. It's a wonder such sites work on any browser.
>
> These sites work in yesterday's mainstream browsers because the first four to five generations of Netscape Navigator and Microsoft Internet Explorer did not merely tolerate non-standard markup and browser-specific code; they actually encouraged sloppy authoring and proprietary scripting in an ill-conceived battle to own the browser space.
>
> Often, nonstandards-compliant sites work in yesterday's browsers because their owners have invested in costly publishing tools that accommodate browser differences by generating multiple, non-standard versions tuned to the biases of specific browsers and platforms…(Zeldman 2003, 29).

The stated mission of the W3C is 'To lead the World Wide Web to its full potential by developing protocols and guidelines that ensure long-term growth for the Web' (World Wide Web Consortium 2012). This mission is guided by the two principles 'Web for All' and 'Web on Everything' both of which deal with inclusion; the inclusion of all people in the first case, and the inclusion of all forms of technology

in the second. The vision of how this mission and guiding principles might be carried out has expanded in recent years. The 'Web of Data and Services' has now been joined by the 'Web for Rich Interaction', which effectively separates standards for content from standards for design and architecture; the Semantic Web falling squarely within the 'Web of Data and Services'. The third leg of the vision has remained largely the same for many years. The 'Web of Trust' remains fundamental to the ongoing success of the Web (World Wide Web Consortium 2011b), and the trust layer at the top of the Semantic Web layer cake shows the importance the W3C has put on building in mechanisms for ensuring trust at the data level all along. Without the implementation of standards, the concepts and technologies behind the Semantic Web would simply not be possible. Currently, the W3C has published a list of specific standards (also referred to as recommendations) addressing every area of development listed in their mission, principles and vision. Within the last five years however, the Semantic Web has gone from being one area under development at the W3C, to something that is now foundational.

2.3 The Semantic Web

In recent years, the W3C has moved away from a traditional technology stack for its standards, to groups of overlapping standards which define an 'Open Web Platform' for application development (World Wide Web Consortium 2011a). These groups include:

- Web Design and Applications
- Web Architecture
- **Semantic Web**
- XML Technology
- Web of Services
- Web of Devices
- Browsers and Authoring Tools

Within the Semantic Web group, there are a variety of standards under development. The Semantic Web is built upon overarching standards like XML, but there are also a group of standards that constitute technologies created specifically for the Semantic Web. The most foundational standards (and the standards used for this research) are:

RDF: The Resource Description Framework (RDF) is made up of several standards completed largely in 2004. This includes the overarching Concepts and Abstract Syntax which define the standard, the Semantics specification defining the precise semantics and inference rules for the standard, the Vocabulary Description Language (also known as RDF Schema or RDFS) which defines the RDF language, and the RDF/XML Syntax Specification which defined the first serialisation (though certainly not the last) format for writing and storing data in RDF format (World Wide Web Consortium 2011c).

OWL: The Web Ontology Language (OWL) is made up of several standards completed largely in 2004, which was superseded by OWL 2 in 2009. OWL is a declarative language (as opposed to a programming language) for modelling ontologies. Ontologies are a formal way of modelling knowledge by creating precise descriptions within a particular information domain, and defining relationships between those descriptions. This allows information readable by humans to have defined meanings, and thereby allow information to be understood by applications in an automated way (World Wide Web Consortium 2009b).

SKOS: The Simple Knowledge Organization Systems (SKOS) also became a standard in 2009. It is a semi-formal way of organising information within a particular domain (as opposed to the fully formal modelling required in OWL), for lighter weight Knowledge Organization Systems (KOS) such as thesauri, classification schemes and taxonomies within the Semantic Web. SKOS is meant to work alongside OWL, providing a way to model light-weight knowledge

domains more easily when the expressivity of OWL is not necessary, but can incorporate elements from OWL as required (World Wide Web Consortium 2009a).

SPARQL: The recursively named SPARQL Protocol and RDF Query Language (SPARQL) became a standard in 2008. The protocol defines how queries should be conveyed and understood by query processers (World Wide Web Consortium 2008c), and the query language defines the syntax and semantics for querying data in RDF. As will be demonstrated in the fourth chapter, SPARQL queries can be used across data that is either in a virtual or native RDF serialisation and can return data as a result set or as a subset of an RDF graph (World Wide Web Consortium 2008a).

In addition, at the time of writing there are several other W3C standards which have been completed for the Semantic Web (World Wide Web Consortium 2011d):

> **GRDDL:** Gleaning Resource Descriptions from Dialects of Languages (2007)
> **RDFa in XHTML:** Syntax and Processing (2008)
> **POWDER:** Protocol for Web Description Resources (2009)
> **RIF:** Rule Interchange Format (2010)

These standards do not correspond directly with the Semantic Web layer cake, as they are particular protocols and technologies developed to allow implementation of the layers, not the layers themselves, nor are they all necessary to every Semantic Web application. The rest of this chapter will focus on the components making up the Semantic Web layer cake as first defined by Tim Berners-Lee, and how the layer cake has grown and changed over the last decade, as the Semantic Web has come to life.

2.4 XML

While the addition of CSS to HTML allowed a critical separation between formatting and structure on the Web, increasingly HTML was being used to add further functionality for which it was not designed, and once again began to show its limitations (Story 2000). With the advent of the W3C, work began on the eXtensible Markup Language (XML) to handle that functionality, along with XHTML (an XML compatible version of HTML) to continue to handle structure. Unlike HTML and XHTML, which are 'markup' languages, XML is a 'metamarkup' language, which means it is a language used to create other languages (Watt 2002, xviii).

As described in the previous section, the battle for standards is relatively recent, and ongoing, but the advent of XML was what began to turn the tide. Again, Jeffrey Zeldman explains its appeal:

> The Extensible Markup Language standard, introduced in February 1998, took the software industry by storm. For the first time, the world was offered a universal, adaptable format for structuring documents and data, not only on the web, but everywhere. The world took to it as a lad in his Sunday best takes to mud puddles (Zeldman 2003, 102).

The most unexpected thing about XML, was its quick adoption by people not necessarily interested in anything to do with the Web. XML offered a fundamentally different way to manage information to that in the past, (with the added advantage of being 'Web-ready').

> Why has XML seized the imagination of so many disparate manufacturers and found its way into their products? XML combines standardization with *extensibility* (the power to customize), *transformability* (the power to convert data from

> one format to another), and relatively seamless data exchange between one XML application or XML-aware software product and another.
>
> As an open standard unencumbered by patents or royalties, XML blows away outdated, proprietary formats with limited acceptance and built-in costs. The W3C charges no fee when you incorporate XML into your software product or roll your own custom XML-based language. Moreover, acceptance of XML is viral. The more vendors who catch the XML bug, the faster it spreads to other vendors, and the easier it becomes to pass data from one manufacturer's product to another (Zeldman 2003, 106).

Couple this with the fact that anyone can add on to an existing application by writing their own application, while not making the data more cumbersome to work with, and the appeal of XML becomes apparent.

2.4.1 The structure of XML

For those used to coding in (X)HTML, XML looks visually similar. (X)HTML is written with predefined tags surrounding text. An (X)HTML document containing the name of someone recording data in an archaeological project would look like:

```
<p>Maurice Moss</p>
```

Any (X)HTML Web browser will understand how to display the text between the tags as a separate paragraph (p stands for paragraph), but it does not understand anything about the fact that the tags contain the name of the recorder. All (X)HTML can do is 'markup' the structure of a document. XML being a 'metamarkup' language which is 'extensible', it can be used to create tags specific to the type of information they will hold (Harold and Means 2002, 4). For example, in an XML document made to hold information about archaeological field recording could use a descriptive tag to show who recorded the information:

<recordedby>Maurice Moss</recordedby>

If it is important to extract, manipulate or compile a list of individuals who were responsible for recording a particular site, the information is now set apart and queryable. At the same time, the XML tag doesn't tell the browser anything about how to structure the information within the document. XML is meant to work with (X)HTML, rather than one replacing the other. While markup languages are predefined by the W3C to be 'read' by current Web browsers, metamarkup languages are defined by the needs of developers working within a particular knowledge domain to satisfy the way their content will be used.

XML tags grouped together for a particular purpose form an 'XML application'. Of course, XML tags are not created arbitrarily, and must be formally defined elsewhere in order to be validated. In order for an XML document to be 'valid' it must be 'well formed' (the XML syntax is written correctly), and include structuring information which corresponds correctly to either a Document Type Definition (DTD) or an XML Schema. DTD is a more limited way of defining XML structure, and is not written in XML syntax. While DTD may be adequate for simple tasks, XML Schema is more appropriate for the Semantic Web. It is powerful, and is written in native XML syntax. Much of its power comes from the use of Namespaces (NS), which allows for absolute precision in declaring to which source of information the schema is referring. This is termed 'disambiguation', and makes it possible to pull data from a wide variety of sources for a single use (Antoniou and van Harmelen 2004, 37-45).

For example, a developer is writing an XML Schema for field recording in an archaeological project. The tags <recordedby></recordedby> would be defined in the schema as the name of the person who recorded the data. Another way to think about the relationship between an XML document and its schema is to picture a map with a key. Say you have a population map of London based on data from 1850. The lowest concentrations of people are labelled in blue and the highest

concentrations are red. The XML document is the map itself. It is within the XML Schema where the colour red is defined as meaning 'area of highest population'. A query written against the XML document asking for all of the data which is labelled in red will return the areas with the highest population.

If an archaeologist working on another project saw <recordedby></recordedby> within the field recording schema, felt this definition of the name of the person responsible for field recording matched their own, and wanted to use it for their own recording project (along with any other parts of the schema deemed appropriate), then the data in that field becomes 'interoperable' with the data from the first project. Interoperable data is much more powerful and flexible, as it can be combined and compared quickly and easily, and allows for new and different interpretations. (Harold and Means 2002, 5). This concept of interoperability is at the heart of the Semantic Web.

2.4.2 XML as the foundation of the Semantic Web

Despite its uptake in many areas outside of the Web, XML was developed as a Web technology first and foremost. In choosing SGML as the basis for all the markup and metamarkup languages for the World Wide Web, Berners-Lee made a very shrewd choice. In Uche Ogbuji's commentary in the XML retrospective issue of the IBM Systems Journal (remember SGML was developed at IBM), he discusses the importance of 'generic coding' and outlines some of the hard lessons which have…

> …taught us that it is extremely valuable to develop data so that it outlives the applications that presently operate on it. XML, used properly can help prevent such crises...[and] generic coding is the foundation of XML and related technologies. One of the most important principles you should adopt in using XML is 'If any aspect of the XML design is too closely tied to the application, consider that a bug.' (Ogbuji 2006).

Because of this, XML is well suited to forward migration, which is particularly important in disciplines like Archaeology. Applications and hardware will continue to change, and data generated by archaeological projects may or may not have the funding to keep up. As a discipline that traditionally generates large amounts of data, more similar to the Sciences, but with funding resources often tied to the Arts and Humanities, Archaeology has to be both creative and careful about how its data is handled, and take into account the best ways to make it available for the future. As will be discussed in the next chapter, archaeological fieldwork, especially excavation, is essentially a destructive process, and its results unrepeatable. For this reason, archaeologists have a special obligation to make sure their data is available for future use.

Trying to unravel 'which came first, XML or the concept of the Semantic Web' in the mind of Tim Berners-Lee is difficult. It is important to note that XML is not *necessary* to the Semantic Web, it is just what was created to fill a particular technological need. At the same time, the vision in Berners-Lee's mind for the Semantic Web would likely not have been possible without XML as the foundational idea of how it could actually work. At this point, however, XML and the Semantic Web seem to be fuelling each other.

> XML is at the heart of many of today's nascent technologies. For example, as search engines improve and the world moves towards the Semantic Web, XML is how webmasters can add meaningful information to their pages. Grid computing and autonomic computing continue to gain ground, and XML figures prominently in these technologies, as well. Database vendors continue to look at storing XML more efficiently, and XML Query Language gains steam…The semantic Web doesn't require XML, but you'd be hard-pressed to see that from the way the technology currently looks. Most information is encoded in some form of XML, whether it is the Resource

> Description Framework (RDF), or independent microformats. This is because of XML's nearly universal readability and understandability (IBM DevelopWorks 2011).

XML with XML Schema is the application independent way of holding and defining data so that it is extensible and transformable, can be customised to meet the needs of a particular knowledge domain and provide the needed functionality not available in HTML. XML on its own does nothing to define the relationships between the data however, which provide context and meaning, and are essential to the Semantic Web. For that, the W3C has developed an XML application called the Resource Description Framework (RDF).

2.5 RDF

If XML with XML Schema provide an extensible and interoperable way to structure and describe data, then RDF and RDF Schema (RDFS) provide the way to define the relationships between that data to give it meaning. When RDF was first released as an official W3C recommendation in 1999, its purpose was to provide metadata (data about data) for XML. Since the most recent version of the specification released in 2004, this has expanded considerably. Rather than just holding metadata (like the name of the person who created the XML document, or the date it was created), RDF became the primary glue that holds the information in an XML document together (Tauberer 2006). RDF puts the data into context.

While XML is a markup language, RDF is not. RDF is a data modelling framework for defining the structure and relationships within a data resource. The different types of data to be modelled are typically referred to as classes, and the relationships between the classes are called properties. Classes are things to be queried, assessed, quantified or manipulated. In archaeology, this might include artefacts, places, soil types, features or contexts. Properties describe the relationships between resources, like 'is fashioned from', 'is in the style of', 'is

nearby', 'is parallel to', 'cuts' or 'is cut by', etc. The classes and properties of resources are asserted by something called a statement, and a statement is made up of something called an 'RDF triple' (Antoniou and van Harmelen 2004, 63-4).

2.5.1 RDF triples

An RDF triple is two classes joined by a property, and is expressed as a 'subject, predicate, object' grouping (World Wide Web Consortium 2004a). The RDF triple is the basic building block of the Semantic Web, and all Semantic Web data can be organised within this simple format. If the purpose of the Semantic Web is to be 'an environment where a huge amount of mundane and specialised information traditionally processed by humans becomes processable by machines', then the RDF triple is where machine processing begins.

> …three pieces of information are all that's needed in order to fully define a single bit of knowledge. Within the RDF specification, an RDF triple documents these three pieces of information in a consistent manner that ideally allows both human and machine consumption of the same data. The RDF triple is what allows human understanding and meaning to be interpreted consistently and mechanically (Powers 2003, 17).

Taking an example from data in the ADS archive *Stone in Archaeology: Towards a digital resource* (Peacock 2005), archaeological information expressed as an RDF statement might look like:

> Ashford Black Marble *was mined at* Arrock Mine

In this statement, the subject is 'Ashford Black Marble' the predicate is 'was mined at', and the object is 'Arrock Mine'.

RDF can be written in a variety of ways, known as serialisations. One of the simplest is called N3 notation (Daconta *et al.* 2003, 89). In N3, the above statement could be written thus:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix sia: <http://dataservice.ac.uk/stoneinarchaeology/ontology/>

<sia:Ashford Black Marble>  <sia:was-mined-at>  <sia:Arrock Mine>
```

The first line defines the location of the RDF syntax on the W3C Website, so any machines processing the data will understand that it is in RDF and how to read it, and the second line defines the location of the (entirely fictitious) *Stone in Archaeology* ontology, so any machines processing the data will recognise the terms and understand the relationships between them. Rather than include the entire Web address for each piece of information, namespace abbreviations are used to create a shorthand which is much more human readable. In this case 'rdf' for RDF and 'sia' for *Stone in Archaeology*. Each piece of the triple can then be written using the appropriate prefix followed by the desired value. The third line shows the triple statement 'Ashford Black Marble was mined at Arrock Mine'. More information could be added easily:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix sia: <http://www.york.ac.uk/stoneinarchaeology/ontology/>

<sia:Ashford-Black-Marble>  <sia:was-mined-at>  <sia:Arrock-Mine>
<sia:Ashford-Black-Marble>  <sia:was-mined-at>  <sia:Rookery-Mine>

<sia:Rookery-Mine>  <sia:is-a-quarry-near> <sia:Bakewell>
<sia:Arrock-Mine>  <sia:is-a-quarry-near> <sia:Bakewell>
```

This group of statements tells us that Ashford Black Marble was mined at both Arrock and Rookery Mines, and that both mines are in quarries near Bakewell. Note that Arrock Mine and Rookery Mine appear as both subjects and objects. This shows the essential way data links together in RDF. Groups of linked triples come together to form what is known as a 'graph'. As graphs are built, new relationships can be inferred. In the example above, it could be inferred that

Ashford Black Marble can be found near Bakewell, even though it is not stated explicitly. In addition to interoperability, data held in RDF in graph format can branch in any direction to form new relationships whenever new data is added, which is very difficult in traditional relational data structures.

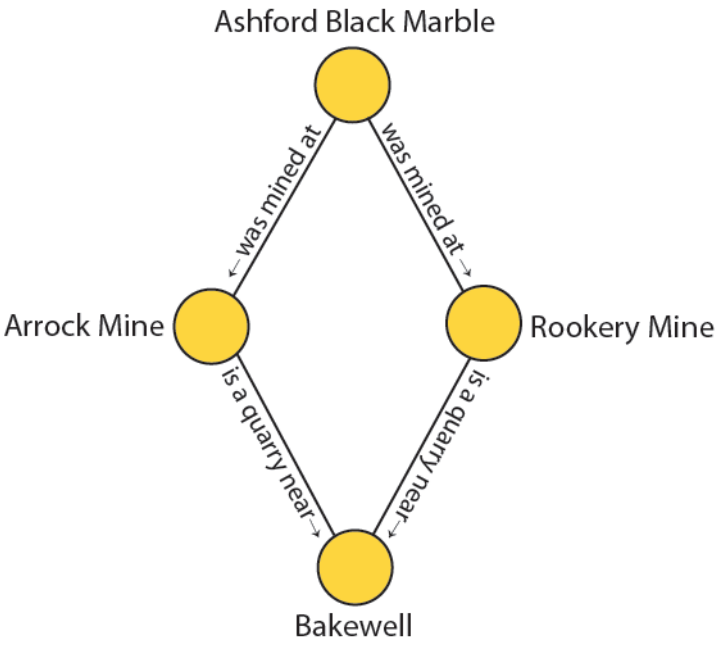


Figure 4: An image of how the preceding N3 triples look linked together to form the beginnings of an RDF graph. It shows how information can be liked together, and new information inferred from data not stated explicitly, like the presence of Ashford Black Marble near Bakewell.

2.5.2 RDF/XML

Another way of expressing RDF is an XML compatible syntax referred to as RDF/XML. It is less human readable than RDF N3 notation, but as the first RDF serialisation format, and the one developed as a standard by the W3C, it has the widest compatibility. When writing RDF by hand developers often use N3 or some other format (like N-Triples or Turtle), and then use an automated tool to convert it to RDF/XML (Daconta *et al.* 2003, 89, 103). The previous N3 statement written in RDF/XML syntax would look like:

```
<rdf:RDF
        xmlns:RDFNsID1='#'
        xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntx-ns#' >

        <rdf:Description rdf:about='#Ashford-Black-Marble'>
                <RDFNsID1:was-mined-at>
                        <rdf:Description rdf:about='#Arrock-Mine'>
                        </rdf:description>
                </RDFNsID1:was-mined-at>
        </rdf:description>

        <rdf:Description rdf:about='#Ashford-Black-Marble'>
                <RDFNsID1:was-mined-at>
                        <rdf:Description rdf:about='#Rookery-Mine'>
                        </rdf:description>
                </RDFNsID1:was-mined-at>
        </rdf:description>

        <rdf:Description rdf:about='#Rookery-Mine'>
                <RDFNsID1:is-a-quarry-near>
                        <rdf:Description rdf:about='#Bakewell'>
                        </rdf:description>
                </RDFNsID1:is-a-quarry-near>
        </rdf:description>

        <rdf:Description rdf:about='#Arrock-Mine'>
                <RDFNsID1:is-a-quarry-near >
                        <rdf:Description rdf:about='#Bakewell'>
                        </rdf:description>
                </RDFNsID1:is-a-quarry-near>
        </rdf:description>

        rdf:resource='http://www.york.ac.uk/stoneinarchaeology/
        ontology' />

</rdf:RDF>
```

It is easy to see why one type of notation might be chosen over the other, although for those well versed in writing (X)HTML and XML, using the RDF/XML syntax will feel familiar.

To add more functionality to RDF statements, reification is an important feature that can be used. In the context of the Semantic Web, reification means *to make statements about statements.* It is a way of introducing an auxiliary object into an RDF relationship without changing the statement from a triple to a quadruple, etc. (Antoniou and van Harmelen 2004, 67). Returning to the first example from the *Stone in Archaeology* archive, it would allow the statement:

> According to the 'Stone in Archaeology' archive, Ashford Black Marble was mined at Arrock Mine.

In RDF/XML this statement reifies as:

```
<rdf:Statement rdf:about= ' According-to-the-'Stone-in-Archaeology'-
archive '>
        <rdf:subject rdf:resource='Ashford-Black-Marble' />
        <rdf:predicate rdf:resource='was-mined-at' />
        <rdf:object>Arrock Mine</rdf:object>
</rdf:Statement>
```

2.5.3 RDF and relational databases

RDF probably feels quite foreign for those used to organising data in a relational database. Thomas Passin does not think the differences are terribly vast however, and as stated in the previous section organising information using RDF has some advantages:

> …any well-designed set of tables can be rewritten in the form of RDF triples…if RDF and relational tables are in some sense equivalent—then why bother with RDF? The database will probably have better performance if the data is perfectly regular, but with RDF, the data doesn't have to be regular. You can always add new triples representing new information. If the triples have a different predicate (the data type represented by the column name), they won't fit into an existing table. This would cause some problems for a conventional database but none for an RDF data store. So, an RDF data store has the

> advantage of flexibility. In addition, you can make statements about the predicates as well as statements about property values, because predicate types are also resources in RDF. In other words, RDF can describe many of the properties of its own data structures (Passin 2004, 27).

Archaeological data is not always regular, so RDF might be worthy of exploration for that reason alone. Whether they are aware of it or not, most people are used to the way a table or a table that is part of a relational database organises data. Even with reification, it is difficult to imagine how organising data in the form of a triple can form the complex associations we experience with a relational database. Passin asks this question as well:

> Is a triple enough to represent all the data you might be interested in? Consider the case of a conventional database… All the items in a row normally belong together, whereas the dismemberment into a collection of triples seems to lose that connection…The reason is simplicity: Triples are smaller and simpler than anything bigger. Data structures within programs can be simpler, because their size will always be the same (Passin 2004, 28).

Data in an RDF triple structure represents a very different way of thinking about data than a traditional relational data structure, and it takes some time to get used to. While it may feel more unstructured and organic than a relational database, it is an equally rigorous way of storing and organising data. RDF has an organic feel about it, much like the Web itself. Data in RDF format allows very disparate data sources to work together in ways that are simply not possible with standalone or proprietary database structures, and as stated in the previous section, RDF also makes it easy to infer new information from existing data, which would either be much more difficult, or simply not possible with a traditional relational database.

2.5.4 RDF Schema

RDF Schema (RDFS) plays quite a different role in relation to RDF, in comparison to the role XML Schema plays to XML.

> The name RDF Schema is now widely regarded as an unfortunate choice. It suggests that RDF Schema has a similar relation to RDF as XML Schema has to XML, but in fact this is not the case. XML Schema constrains the structure of XML documents, whereas RDF Schema defines the vocabulary used in RDF data models. In RDFS we can define the vocabulary, specify which properties apply to which kinds of objects and what values they can take, and describe the relationships between objects (Antoniou and van Harmelen 2004, 62).

As RDF plays a role similar to a relational database, RDF Schema is more similar to a relational database schema. 'The RDF Schema provides the same functionality as the relational database schema. It provides the resources necessary to describe the objects and properties of a domain-specific schema—a vocabulary used to describe objects and their attributes and relationships within a specific area of interest' (Powers 2003, 86). Because RDF Schema has a predefined vocabulary, there are a core group of classes and properties that make up RDF Schema 'elements' (World Wide Web Consortium 2004c).

RDFS classes	**RDFS properties**
Resource	range
Class	domain
Literal	type
Datatype	subClassOf
XMLLiteral	subPropertyOf
Property	label
	comment

Below is a very simple example of RDFS with a class (Class) and its associated property (subClassOf) from the *Stone in Archaeology* archive data, stating that Ashford Black Marble is from Arrock Mine:

```
<rdfs: Class rdf:about='Ashford Black Marble'>
        <rdfs:subClassOf rdf:resource='Arrock Mine' />
</rdfs:Class>
```

RDFS has now made the hierarchical relationship between Ashford Black Marble and Arrock Mine explicit. To further refine an RDFS vocabulary, constraints can be introduced. Again, constraints in RDFS provide similar functionality to those in relational databases. As you can specify a data type constraint in most programming languages (a particular field can only accept data in text, integer, currency, binary, etc. format), so you can specify constraints in RDFS. While XML with XML Schema and RDF with RDF Schema give a way of structuring information, and defining the relationships between that information, there must be a way to specify the overarching concepts about the data, which express the meaning we are trying to convey. For this, ontologies are required.

2.6 Ontology

Much like 'semantic', the term 'ontology' is borrowed from another discipline, in this case philosophy, and specifically metaphysics. Far more tangible than the study of the nature of existence (and surely less daunting), as used in computer science, an ontology is meant to first describe and represent an area of knowledge and then define the common words and concepts within that area of knowledge (Daconta *et al.* 2003, 186). While defining the common terms and relationships (in the sense of shared, rather than commonplace) to be used (more formally termed classes and predicates), may seem like the easiest part of the task, in practice, it is likely the most difficult. Common words means an agreed upon terminology, which means those who wish to use the ontology must agree as to what they are and what they mean. In the case of a broad and complex discipline

like archaeology, which spans such large spatial and temporal vistas, using many different languages both to convey current research and to understand communication in past cultures, the task is considerable. An ontology meant to describe and represent a particular area of knowledge, like archaeology, is called a domain ontology.

After more than 10 years of work, the first official ontology to attempt to describe the Cultural Heritage domain was accepted by the International Organization for Standardization (ISO) in 2006. At the time of this writing, the most recent official version was released in January of 2010 (Crofts *et al.* 2010). Created in partnership with *Le comité international pour la documentation des musées* (CIDOC), the CIDOC Conceptual Reference Model (CRM) set out to create an ontology for the entire Cultural Heritage sector, with an eye to Semantic Web applications.

> The CIDOC CRM is intended to promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to. It is intended to be a common language for domain experts and implementers to formulate requirements for information systems and to serve as a guide for good practice of conceptual modelling. In this way, it can provide the 'semantic glue' needed to mediate between different sources of cultural heritage information, such as that published by museums, libraries and archives (CIDOC CRM 2011).

While the CIDOC Documentation Standards Working Group (DSWG) were busy creating the CRM, the W3C was equally hard at work creating technologies to make use of the domain ontologies under development. The W3C formed the Web Ontology Working Group to create a more powerful ontology modelling language, which resulted in the Web Ontology Language (OWL) (Antoniou and van Harmelen 2004, 109). OWL was based on the American DARPA Agent Markup

Language (DAML) and the European Ontology Inference Layer (OIL), OWL built upon them both to provide a more universal way of representing knowledge. 'Where earlier languages have been used to develop tools and ontologies for specific user communities (particularly in the sciences and in company-specific e-commerce applications), they were not defined to be compatible with the architecture of the World Wide Web in general, and the Semantic Web in particular' (World Wide Web Consortium 2004b).

2.6.1 Types of ontologies

There are several types of ontologies, and which is most appropriate for a particular task is dependent on how weak or strong the ontology needs to be. Taxonomies, which are sufficiently strong for relational databases, and Thesauri, which are sufficiently strong for entity-relational (ER) databases, are not strong enough for the Semantic Web. The CIDOC CRM is a conceptual model, which is considerably stronger, though not as strong as the (currently hypothetical) next step to a true Semantic Web, known as local domain theory (Daconta *et al.* 2003, 166).

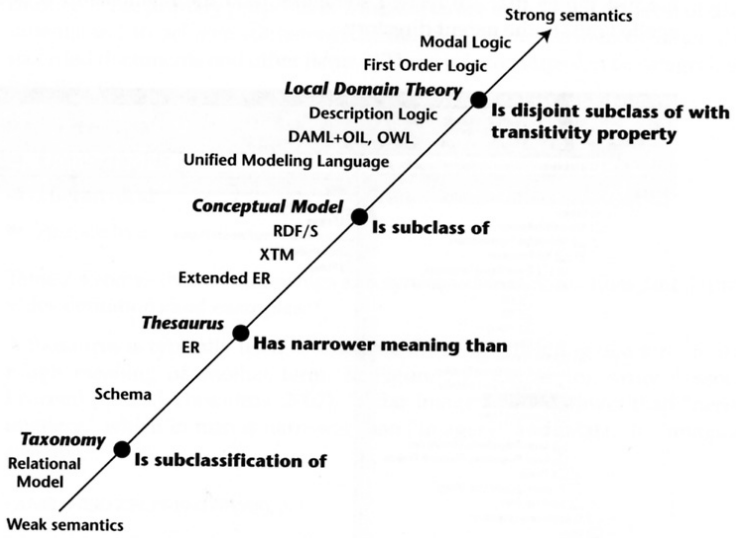


Figure 5: The ontology spectrum. The strength or weakness of an ontology is governed by how richly it can express meaning. An ontology is considered weak if it can express only simple meanings, whereas it is considered strong if it can express meanings that are 'arbitrarily complex' Reproduced from Daconta *et al.* (2003, 157). The ontology spectrum as defined by Daconta is one way of expressing the differences between types of ontologies, but not necessarily the only one (Doug Tudhope pers. comm. November 2011).

2.6.2 An archaeological ontology

OWL became an official W3C recommendation in 2004, and the CIDOC CRM became an ISO standard in 2006. With the completion of OWL and the CIDOC CRM, both of the necessary components to allow rich semantic modelling with archaeological data were present. The most comprehensive example thus far has come from the Centre for Archaeology (CfA) at English Heritage in the form of the Revelation project. Revelation began as an assessment exercise to create 'a coherent digital information system that will make the capture, analysis and dissemination of CfA research faster and more effective' (May and Cross 2004, 166), and became a project in ontological modeling. After assessing the particular needs of a complex archaeological organisation, they determined themselves to be:

> …a rather disparate grouping, or 'archipelago', of diverse, specialised, but rather isolated and independent information systems and databases. In many cases, due to their age, these systems do not have very clear mechanisms to enable the sharing of data either between the different data 'islands' within the CfA or with the outside world. Another outcome of this initial work from Revelation was the recognition that, whereas the conventional modelling work had proved quite successful in revealing gaps existing between systems, it did not readily enable the modelling of likely solutions, i.e. how the information held in different systems could be shared.
>
> What was needed was an approach to modelling which would produce a more conceptual overview of all the information being created. Such a model needed to include how existing data items would continue to be represented. But it should also show the conceptual relationships that pertained between data, thus allowing construction of a more complete picture of how all the data fitted together. It was at this point that the idea of using an ontological approach to modelling was considered…(Cripps *et al.* 2004, 3).

The CfA continued to refine and test the ontological modelling of their archaeological processes and felt their results were sufficiently universal that archaeologists outside of the CfA could make use of it. As such, they made the results available to anyone else wishing to use them, though as their work was carried out in the UK and uses a specific type of archaeological recording system known a 'single context recording', it was not envisaged as a universal domain ontology, but as something which could certainly be useful for other UK archaeologists (single context recording being the most common system in use). With that assumption, they began working with the University of Glamorgan to test the ontology, now called the CRM-EH, with a project called Semantic Technologies for Archaeological Resources (STAR). In addition to taking the modelling done by the CfA and building it into an actual ontology in RDF, STAR included an interoperability demonstrator using several archaeological data sets from different sources to show how mapping to a common ontology allowed them to be used together. In addition, the CRM-EH is an extension of the CIDOC-CRM specific to Archaeology, thereby retaining coarse grained interoperability with any other ontology mapped to the CIDOC-CRM, and therefore the entire Cultural Heritage sector (May *et al.* 2008). This concludes the brief overview of the first three Semantic Web layers, which are now firmly in existence. The next three layers involve the future vision of the Semantic Web.

2.7 Logic, Proof and Trust

The upper half of the Semantic Web layer cake is more hypothetical. Much like creating a single website not connected to anything else does not a World Wide Web make, structuring your data in XML, creating sufficiently expressive relationships using RDF and OWL, and using an ontology to describe your knowledge domain…does not a Semantic Web make. It is only when these creations begin to interact with those created by others, does the Semantic Web begin to form.

It is important to note, there is much overlap between the three upper layers, and in some ways trying to pick them apart seems to make the various concepts even more complex. This is also reflected in the literature. It is still sparsely discussed, with attempts being made by computer scientists and IT professionals to explore small subject areas, with just a few groups or individuals attempting any sort of holistic approach. This section will attempt to bring together some of the basic ideas surrounding the upper layers of the Semantic Web, but must be necessarily uneven and piecemeal, as work is ongoing in this area. Logic, proof and trust are the means by which the Semantic Web will start to knit together the disparate efforts within archaeology to be part of the 'Web of Data'.

Tim Berners-Lee knew that he would need overarching rules of logic that could be applied to the Semantic Web, but while the lower half of the layer cake was made up of technologies created under the auspices of the W3C, for the upper layers he states he was now stepping into research with an established history outside of his area of direct expertise. In order for rules of logic to be compatible with his Web ethos of 'as close as possible to no rules at all' he felt:

> …a universal design such as the Semantic Web must be minimalist. We will ask all logical data on the web to be expressed directly or indirectly in terms of the Semantic Web - a strong demand - so we cannot constrain applications any further. Different machines which use data from the web will use different algorithms, different sets of inference rules. In some cases these will be powerful AI systems and in others they will be simple document conversion systems. The essential thing is that the results of either must be provably correct against the same basic minimalist rules (Berners-Lee 2009a).

The application of formal logic to Semantic Web data facilitates some of its most important features, as it sets the rules for how different aspects of the Semantic

Web should interact. This includes the ability to use inference to create new information that has not been explicitly stated, specifying ontologies, describing what may be said about a particular domain and how it should be understood (knowledge representation), the detection of contradictory statements, and interoperability (Passin 2004, 128-9). Once the overarching rules of logic are defined and made explicit, it becomes possible to query the veracity of what has been returned from heterogeneous and dispersed data sources. This querying process is what makes up the Semantic Web layer known as proof. Rather than a query producing a subset of the data, proof returns the result of a reasoning process and the associated information about that information. This is important when the data is not under the direct control of the user:

> The main difference between a query posed to a 'traditional' database system and a Semantic Web system is that the answer in the first case is returned from a given collection of data, while for the Semantic Web system the answer is the result of a reasoning process. While in some cases the answer speaks for itself, in other cases the user will not be confident in the answer unless she can trust the reasons why the answer has been produced. In addition it is envisioned that the Semantic Web is a distributed system with disparate sources of information. Thus, a Semantic Web answering system, to gain the trust of a user must be able, if required, to provide an explanation or justification for an answer. Since the answer is the result of a reasoning process, the justification can be given as a derivation of the conclusion with the sources of information for the various steps (Antoniou *et al.* 2008, 663).

Complex as this is, Sergej Sizov refers to proof in the Semantic Web in a way that should make archaeologists quite comfortable. He calls it the 'Web of Provenence', and describes it as 'what' information:

> There are several kinds of 'what' information. For example, *data-what* describes the information and knowledge sources (such as which document was used for information extraction). *Transformation-what* describes how the system manipulates objects or data (such as which filtering algorithms it applied). *Personalization-what* describes the human influence on particular decisions (such as an expert's decision to include facts with low extraction confidence in the knowledge base). Finally, *infrastructure-what* describes the environment (such as parametrization of the natural language processing algorithm used, stop-word lists, and lemmatization settings) at knowledge acquisition (Sizov 2007, 94).

Understandably, the top Semantic Web layer is the most hypothetical. Trust is not something that will magically occur if all the lower layers are correctly in place. In the same forward where Tim Berners-Lee introduces the 'Oh, yeah' button mentioned at the beginning of this chapter, his vision of the 'Web of Trust' as he saw it in 2003 was still very much an outline. He describes it as something that will be:

> …a set of documents on the Web that are digitally signed with certain keys and contain statements about those keys and about other documents. Like the Web itself, the Web of trust will not need to have a specific structure, such as a tree or a matrix. Statements of trust can be added in such a way as to reflect actual trust exactly. People learn to trust through experience and through recommendation. We change our minds about who we trust and for what purposes. The Web of trust must allow us to express this (Berners-Lee 2004, xviii).

While all areas of trust will be important to archaeology, of most importance will be trust in the content of Semantic Web data. Archaeology is not a large

and anonymous world like e-commerce. It is likely we will know the people and institutions producing the data we wish to use, if not personally, then by reputation. It is the actual data itself that will require the most scrutiny and attention in the Semantic Web.

2.7.1 Logic

There are many different types of logic, but at the strongest end of the ontology spectrum as defined by Daconta *et al.* (Figure 5), are first order logic (FOL) and modal logic. First order logic (also known as predicate logic) allows 'statements about things and collections of things, their types and properties, and to qualify them, describe them, or relate them to other things' (Passin 2004, 136). Modal logic is one of the ways to take FOL a step further. There are different types of modal logic, 'in which statements may be contingent in various ways instead of just being true (or false)—that is, they might be true but aren't necessarily true.' (Passin 2004, 137).

Logic is used to set up the overarching rules that turn the controlled vocabulary and relationships of an ontology into a meaningful representation *language* that actually communicates knowledge. This is the beginning of what is referred to in Computer Science as intelligence.

> Before any system aspiring to intelligence can even begin to reason, learn, plan, or explain its behavior, it must be able to formulate the ideas involved. You will not be able to learn something about the world around you, for example, if it is beyond you to even express what that thing is. So we need to start with a *language* of some sort, in terms of which knowledge can be formulated (Brachman and Levesque 2004, 15).

Logic is expressed in a knowledge representation language, which consists of syntax, semantics and pragmatics. Syntax sets out how information is organised into sentence-like structures that are considered 'well-formed' and therefore

communicate the correct information. Semantics defines the meaning of well-formed syntax. Just because the syntax of a 'sentence' is correct, does not necessarily mean it communicates an idea correctly. Logic semantics provide a way to *check* that the meaning is what was intended. If an idea is expressed in a manner that is considered well-formed, and consistent with the semantic ideas for that language, then pragmatics define the use of the 'sentence' within the greater paragraph, chapter, etc. Once the sentence is put into context, and knowledge begins to be formulated, then *inference* becomes possible (Brachman and Levesque 2004, 15-6).

Inference is one of the most important Semantic Web concepts. By applying logic to Semantic Web ready data, we allow machine processing to begin. In addition to the example in section 2.5, a very simple illustration of archaeological inference can be shown using a stratigraphic relationship within a unit. If layers five and six in the unit lie wholly below layer four, and layer four lies wholly below layer three, then we can infer that layers five and six lie wholly below layer three. This may seem ridiculously simplistic, as this sort of task is easy for the human mind, but it is extremely difficult for an artificial mind.

With the introduction of inference, concepts long associated with Artificial Intelligence (AI) begin to be talked about in earnest with regard to the Semantic Web. This is where the Semantic Web reaches into the area of computer science known as Knowledge Representation and Reasoning, which is the part of AI concerned with thinking and intelligence. Brachman and Levesque differentiate the way computer scientists look at the concept of intelligence in contrast with other disciplines thus:

> Instead of asking us to study humans or other animals very carefully (their biology, their nervous systems, their psychology, their sociology, their evolution, or whatever), [knowledge representation and reasoning] argues that what we need to study

> is *what humans know.* It is taken as a given that what allows humans to behave intelligently is that they know a lot of things about a lot of things and are able to apply this knowledge as appropriate to adapt to their environment and achieve their goals. So in the field of knowledge representation and reasoning we focus on the knowledge, not on the knower. We ask what *any* agent—human, animal, electronic, mechanical—would need to know to behave intelligently, and what sorts of computational mechanisms might allow its knowledge to be made available to the agent as required (Brachman and Levesque 2004, xvii).

Brachman and Levesque also define the important tension that runs throughout knowledge representation and reasoning, and therefore throughout the higher levels of Semantic Web design, controlling every fundamental decision made during the construction of Semantic Web ready data. They refer to this tension as:

> …the interplay between representation and reasoning. It is not enough, in other words, to write down what needs to be known in some formal representation language; nor is it enough to develop reasoning procedures that are effective for various tasks…knowledge representation and reasoning is best understood as the study of how knowledge can at the same time be represented as comprehensively as possible and be reasoned with as effectively as possible…There is a tradeoff between these concerns… (Brachman and Levesque 2004, xvii-iii).

As with any traditional database containing archaeological data, there is always a balance to be struck between how tightly or loosely to structure the way the data is input and manipulated. If done well, it meets the needs of the research design by making information retrievable at a level of specificity that it provides the necessary data. Too much specificity returns data that is cumbersome and

inefficient, while too little results in insufficient information to answer the research questions. Archaeologists will have to continue to work carefully to find the right balance.

2.7.2 Proof

The role of proof in the Semantic Web is even more hypothetical, and until there are clear decisions about how logic will be applied, it will continue to be so. Returning to Sizov's 'Web of Provenance', for archaeological data, the most relevant area is what he terms 'database provenance.'

> Database systems usually consider provenance as describing the data's origins and the process by which it arrived as a query answer. The established terminology distinguishes between where-provenance, why-provenance, and how-provenance:
>
> - *Where* is where the given fact or statement is physically serialized in one or more RDF statements (that is, 'where does a given piece of data come from?').
> - *Why* is the collection of facts or statements that contributed to produce the query answer, such as a composed statement ('which facts contributed to this answer?').
> - *How* is how the query result was produced ('how did facts contribute to the answer?') (Sizov 2007, 95-6).

For archaeological data, we would certainly want to know all of these things. We need to know where data comes from, and about the credentials of the people and organisations involved who produced the data. Secondly, we need to know the specific criteria that led to particular information being returned, such as which parts of the database were used. If we made a Semantic Web query for all available Anglo-Saxon information from North Yorkshire, how do we know this is what was returned? Thirdly, we need to know how the information was chosen. Does information returned as Anglo-Saxon include data which is Anglo-

Scandinavian (which could be an overlapping time period and region, but refer to people of different origin), or is it considered to be different, and who made the decision, and why? These would be basic to a proof query of archaeological information.

There must also be specific ways to check the veracity of the choices being made for us, and these will take the form of proof checking mechanisms. A good explanation of how this would actually work has been made by Aaron Swartz:

> Once we begin to build systems that follow logic, it makes sense to use them to prove things. People all around the world could write logic statements. Then your machine could follow these Semantic 'links' to construct proofs…While it's very difficult to create these proofs (it can require following thousands, or perhaps millions of the links in the Semantic Web), it's very easy to check them. In this way, we begin to build a Web of information processors. Some of them merely provide data for others to use. Others are smarter, and can use this data to build rules. The smartest are 'heuristic engines' which follow all these rules and statements to draw conclusions, and kindly place their results back on the Web as proofs, as well as plain old data (Swartz 2002).

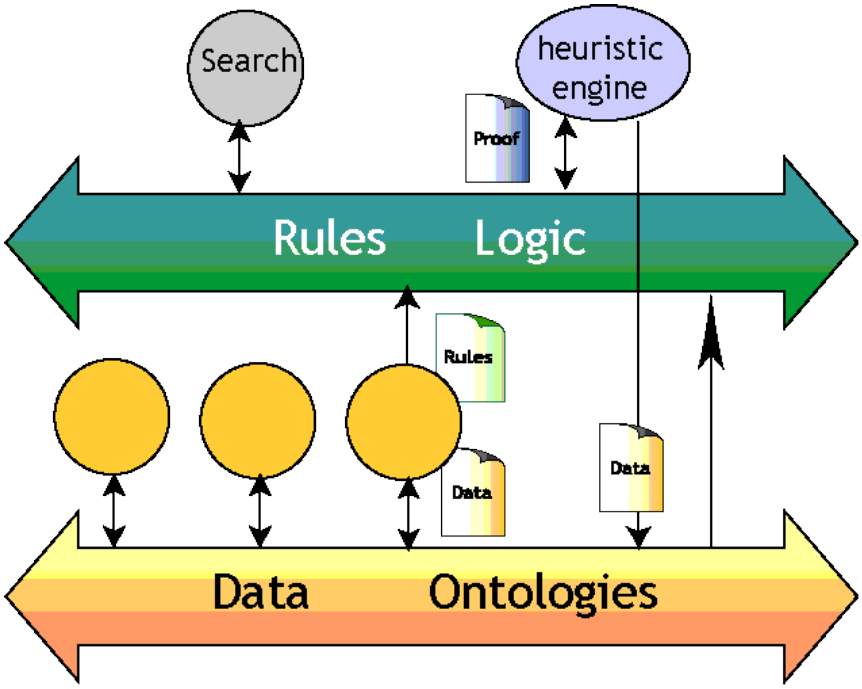


Figure 6: Image created by Tim Berners-Lee, showing the relationships of logic and proof to the lower layers of the Semantic Web. Reproduced from the W3C Website, http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide14-0.html.

Tim Berners-Lee and several others are looking toward what is termed a *Policy-Aware* Web to begin to solve these problems. In *Creating a Policy-Aware Web: Discretionary, Rule-based Access for the World Wide Web* there is discussion about the rules which will need to be considered when creating proof checking mechanisms:

> The lack of *policy awareness* in today's Web infrastructure makes it difficult for people to function as they normally would in informal or ad hoc communities. Thus, policy awareness is a property of the Semantic Web that will provide users with readily accessible and understandable views of the policies associated with resources, make compliance with stated rules easy, or at least generally easier than not complying, and provide accountability when rules are intentionally or accidentally broken (Weitzner *et al.* 2004).

The paper goes on to discuss how this might be accomplished, and while much of it remains hypothetical, rule languages are being created for constructing proofs. There are languages based in XML (e.g. RuleML), and there is a group at Stanford creating the Proof Markup Language (PML) specifically for the Semantic Web. PML is based in OWL and is meant to provide:

> …a means of describing a justification as a sequence of information manipulations used to generate an answer. Such a sequence is referred to as a Proof. A PML proof can represent many kinds of information manipulations ranging from formal logic derivations to natural deduction derivations to database and information retrieval operations to the natural language processing performed during information extraction (Pinheiro da Silva *et al.* 2006).

PML is part of Stanford's larger project, known as Inference Web (IW). The IW is 'a Semantic Web based knowledge provenance infrastructure which supports interoperable explanations of sources, assumptions, learned information, and answers as an enabler for trust' (Inference Web 2011), and consists of:

- **Provenance** - if users (humans and agents) are to use and integrate data from unknown, uncertain, or multiple sources, they need provenance metadata for evaluation.
- **Interoperability** - more systems are using varied sources and multiple information manipulation engines, thus increasing interoperability requirements.
- **Explanation/Justification** - if information has been manipulated (i.e., by sound deduction or by heuristic processes), information manipulation trace information should be available.
- **Trust** - if some sources are more trustworthy than others, trust ratings are desired (Inference Web 2011).

IW is the kind of environment that will create the sort of 'heuristic engine' Berners-Lee and Swartz talk about when they discuss proof. In addition to PML, IW consists of a toolkit, some of which is still very hypothetical. This includes a registrar called IWBase, which is 'an interconnected network of distributed repositories of proof and explanation meta information.' It also includes an 'explainer for abstracting proofs into explanations…[a] browser for displaying proofs…and planned future tools such as proof web-search engines, proof verifiers, proof combinators and truth maintenance systems' (McGuinness and Pinheiro da Silva 2004). This is all complicated stuff, and is a level of specialisation beyond what most Web practitioners will need to think about. Even Berners-Lee has admitted he is learning (Berners-Lee 2009a), but it will be important to at least understand these general principles, and stay abreast of how logic and proof develop, so we may all be working towards the top of the Semantic Web layers, which is trust.

2.7.3 Trust

The concept of authentication and digital signatures is where most of the work in the area of trust has been focussed, but there is growing consensus that these issues are just scratching the surface (Artz and Gil 2007, 58; Golbeck 2008, 1640; Hartig 2008). Projects like the work of Yolanda Gil and Donovan Artz at the University of Southern California are investigating trust in ways that might be of greatest importance to archaeologists. Rather than looking at whether a person or transaction is trustworthy, they are interested in how content on the Web can be subject to trust analysis. While interoperability is very appealing for archaeologists wishing to pose new questions by using multiple data sets, unless the data itself is trustworthy, and the reasons can be articulated, it is of little use. Gil and Artz define the challenges thus:

> Content trust is often subjective, and there are many factors that determine whether content could or should be trusted, and in what context. Some resources are preferred to others

> depending on the specific context of use of the information (e.g., students may use different sources of travel information than families or business people). Some resources are considered very accurate, but they are not necessarily up to date. Content trust also depends on the context of the information sought. Information may be considered sufficient and trusted for more general purposes. Information may be considered insufficient and distrusted when more fidelity or accuracy is required. In addition, specific statements (content) by traditionally authoritative entities can be proven wrong in light of other information. The entity's reputation and trust may still hold, or it may diminish significantly. Finally, resources may specify the provenance of the information they provide, and by doing so may end up being more trusted if the provenance is trusted in turn (Gil and Artz 2007, 228).

Gil and Artz go on to further define challenges specific to content trust as they see it, and begin to do some modelling for possible solutions, but clearly feel this is a neglected area that has received little attention. While Gil and Artz seem to be pursuing other related avenues recently, Olaf Hartig has made this his central area of research and has taken it further. In his recent paper *Trustworthiness of Data on the Web* he has begun to address this issue with practical solutions based in RDF, which he calls the *tRDF Project.* His main aim is to create data whose trustworthiness can be analysed down to the individual RDF statement (this is known as fine granularity), and has already begun modelling in tRDF (Hartig 2008).

As archaeologists currently use the Web, we make both overt and subtle decisions about how trustworthy the content in a site is all the time. The form of proof we use is frequently our own experiences or research. Using the *Stone in Archaeology* archive as an example, the homepage for the archive indicates the data and

relationships were created by staff at the University of Southampton, specifically Prof David Peacock and Kathryn Knowles. If you are not familiar with the individuals involved in the project, then the fact the information was produced by staff at a reputable university may be influential. In addition, the work was funded by the Arts and Humanities Research Board (AHRB, now the Arts and Humanities Research Council, AHRC). For those familiar with funding bodies, the fact that the project was deemed worthy of funding by a national research council might also be proof that the information can be trusted (Peacock 2005). The archive appears as part of the ADS website, which also provides verifiable credentials on its homepage. It is also funded by the AHRC, and is a project of the equally reputable University of York (Archaeology Data Service 2011a).

Once we are satisfied with the credentials of the people and institutions involved with producing the content, then the appropriateness of the data itself for the task at hand becomes the issue. The Overview section of the Stone in Archaeology archive makes the way the data was meant to be used explicit:

> This database allows the identification of stone samples by searching on the distinctive physical properties of a stone. The results of the search can be backed up by macroscopic and thin-section photomicrographs of each sample and any geologically relevant information. The resource also provides information regarding the use, quarry location/vicinity and distribution of the stone throughout various periods of history. The resource's ability to be manipulated in many different ways is one of its strengths (Peacock 2005).

This shows whether the level of specificity and type of information needed is correct for the way the data will be used. It also indicates the scientific processes used to verify the data. Next, an archaeologist would want to know how the data itself is described and organised. The archive includes a list of terminology and

definitions for the various properties used to describe the stone, and lists all of ways to search the archive. If all of these areas are found to be satisfactory, then the data will likely be deemed trustworthy.

All of these very basic criteria, and much more, will have to be satisfied before archaeologists will hand over their trust of a data resource to the Semantic Web, and allow its use to be automated. Once archaeologists do begin to make the information available in a consistent, explicit and machine-readable manner however, much more will become possible. Christopher Walton describes trust as an 'umbrella term' which covers a wide range of interrelated issues' (2007, 243). There is much work going on in the commercial sector around traditional security issues that have been part of Web development for many years, but archaeologists will continue to be more interested in data reliability, and the context surrounding any inferences made about that data, so we should continue to watch as the topmost layer of the Semantic Web develops.

2.8 Beyond the 'layer cake'

There was a period of rumination after *Weaving the Web* was published, and Semantic Web texts aimed at a popular audience began to appear during 2003 and 2004 from a variety of technology publishers (Alesso and Smith 2004; Antoniou and van Harmelen 2004; Daconta *et al.* 2003; Davies *et al.* 2003; Fensel *et al.* 2003; Geroimenko 2004; Passin 2004), all of which were attempts to define the Semantic Web, and the new technologies and languages upon which it was based. These texts were meant to explain the framework necessary to create the solutions for programming Semantic Web-based applications. After another period of rumination (and considerable work and debate), practical texts aimed at Web developers, focussing on actual programming and workflows began to appear around 2008 (Allemang and Hendler 2008; Hebeler *et al.* 2008; Hitzler *et al.* 2010; Segaran *et al.* 2009), and are moving practical concepts outside the realm of Computer Science into more mainstream demonstration.

While focussing on the popular texts published about the Semantic Web is certainly not a comprehensive way of tracking its practical development over the last decade (there were an equal number of specialist and academic texts published throughout this period), and as might be expected, there is extensive (though more haphazard) information published throughout the Web, the subject and timing of these publications serves as strong illustrative punctuation for its growth. There are currently more than a dozen new Semantic Web texts aimed at Web developers recently announced as being in press, doubling the number of total publications within the next couple of years, and showing the speed with which popular momentum and demand is suddenly gaining in this field.

2.8.1 The 'layer cake' 10 years on

Another means of exploring the growth of the Semantic Web from a theoretical construct to a practical solution is through the 'layer cake' or 'technology stack' graphic (Figure 1 shows the original version created by Berners-Lee in 2001). There are many variations of Berners-Lee's image, which continues to evolve and change, but has persisted because it was a way to give 'some illustration to the un-illustratable' (Zacharias 2007). A humorous history of the 'layer cake' was presented at the 2009 International Semantic Web Conference (ISWC) dinner by Jim Hendler, showing how complicated the current Semantic Web model has become (Figure 7), but it had a serious message. Ten years on, much work has been done, many new technologies have been created to realise (at least the lower half) of the 'layer cake', that the work is ongoing, and its momentum increasing. The fact that 'user interface and applications' now sits like a boom over the top of the graphic above the still largely untouched layers of logic, proof and trust, shows the distance still left to travel before the Semantic Web can be considered even partially complete. The concepts 'user interface and applications' were not even part of the original graphic, so far away were the ideas which would lead to any practical implementation. As the current graphic indicates, what practical exemplars do exist have grown sideways through the new technologies. Just as many people must publish linkable pages before a World Wide Web can exist, so

must a critical mass of data be published before the upper layers of logic, proof and trust can be activated, before a Semantic Web can exist.

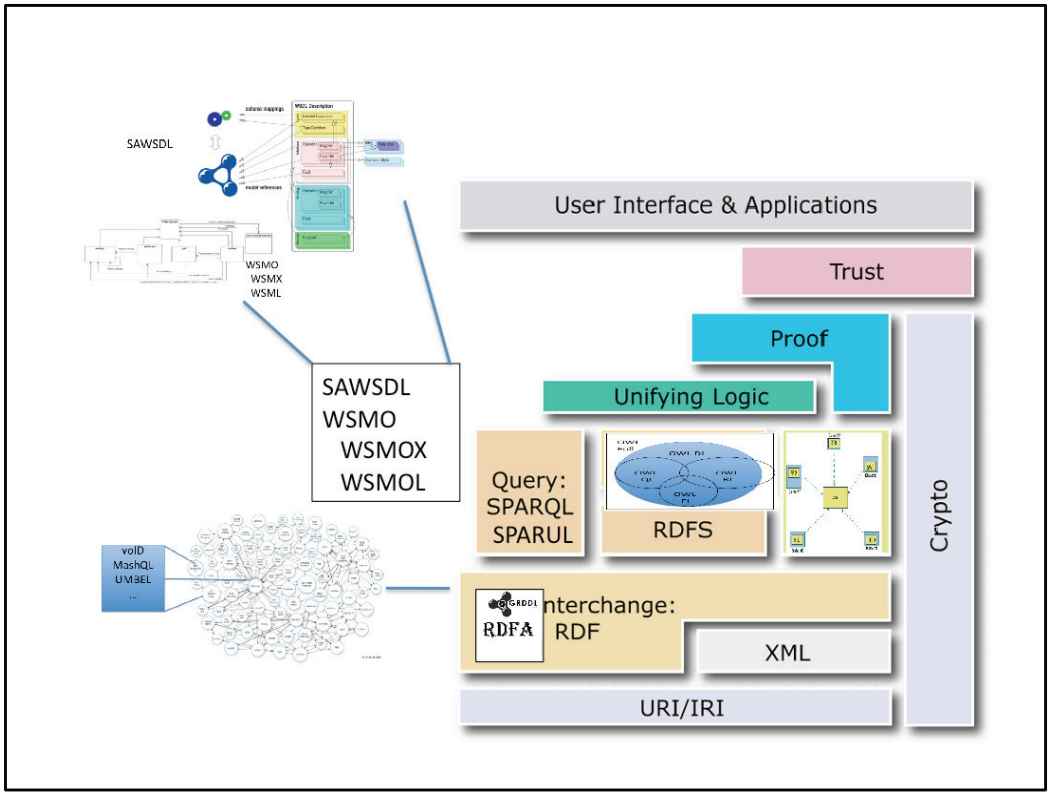


Figure 7: The Semantic Web 'layer cake' or 'technology stack' as recently presented by Jim Hendler at the International Semantic Web Conference (ISWC) in 2009. While the image was part of a humorous presentation about the history of the graphic during the conference dinner (he chose to narrate the slides using Seussian verse), it was also a serious depiction of how much the Semantic Web has grown and changed over the last 10 years. The new syntaxes (Turtle, Manchester, Structural OWL and N3) were also added in a later slide (Hendler 2009).

More than 10 years on, and still awaiting the wave of practical exemplars, the incorporeal nature of the Semantic Web (the incorporeal meets the un-illustratable), is still difficult to talk about, and much of the problem has been down to semantics—in the traditional sense. There is as much confusion as ever as to whether the thing *called* the Semantic Web is in the process of success or failure, or if it just needs to undergo some rebranding. Interestingly, there seems to be little disagreement that the potential of the Web is being hampered by its document-based format; that a raw-data structure where new relationships can be built would transform the Web and unlock that potential (which is what the

Semantic Web was always supposed to be), but this still seems to get lost in the fight over terminology. Web developers (or people who write about Web development) seem to have exceptionally short attention spans, and in an industry where the time it takes for a new technology to be adopted widely is considered an important measure of its worth, they may do the Semantic Web a disservice. That said, the name Semantic Web was an oblique choice by Berners-Lee in the first place, and something more straightforward probably would have eased the situation, but it is hard to argue with a visionary.

A better choice, and the term heard more and more frequently is Web of Data. Not as elegant, but a straightforward term for what needs to be a straightforward concept. Other largely synonymous (though how synonymous is also a source of debate) terms like Web 3.0 abound, but ultimately distract from rather than clarify what is actually happening. The title Semantic Web denotes the particular vision of Tim Berners-Lee, whereas Web of Data is an informal concept used by Berners-Lee and many others as a descriptor for what the Semantic Web is. The terms Semantic Web and Web of Data (to the exclusion of other terms) will be used for the duration of this thesis, in the manner set out by Tom Heath:

> Personally I use the term Web of data largely interchangeably with the term Semantic Web, although not everyone in the Semantic Web world would agree with this. The precise term I use depends on the audience. With Semantic Web geeks I say Semantic Web, with others I tend to say Web of data – it's not about rebranding, it's about using terms that make sense to your audience, and Web of data speaks to people much more clearly than Semantic Web (Heath 2009).

2.8.2 The Rise of Linked Data and SPARQL

Adding to this terminological confusion, is the relatively new concept of Linked Data, which was coined by Berners-Lee in 2006. Often perceived as being distinct

from the Semantic Web, in reality it is just a set of best practices for publishing data in a way which makes it part of a single global information space (Bizer *et al.* 2008, 2). These best practices define a way of taking data out of proprietary data silos (individual databases and documents), and by giving each piece of raw data its own unique address (in the form of a Uniform Resource Identifier or URI) it becomes uniquely identifiable and its location resolveable, and therefore linkable and manipulatable. Linked Data was never meant to be a replacement for the Semantic Web, or take it in a fundamentally different direction, it is merely the practical way this area of Semantic Web technology is developing (Heath 2009). The fact that it seems as though it has a life of its own, is because it has been so visibly successful.

Represented in Figure 7 as the part of the image made up of interlinked circles (and resembling a swarm of bees), Linked Data has enjoyed the hype and enthusiasm the Semantic Web has been waiting for for years. Almost as a collective sigh of relief that at last Semantic Web developers finally have a way to show their work, Linked Data has seen rapid uptake, especially by those with a mandate to make their data available (including the US and UK governments), through the W3C Linking Open Data project (World Wide Web Consortium 2010). That few resources yet exist to harvest and use that data in meaningful ways is another issue, but it is a first step in making the Semantic Web tangible which seems to be catching on.

The key to developing those resources is now part of the updated layer cake as well. Now that SPARQL is available, the protocol necessary for developing interfaces for querying Semantic Web data has been used to create 'SPARQL endpoints' for that purpose. The SPARQL query language allows queries to be written within a SPARQL endpoint that then returns the desired data. So even if full Semantic Web implementation is not yet available, Linked Data can now be queried with SPARQL and users are finally getting to see the Semantic Web in action.

2.9 Conclusion

In December of 2007, an article titled *The Semantic Web in Action: Corporate applications are well under way, and consumer uses are emerging,* was written in Scientific American as a follow up to the original piece written by Tim Berners-Lee and his co-authors in 2001. During the six years between the articles, much had changed. We are now swimming in what Dale Dougherty, the vice-president of O'Reilly Media, coined Web 2.0. As much a commercial designation as a change in technology, Web 2.0 referred to the hope that *something* would follow the burst of the Dotcom Bubble in 2000, but what this was had yet to be been defined (Vossen and Hagemann 2007, xi). In many ways that is still the case, but references to Web 2.0 are now generally accepted to mean the many forms of social media based on user created Web content.

This can take several forms. Blogs are centred on an individual opinion that can be commented on by others, a wiki can organise a group of opinions or information on a particular topic, you can rate and review something you purchased on a commercial website to help influence other consumers, you can participate in a social network like Facebook, or you can help identify the contents of a photograph in Flikr using tagging. The mainstream availability of technology, combined with a generation coming of age who have used the Web since childhood, has created a massive surge in the Web over the last several years, moving from the ground up. The Semantic Web was a mature vision in the mind of Tim Berners-Lee long before the burst of the dotcom bubble, but it was always based on a top-down approach. Confusion about terms like Web 2.0, Web 3.0, the social web and the Semantic Web, etc. has led to erroneous ideas that these are ideas in competition. In reality, the social web and the Semantic Web have much to contribute to each other and will leave the Web stronger in the end.

The Scientific American article written by Berners-Lee in 2001 depicted a foreign and rather unsettling world at the time it was published, even to Web enthusiasts. The

idea of trusting unseen machines to help make decisions about even mundane Web interactions and information was disconcerting. Today this is no longer the case, and that the contributions made by Web 2.0 account for much of the reason. The level of trust and effort we all seem willing to invest, in order to collaborate with friends and strangers alike, is astounding. Of course, trusting a person and trusting the automated parameters created by a group of people are not the same thing, but it is the real desire for collaboration to make the whole greater than the sum of its parts, which sits at the heart of both Web 2.0 and the Semantic Web.

At the same time, there can be no doubt that the Semantic Web is surging ahead as well, as evidenced by the 2007 follow-up article in Scientific American. Once again, an article marks a watershed moment. Speaking about the same period of time between the publication of the previous article by Berners-Lee and the present, the authors are clear: 'Since then the sceptics have said the Semantic Web would be too difficult for people to understand or exploit. Not so. The enabling technologies have come of age. A vibrant community of early adopters has agreed on standards that have steadily made the Semantic Web practical to use' (Feigenbaum *et al.* 2007, 91).

The majority of the examples given in the article involve current uses of Semantic Web elements in the healthcare industry, but they also cite one of the best examples of the Web 2.0 working in common with the Semantic Web:

> Consumers are also beginning to use the data language and ontologies [of the Semantic Web] directly. One example is the Friend of a Friend (FOAF) project, a decentralized social-networking system that is growing in a purely grassroots way. Enthusiasts have created a Semantic Web vocabulary for describing people's names, ages, locations, jobs and relationships to one another and for finding common interests among them. FOAF users can post information and imagery in any format

> they like and still seamlessly connect it all, which MySpace and Facebook cannot do because their fields are incompatible and not open to translation (Feigenbaum *et al.* 2007, 93).

The reason behind the success of the Web thusfar will likely continue to propel it further. The willingness of Tim Berners-Lee to let the natural way people communicate flow over him and inform his thinking when he created the Web, rather than trying to create a rigid structure and asking users to conform, is ultimately in keeping with the Web 2.0 ethos and will help to propel the Semantic Web into the mainstream as well.

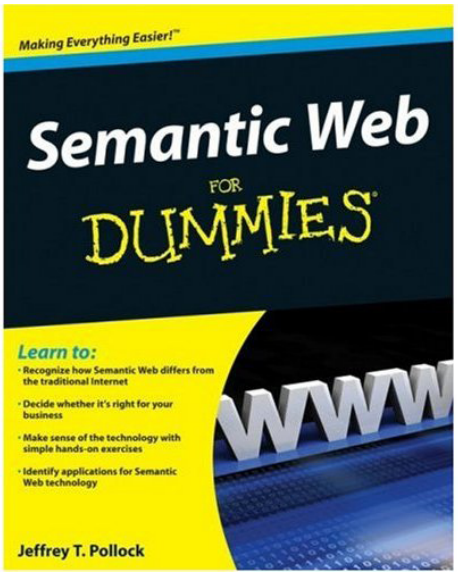


Figure 8: While there are still relatively few mainstream texts dedicated to the Semantic Web, the *Semantic Web for Dummies* was published recently, which can certainly be construed as a sign of mainstream acceptance.

The 2007 Scientific American article also addresses an issue of great concern to the further development of the Semantic Web, which is inclusion in, but never dominance of, the W3C by the commercial sector:

> As applications develop, they will dovetail with research at the Web consortium and elsewhere aimed at fulfilling the Semantic

> Web vision. Reaching agreement on standards can be slow, and some sceptics wonder if a big company could overtake this work by promoting a set of proprietary semantic protocols and browsers. Perhaps. But note that numerous companies and universities are involved in the consortium's semantic working groups. They realize that if these groups can devise a few well-designed protocols that support the broadest Semantic Web possible, there will be more room in the future for any company to make money from it (Feigenbaum *et al.* 2007, 97).

This all feels miles away from the Wild West Web of the browser wars, and a bit of law and order has clearly rolled into town. By combining Web 2.0 with a Semantic Web that is becoming more mainstream, something quite other may form. If the grassroots Web 2.0 is the stalagmite pushing its way up from the bottom of the cave and the Semantic Web is the stalactite reaching down from the cave ceiling, when they meet in the middle to form a column what then will appear? Will that be Web 3.0? Perhaps.

When asked about Web 2.0 by a reporter for the International Herald Tribune, Tim Berners-Lee:

> …shrugs at the use of the term 'Web 2.0' - a Silicon Valley buzzword to describe the Internet since the dot-com bust of the turn of the century - he does say he sees a new level of vigour across the network…'People keep asking what Web 3.0 is,' Berners-Lee said. 'I think maybe when you've got an overlay of scalable vector graphics - everything rippling and folding and looking misty - on Web 2.0 and access to a semantic Web integrated across a huge space of data, you'll have access to an unbelievable data resource (Shannon 2006).

All of the pieces are in place for the discipline of Archaeology to begin taking advantage of this ‘unbelievable data resource’ on many different levels, and the Semantic Web is now moving beyond theory into practice. How that practice might be applied within archaeology, and specifically to the data derived from field drawing, is the subject of the rest of this thesis.

Chapter Three

Archaeological Field Drawing: The Significance and Evolution of the Visual Archaeological Record

> *Archaeological draughtsmanship involves the construction of technical cryptograms, and as in all ciphers these must be made according to rules carefully observed by both transmitter and recipient. As symbol, all illustration is a transcript of reality...The draughtsman's illustrations are no more passive agents of communication than the author's words they complement and expand. A drawing must say something or it is failing in its primary purpose, exactly as a sentence or a paragraph of text must say something economically or elegantly, in clarity or in confusion.*
>
> –Stuart Piggott (1965, 165)

> *The eye travels along the path cut out for it in the work.*
>
> –Paul Klee (1953, 33)

3.1 Introduction

Field drawing in archaeology is about transformation. In the most extreme case, that of traditional excavation, the visual record *becomes* the archaeological resource. As each layer is visually recorded, it is then destroyed, carefully and methodically, but irrevocably. The physical entity is seen, understood and interpreted through many different eyes, and then channelled through the action of as many hands, into disparate two-dimensional records (Reilly 1991, 135). The individuals creating the two-dimensional records resemble the blind men of Indostan (from the historically sourced poem by John Godfrey Saxe), each describing different parts of the same elephant. It is up to the project director to attempt to understand and reconstruct the elephant as a whole (or at least as much as the excavation has revealed), but which has been transformed by others who saw it only in part.

The visual record is transformed again during the post-excavation analysis derived from the recorded data. For Jonathan Bateman, understanding the interpretative processes occurring during these transformations is critical:

> The intimate relationship between the destructive and creative processes that are excavation, and the archaeological drawings that both drive and witness them, puts the act of drawing at a conceptually crucial stage in the archaeological production process. The potency of this interpretive step, becomes inextricably intertwined with both previous and later interpretative and creative stages, such as the physical excavation itself and the writing of narratives of the past based on that excavation (Bateman 2006, 69).

It is important to consider this transformation when looking at the visual archaeological record with regard to the Semantic Web. Field drawing in archaeology has its own distinct history, but also incorporates many other disciplines, including ideas from the realms of art, visual cognition, draughtsmanship, design, quantitative communication and computing. At the same time, the visual records created as part of the process of archaeological fieldwork are distinct from other types of visual archaeological recording, such as photography or illustration.

The interpretative process of transforming an archaeological resource during excavation and visual recording is what the foundational Bauhaus drawing teacher, Paul Klee would have described as both a ‘productive’ and ‘receptive’ act. Klee believed:

> The work as human action (genesis) is productive as well as receptive. It is continuity. Productively it is limited by the manual limitation of the creator who only has two hands). Receptively, it is limited by the limitations of the perceiving

> eye. The limitation of the eye is its inability to see even a small surface equally sharp at all points. The eye must "graze" over the surface, grasping sharply portion after portion, to convey them to the brain which collects and stores the impressions (Klee 1953, 33).

This description of the loop (or continuity) experienced by those attempting to use their hands to convey information visually, is appropriate for those working in archaeological field recording. Even as an individual attempts to represent what they see, so it can be collated, understood and interpreted (in concert) by another, they are also limited by having to 'graze' over their own field unit and try to make sense of what they, as individuals, are perceiving only in part. These many perceptions are then taken by the few tasked with combining and interpreting them, in order to then distil what can be understood. This constriction is then released and transformed again, as archaeologists attempt to communicate to those not involved in the process of gathering the information, what that information might mean.

Visual communication has been an integral part of the discipline of Archaeology from its very start. Indeed, it is the visual that captures the mind, and accounts for much of its popularity. Drawings, photographs and reconstructions all fuel the imagination and ask us to ponder how things might have been in the past, and what brought us to the time and place we now inhabit. It does so in a vastly different cognitive way than textual information, and without it, the communication of ideas in archaeology is incomplete. In our text-centric world, visual language is often marginalised, and this too has been the case for archaeology (Gamble 1997, xvi).

The success or failure of the communication of the visual field record depends on many factors common to visual communication generally. These factors have been explored comprehensively across several seminal volumes by Edward Tufte,

not the least of which is the transformation of information once perceived in three-dimensions, to make it understandable when translated into two-dimensions:

> We envision information in order to reason about, communicate, document, and preserve that knowledge—activities nearly always carried out on two-dimensional paper and computer screen. Escaping this flatland and enriching the density of data displays are the essential tasks of information design. Such escapes grow more difficult as ties of data to our familiar three-space world weaken (with more abstract measures) and as the number of dimensions increases (with more complex data). Still, all the history of information displays and statistical graphics—indeed of any communication device—is entirely a progress of methods for enhancing density, complexity, dimensionality, and even sometimes beauty (Tufte 1990, 33).

How well archaeologists are able to use new technologies like the Semantic Web to convey the understanding to be harvested from the transformation of an archaeological resource from the 'three-space world' into 'flatland' will be a test of its success. The more abstract we make the archaeological data, the more we weaken the link to the original three-dimensional resource. Including visual archaeological data alongside the textual within technologies like the Semantic Web is necessary to creating a full picture of the archaeological resource, but it is important to consider the distance we are travelling from 'three-space' to 'flatland'. It is also important to remain mindful of the differing processes of creation and perceptual nature of this visual information from that which is text-based. Whether hand drawn on *permatrace,* vectorised or 'born digital', the gathering of archaeological field data begins with physical work which engages the body, and 'as much as the hand enters thinking, then thinking can be of the hand' (Rosenberg 2008).

Even in what must certainly be termed a digital age, digital methods have not replaced the creation of visual archaeological records by hand drawing. While digital survey equipment is commonplace for creating much of the large scale data capture across a site, resulting in a visual record, the intimate recording within a stratigraphic unit is still largely the purview of pencil on *permatrace.* Innovative attempts are being made to undertake digital capture of primary data, but it remains problematic (Rains 2007, 2), or requires technology beyond the financial reach of many archaeologists. At the same time, digital practitioners encounter the same problems in the field as their analogue colleagues. Sun, rain, wind, heat and cold all produce challenges to good recording practise, whether it is frozen fingers or an overheating computer processor.

Modern visual archaeological recording is usually a mixture of data source types, some (now commonly described as) 'born digital' and some digitised later from analogue sources using retrospective conversion (Hopkinson and Winters 2003). The process of digitising field drawings can be costly and arduous though, and often only a small percentage are deemed sufficiently important to justify the work. While scanning images (rasterisation) can make them easier to access, distribute and store, true retrospective conversion into vector format (vectorisation) can increase the potential functionality of the drawings greatly. As archaeology moves forward and the challenges of technological limitation and cost lessen, the use of digital survey, drawing and terrestrial 3D laser scanning will likely produce an increasingly vector-based primary data record, which in turn will make it more accessible for use with the Semantic Web. Indeed, perhaps the usefulness of vector data within the Semantic Web might be a motivation for the time and expense of applying retrospective conversion to legacy data. A history of the process of transformation known as archaeological field drawing, its increasingly digital application, its importance to modern archaeological practice, and its potential for inclusion within the Semantic Web are the subjects of this chapter.

3.2 A brief history of archaeological field drawing

In the 1960s, the editors of *Antiquity* asked Prof Stuart Piggott and Dr Brian Hope-Taylor to write several articles about archaeological draughtsmanship entitled *Archaeological Draughtsmanship: Principles and Practice*. The articles were an attempt to document both the history of draughtsmanship within the discipline, and current ideas of good practice from two of its foremost practitioners. Unfortunately, only three articles in the series were ever published; one per year from 1965 to 1967. Piggott's history, titled *Principles and Retrospect* was followed by Hope-Taylor's *Ends and Means* and *Lines of Communication,* the last two of which focus primarily on how to draw in order to get the best results from the printing technology of the time. These articles were meant to be followed by further installments covering 'the rendering of excavated plans and sections; use of conventions, mechanical and hand-drawn tints; and the composition and orchestration of archaeological drawings [and] lettering, construction and reconstructions, and drawing of small finds' (Hope-Taylor 1967, 181). For whatever reason, the series simply stops, leaving what would have been an interesting and important snapshot of current thought with regard to field recording at that time, unfinished. This is unfortunate, as very little seems to have been written about archaeological field drawing as a subject in its own right.

3.2.1 Field drawing in the 17th and 18th centuries

Piggott's history traces archaeological field drawing to its antiquarian beginnings, where it was important to provide 'an accessible corpus of material from which typological and taxonomic systems could be developed from criteria more suitably presented visually than in words' (Piggott 1965, 171). Antiquarianism in Northern Europe rose from the interest in nationalism that was part of the Reformation. It caused a shift in the use of antiquity as a means to create proof of biblical links to other parts of Europe, to a means of developing a legitimate national identity separate from Roman Catholicism (Trigger 1989, 45-6).

During the 17th century, antiquarians such as John Aubrey began to produce what we would now certainly recognise as prototypical plan drawings resulting from survey, as with his famous work at Avebury. Olof Rudbeck at Uppsala was using vertical section drawing at Gamla Uppsala to create a relative dating sequence (Trigger 1989, 49). Unfortunately, his dating of the tumuli stemmed from his desire to provide evidence of Sweden being the lost island of Atlantis, and that Gamla Uppsala was the centre of the civilisation, but it still represents a legitimate attempt to break away from an entirely biblical explanation of the world.

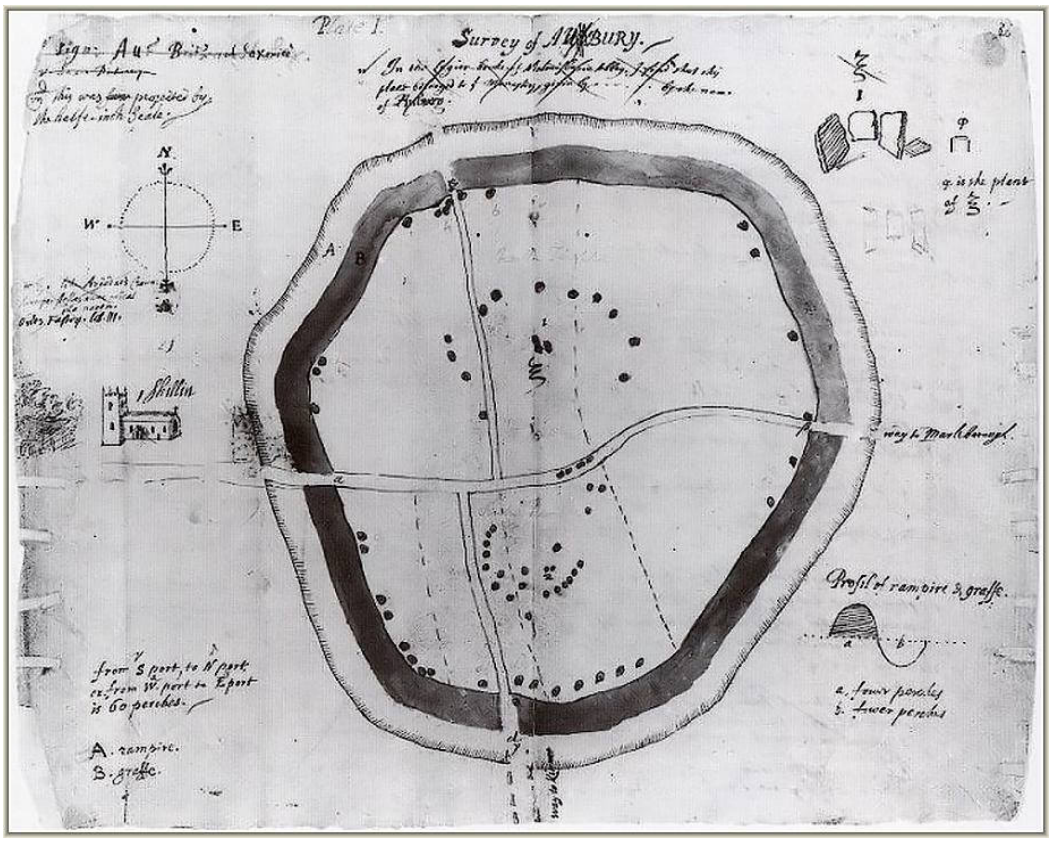


Figure 9: John Aubrey's famous drawing of Avebury, c. 1675. This drawing is an orthographic projection in plan view, which archaeologists continue to use today. Reproduced from http://www.avebury-web.co.uk/aubrey_stukeley.html.

During the 18th century, antiquarianism began to embrace the more scientific principles associated with the Enlightenment. As Northern Europe began to explore the idea that what they saw in the landscape included information from a pre-Roman past, William Stukeley and others first attempted relative dating

for sites for which there was no associated written record (Trigger 1989, 61-4). Stukeley produced drawings of Avebury and Stonehenge, but was also incorporating Romanticism, which resulted in a move away from a more scientific documentary approach, to drawings resembling the 'prospect' view of landscape painting. Such that 'plans were drawn and engraved according to the prescriptions of estate surveyors and cartographers; small antiquities were illustrated as if they were the butterflies or petrifactions or prodigies of nature which might well have accompanied them in the cabinet of curiosities of an ingenious gentleman' (Piggott 1965, 171).

3.2.2 Field drawing in the late 18th and 19th centuries

The work of William Cunnington and Sir Richard Colt Hoare in Wiltshire in the late 18th century continued to use a more scientific approach, and attempted to use stratigraphy to establish relative dates for pre-Roman sites, but also without real success. Without a systematic way of establishing a chronology for Northern Europe, prehistory was left open for any sort of speculation useful to furthering the beliefs of a particular social ideology (Trigger 1989, 67). There was considerable ebb and flow of new ideas and approaches during this time, but generally, progress seemed to follow a much more individualistic and erratic path than other disciplines moving from antiquarian to scientific methods (Roskams 2001, 10-2). The result was little real change in method across the discipline during the 18th and early 19th centuries (Piggott 1965, 171). Even when sound stratigraphic methods put clear chronologies for dating directly in view, as with the mid-19th century work of Jacques Boucher de Perthes in the Somme Valley, the conclusion was so shocking as to require a catastrophic explanation plainly contradicting the evidence. The idea that humans and mammoth existed at the same time and place could only be a mistake (Trigger 1989, 91-2).

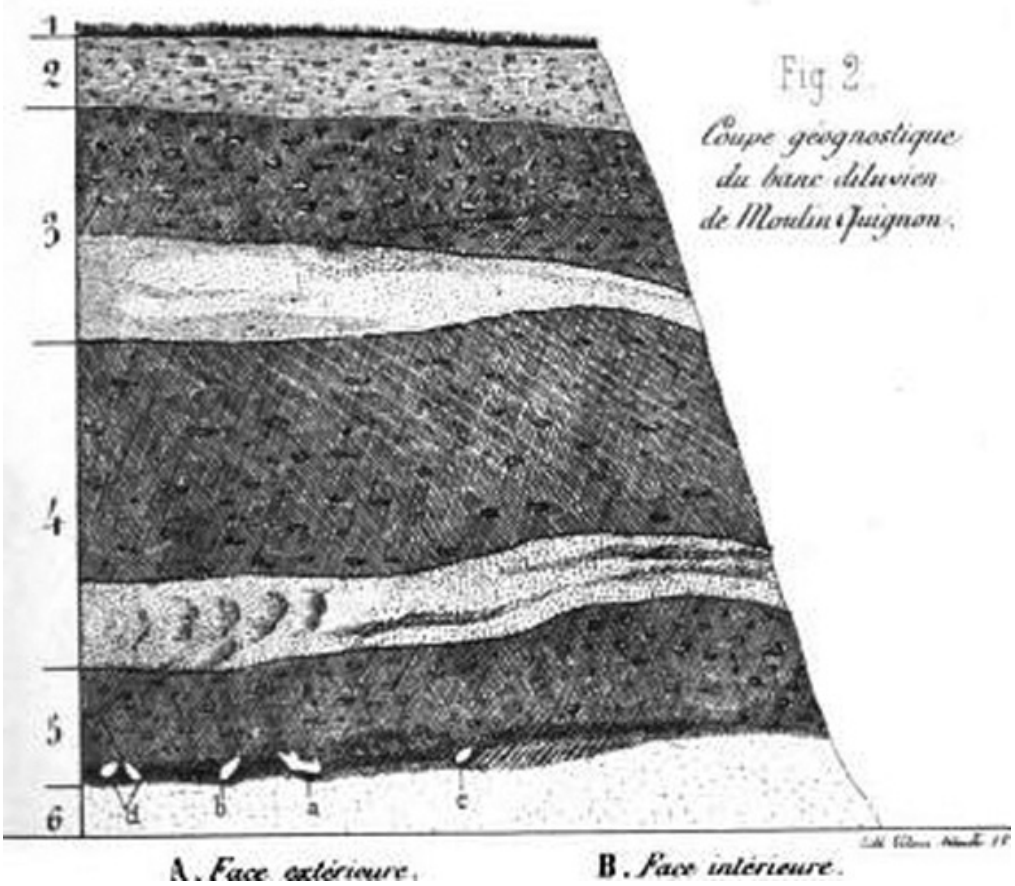


Figure 10: Section drawing showing the placement of human jaws bones found in proximity with stone tools used for hunting mammoth. Reproduced from *Antiquités Celtiques et Antédiluviennes: Mémoir sur L'industrie Primitive et les Arts a Leur Origine* (Boucher de Perthes 1864, 179).

Meanwhile, in the Mediterranean and Near East, where archaeology continued work side by side with historical documentation, methodologies were beginning to change. The density of occupation in these sites called for better ways to discern small temporal changes, and this resulted in advances in several areas, but particularly in stratigraphy, and therefore section drawing. Meticulous section drawing was pioneered by Guiseppe Fiorelli working in Pompeii, and then refined by Alexander Conze and Ernst Curtius while working at Samothrace and Olympus, respectively. Conze and Curtius were the first to be cognisant of the fact that, because excavation destroyed the site, they should attempt to create a written record to replace the archaeological resource as they destroyed it (Trigger 1989, 196). Techniques developed in the Near East made their way back to Northern Europe around the time General Pitt-Rivers was working to revolutionise

archaeological practice, and modern field recording began as we understand it (Adkins and Adkins 1989, 5; Piggott 1965, 172-4; Trigger 1989, 197).

Pitt-Rivers' recording was not just meticulous, but formed the core of his work. Not only was he creating a complete record to replace what he was destroying through excavation, he believed that the complete record should be published so that other archaeologists would be able to ask their own questions of the data (Trigger 1989, 199). That record was also a primarily *visual* record for the first time. Piggott discusses how Pitt-Rivers chose to create illustrations that were:

> …not ancillary, but the main matter of the reports, the text being a comment on the plates...A dictum attributed to Pitt-Rivers—'Describe your illustrations, do not illustrate your descriptions'—seems unfortunately not to be traced in his published works and I sometimes wonder whether I did not invent it myself. Whatever the source, like other apochryphal aphorisms of great men, it is at least in character, and makes not bad summary of his methods (Piggott 1965, 174).

The no-nonsense approach of Pitt-Rivers is evident in Piggott's example of a barrow plan. The plan takes up the whole of the page, and topographic contour lines clearly show the changes in elevation in the landscape. He chose not to use hachured survey however, which Mark Bowden feels was 'indicative of his inability to analyse relations between earthworks from surface evidence' (1991, 157). In a semitransparent overlay, the outline of the barrow is clearly visible in relation to the elevation, and other information such as a later inhumation burial, and several pottery find sites are shown. Scale and direction are clear, but do not encroach on the drawing either in size or line weight. Descriptive information is terse and simple. The plan title, location of the barrow, likely era and type of monument, name of the excavator and a brief key to the types of pottery found are all the plan includes. Pitt-Rivers understood the power of the illustration to

communicate, but while he created effective plan drawings, his section drawings were more problematic. His decision to dig in spits rather than layers meant he was unable to clearly understand his stratigraphic relationships throughout his career (Bowden 1991, 94).

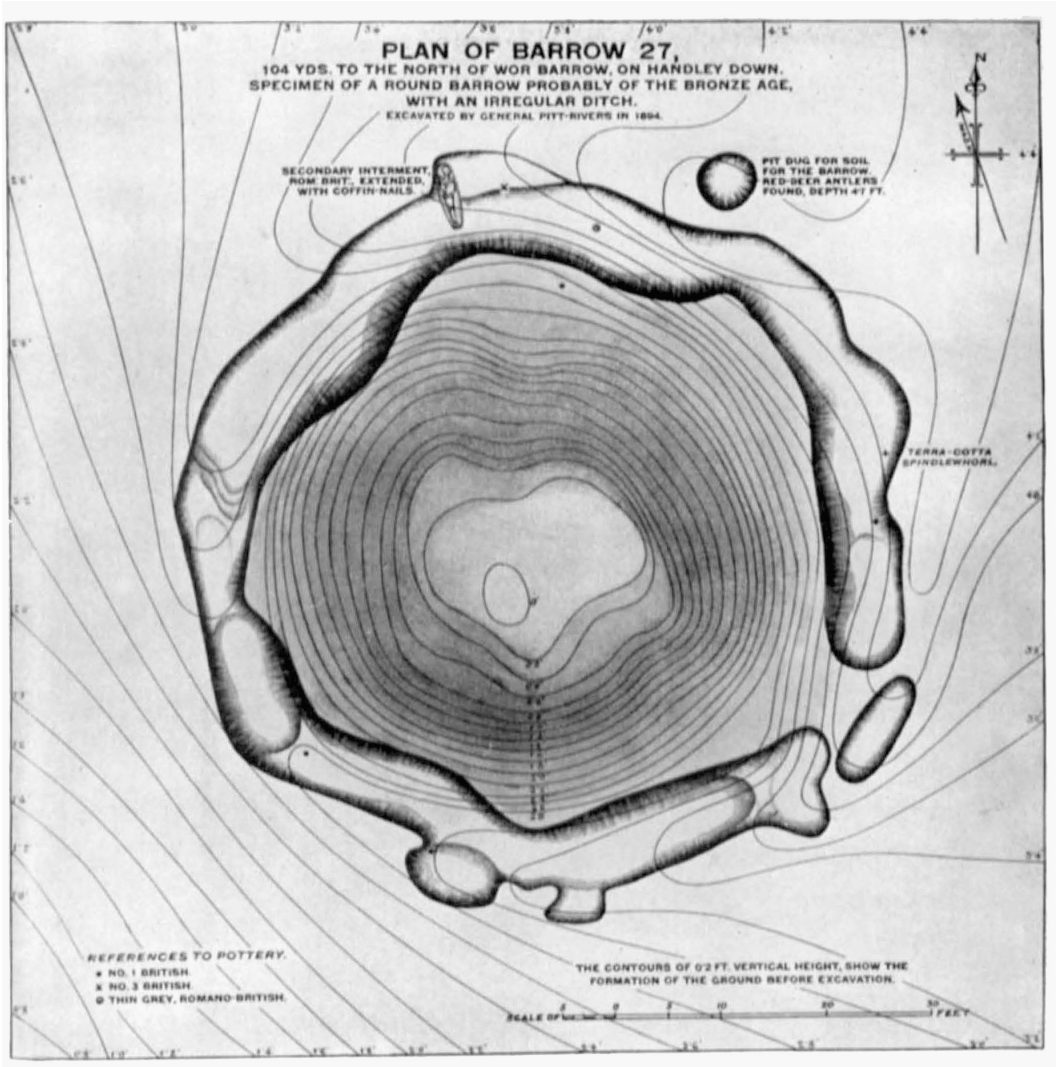


Figure 11: Plan drawing by General Pitt-Rivers from his Excavations in Cranborne Chase, reproduced from Stuart Piggott's *Archaeological Draughtsmanship: Principles and Practice* (Piggott 1965, plate XXXIV).

In addition, Pitt-Rivers sometimes abandoned detailed stratigraphic section drawings for an 'average section', where finds were distributed through a general profile, which further hindered his attempts at establishing dating chronologies (Bowden 1991, 128).

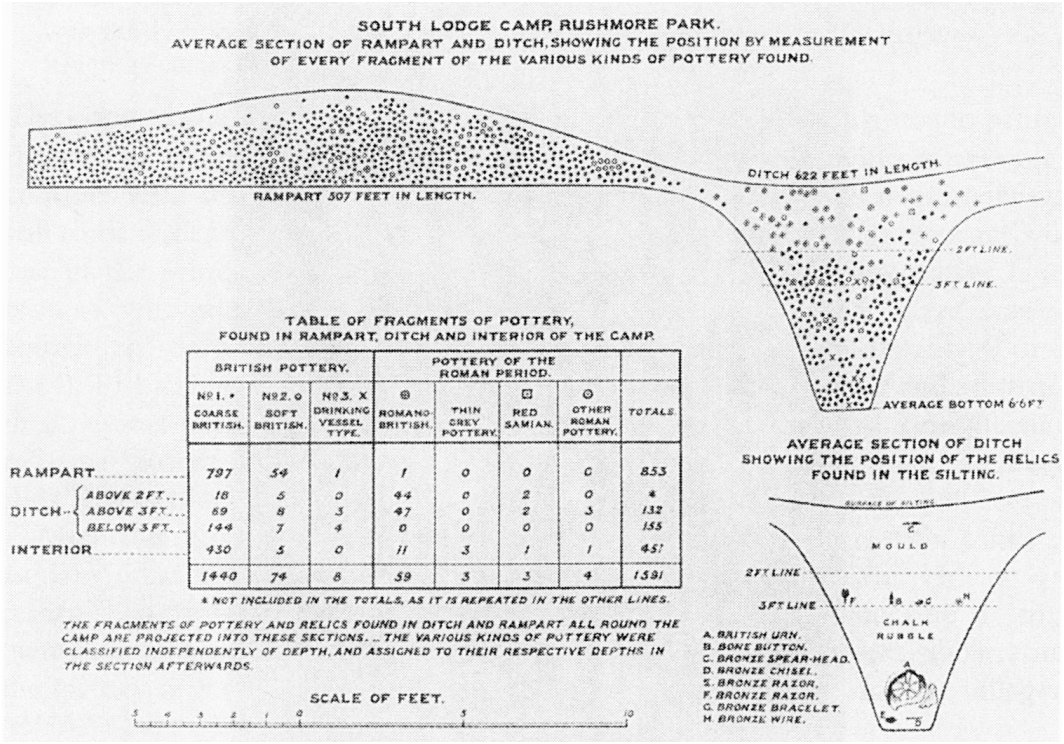


Figure 12: Section drawing by Pitt-Rivers, showing the use of 'average sections', reproduced from Mark Bowden's *Pitt Rivers* (Bowden 1991, 128).

3.2.3 Field drawing in the 20th century

It is often said that Pitt-Rivers was ahead of his time, and this is also clearly evident in the later efforts of Heywood Sumner. Like Stukeley and the Romanticists before him, Sumner was influenced by current thinking in art, and incorporated elements from the Art Nouveau movement into his field recording. Piggott describes the 'danger that lay in wait for those who could be adversely affected by the quality in Sumner's drawings that trod the tightrope between apt decoration and arty-crafty awfulness' (1965, 175). Sumner's plan of Hambledon Hill is a gracefully noisy attempt at bringing together plan views from different parts of the monument at different scales, along with multiple section drawings of the earthworks taking up any spare space on the page. The hachure lines are more effective in showing the subtleties in the complex earthworks, but Pitt-Rivers would likely have had something to say about the use of connective cursive and the floral motif used in the cardinal points.

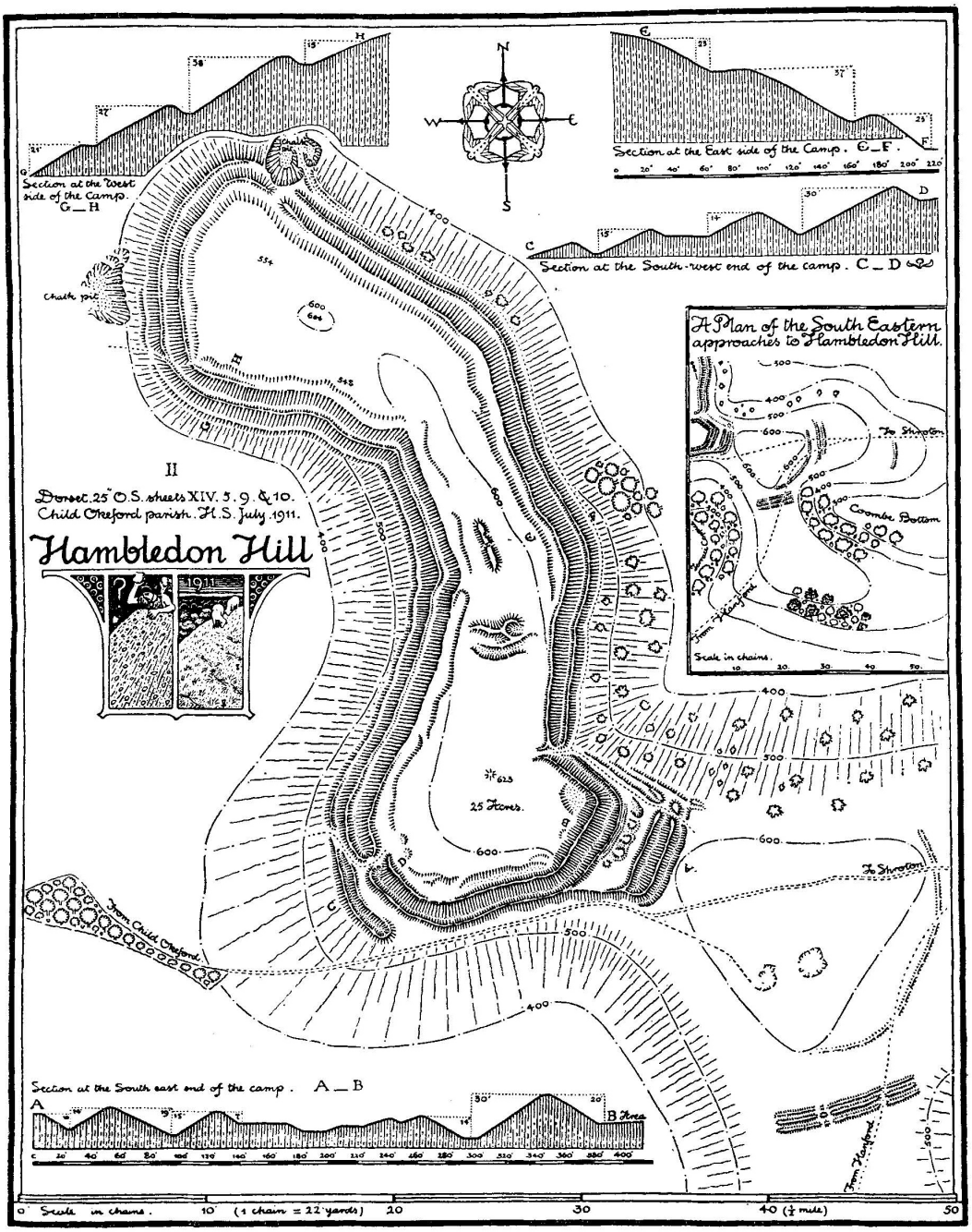


Figure 13: Plan drawing by Heywood Sumner of Hambledon Hill, reproduced from Stuart Piggott's *Archaeological Draughtsmanship: Principles and Practice* (Piggott 1965, 173).

In contrast, Robert Gurd, whose work was roughly contemporary to Sumner's, was able to include some of the artistic conventions of the day, but with more readable results. Far less well known, and working almost entirely in Sussex, Gurd was a railway draughtsman used to taking complex and heterogenous information and presenting it clearly. While lacking the visual path to guide the

eye that is usually part of visual art training, his plans are still exemplars of good communication:

> His maps, often very busy with detail, are always well balanced, calm and easy to read; the hierarchy of information is good. Plan, title and annotation, scale and frame never compete for attention...Yet amongst his contemporaries, and tragically even today, we can find examples of maps and field plans which are difficult to 'read' or interpret because they lack balance (Goddard 2000, 8).

Gurd was best known for his pottery illustrations and is the uncredited artist for around half of the pottery illustrations for Mortimer Wheeler's excavation report for Maiden Castle. Gurd died unexpectedly during the preparation of the report, and Seán Goddard believes the rest of the illustrations were taken over by the report author, Stuart Piggott. Goddard also believes that Gurd was a great influence on Piggott (2000, 12), who went on to influence an entire generation of archaeologists. Piggott's own contribution to field drawing is entirely neglected in his 1965 history of archaeological draughtsmanship.

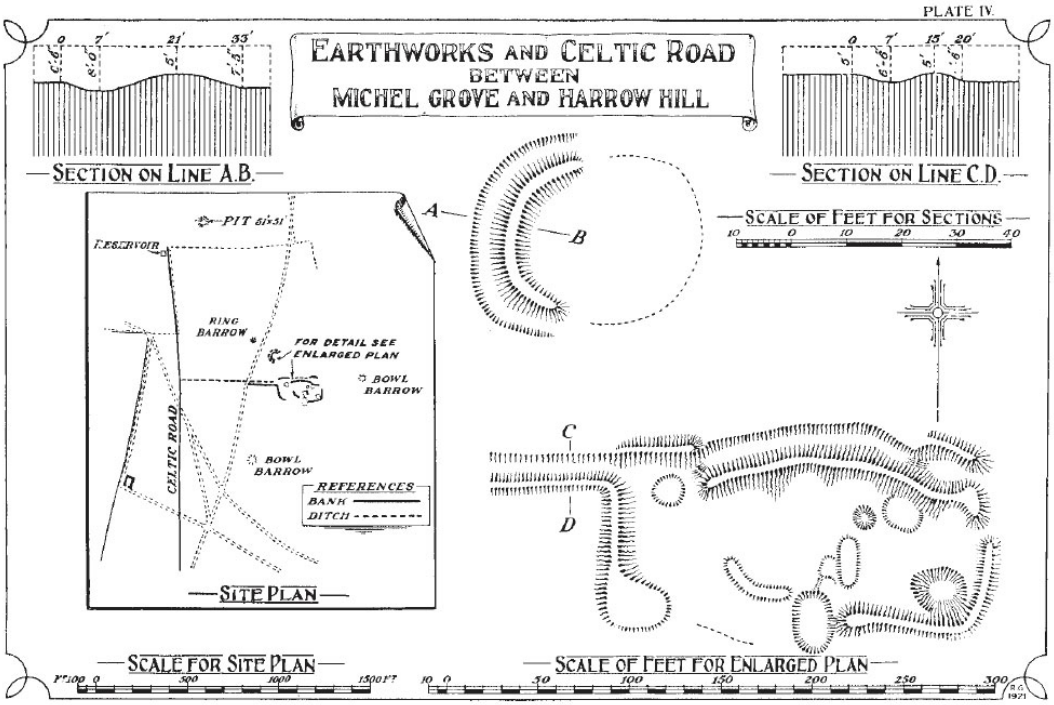


Figure 14: Plan drawing of earthworks by Robert Gurd. Reproduced from *The importance of illustration in archaeology and the exemplary work of Robert Gurd* by Seán Goddard (2000, 8).

Perhaps modesty forbade it, but Piggott's own work shows a wonderful synthesis and understanding of the artists with whom he worked, like Sumner and Gurd (though visually he seems to have taken his cues more from Gurd than Sumner), and his careful study of the work of Aubrey, Stukeley and Pitt-Rivers. He retains the arts and crafts feel in his lettering, and his work manages to be highly detailed and precise, while maintaining Gurd's 'well balanced, calm and easy to read' (Goddard 2000, 8) aesthetic. His plan of the cairn at Cairnpapple Hill surpasses them all (Piggott 1947-8, 82). The information is complex, but the annotation never obscures the plan itself, it is visually easy to separate the cairn and kerbstones from the larger henge, and the above ground features from the areas where excavation took place.

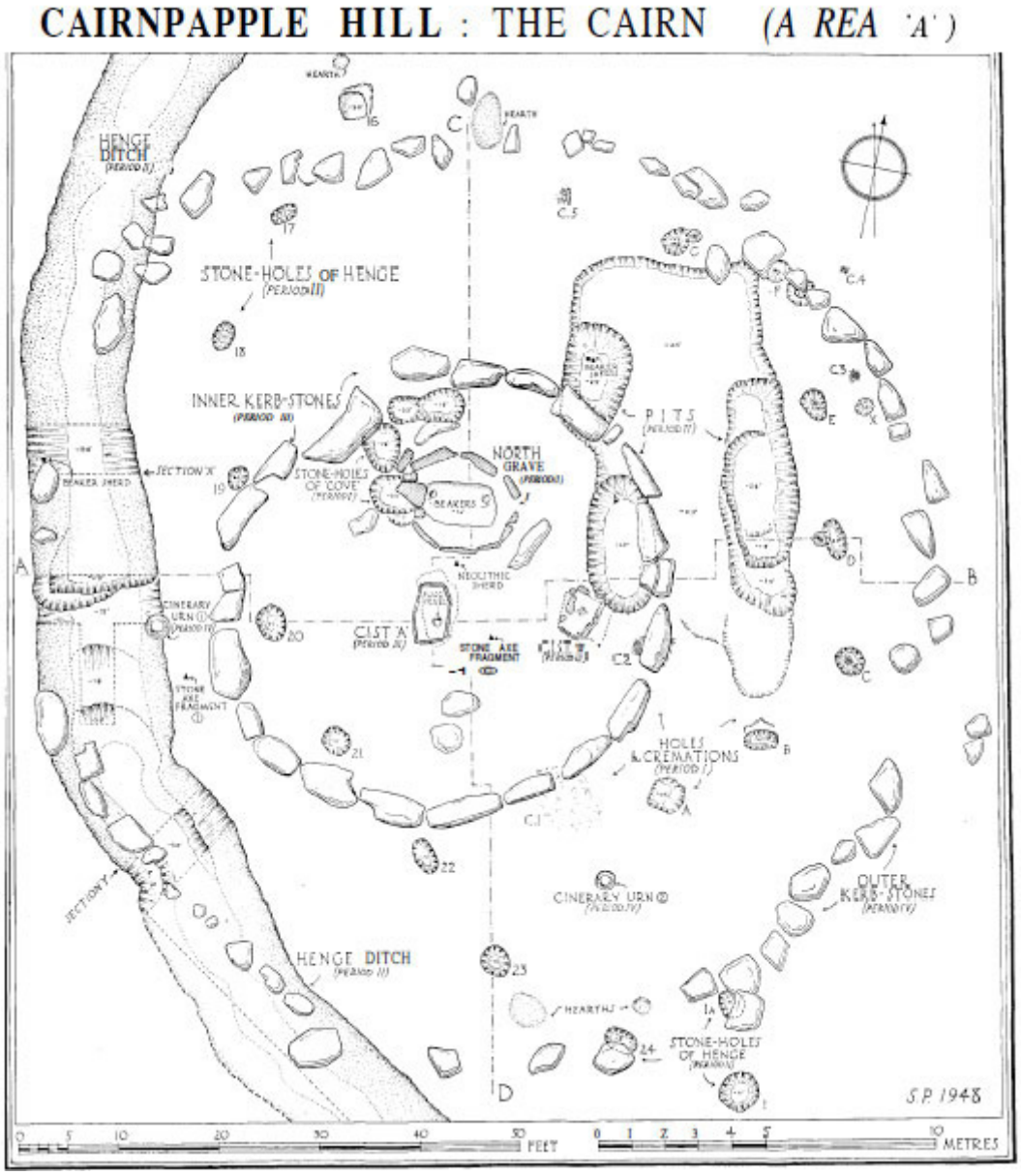


Figure 15: Plan drawing of the cairn at Cairnpapple Hill. Reproduced from *The excavations at Cairnpapple Hill, West Lothian* by Stuart Piggott (1947-8, 82).

Clearly, the 20th century was producing artists, draughtsman and archaeologists all forming a modern understanding of field recording, but the image those writing about the history of the field seem to hold as the standard of perfection, combining both clarity and aesthetic, is Mortimer Wheeler's 'Section from Segontium' from 1922. Lesley and Roy Adkins and Stuart Piggott agreed on the importance of this single drawing:

> The modern approach to illustration, with the conscious realisation that the purpose of the illustration is to convey not only information but also an interpretation of that information... irrespective of differing styles and approaches, the best archaeological illustrations have been based on the principles so clearly demonstrated in that drawing (Adkins and Adkins 1989, 5).

> It stands for, and was among the most immediate founders of a tradition which British archaeological draughtsmen have in the main followed since the 1920s. It was a statement of a new code, a relational model presenting the excavator's interpretation clearly and unhesitantly; the sentence spoken with inflexions of authority; the drawing of a man who had made up his mind (Piggott 1965, 175).

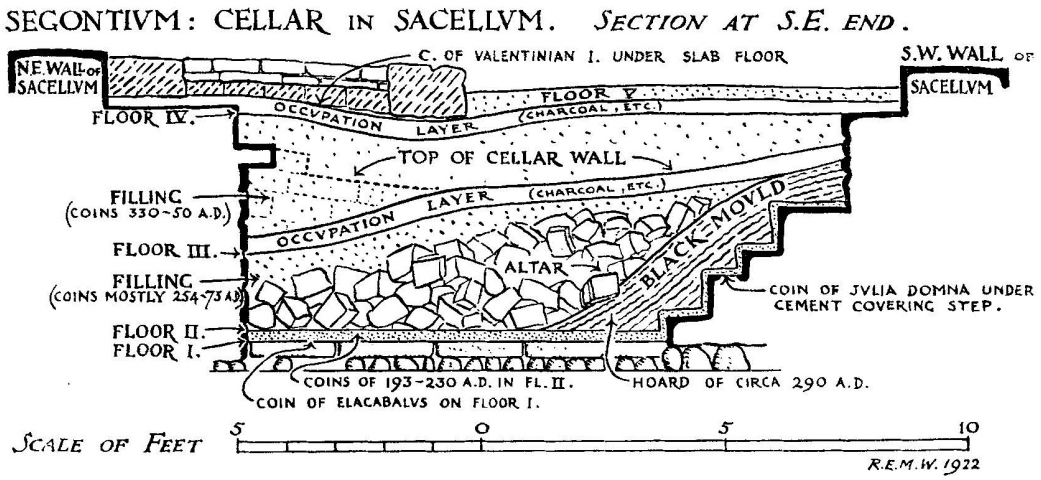


Figure 16: Drawing by Mortimer Wheeler of a section across the cellar in Sacellum at Segontium. Reproduced from Lesley and Roy Adkins' *Archaeological Illustration* (1989, 6).

Edward Harris comments on Piggott's aesthetic evaluation with a further practical explanation: 'it may be suggested that the drawing broke with tradition in having the interfaces between strata properly defined...he also began to number the layers of soil in sections and in the records, which was definitely a landmark decision' (Harris 1989, 11).

The statement by Adkins and Adkins in 1989, that a drawing created in 1922 still epitomised the modern approach to archaeological illustration, is extraordinary. At a time when the discipline was entering the digital age, Wheeler's effort was just as powerful as when it was first published. Brian Hope-Smith sums up the vital substance, to which we should all aspire: 'The crucial point, in reality is not how the line is to be drawn, but where it is to be drawn, and the success or failure of a drawing is actually determined before pen is set to paper. There is but one simple qualification for success, and that is precise understanding of the idea or the image to be transmitted...Once the vital process of thought is complete, there is no difficulty in placing lines in the right places' (Hope-Taylor 1966, 107-8).

3.3 Modern field drawing

Archaeological field drawing today has become a more systematic process, with generally agreed upon methods and developed concepts of good practise. Field drawing as currently undertaken can be said to be part of a larger system of field recording. Field recording also includes other types of data capture, as in surveying a site's surface characteristics, gathering remote sensing data, obtaining photographs to be used in photogrammetry, or taking measurements from a standing structure. All are different methods for recording the important information about a site, so that it may be well understood during its post-excavation analysis.

Field drawing differs from other forms of recording in both intention and execution. With the exception of excavation photography (or perhaps building recording for a structure about to be removed), field drawing is the only form of field recording used expressly as an interpretive replacement for the archaeological record as it is destroyed. Field drawing is also the last form of recording where much of the work is still done by hand. While there are experimental attempts to change this, the forms of field drawing for which there are still no adequate automated substitutes, are plan and section drawing (Rains

2007, 2). The tools used to create these drawings have changed little since the advent of field drawing itself. The introduction of waterproof and dimensionally stable drafting films like *Permatrace* have made field drawings more durable and long-lasting (Adkins and Adkins 1989, 11; Hawker 2001, 47), and the use of calculators and electronic distance measurers (EDMs) to aid the recording of distances and levels (Hawker 2001, 46), are the newer technologies currently used to create them. Even so, the primary tools are still a pencil, something with which to erase it, something to sharpen it, and something upon which to apply it.

3.3.1 Drawing conventions

All field drawings use some sort of drawing convention, which consists of the agreed upon standards and formatting used across all drawings in a particular project; what Helen Wickstead refers to as 'Collective Drawing' (2008, 21). This can include everything from basic information like the site code, the drawing number, the style of north arrow, drawing scale, names of those involved in creating and checking the drawing, and grid referencing. For the drawing itself, conventions must be established to show the edges of the unit. Since the edges of a unit can be actual, arbitrary (when intrusions to the unit are present), or uncertain (when the edges are difficult to establish), it is necessary to use conventions to communicate these differences consistently (Roskams 2001, 135-6). Within a unit drawing, conventions must be used to show the interior surface of the unit so they may be interpreted in the same way across the site. Conventions vary according to the preferences of different archaeological field units and supervisors, but as long as the conventions are consistent (and used correctly), they can be interpreted properly during post-excavation.

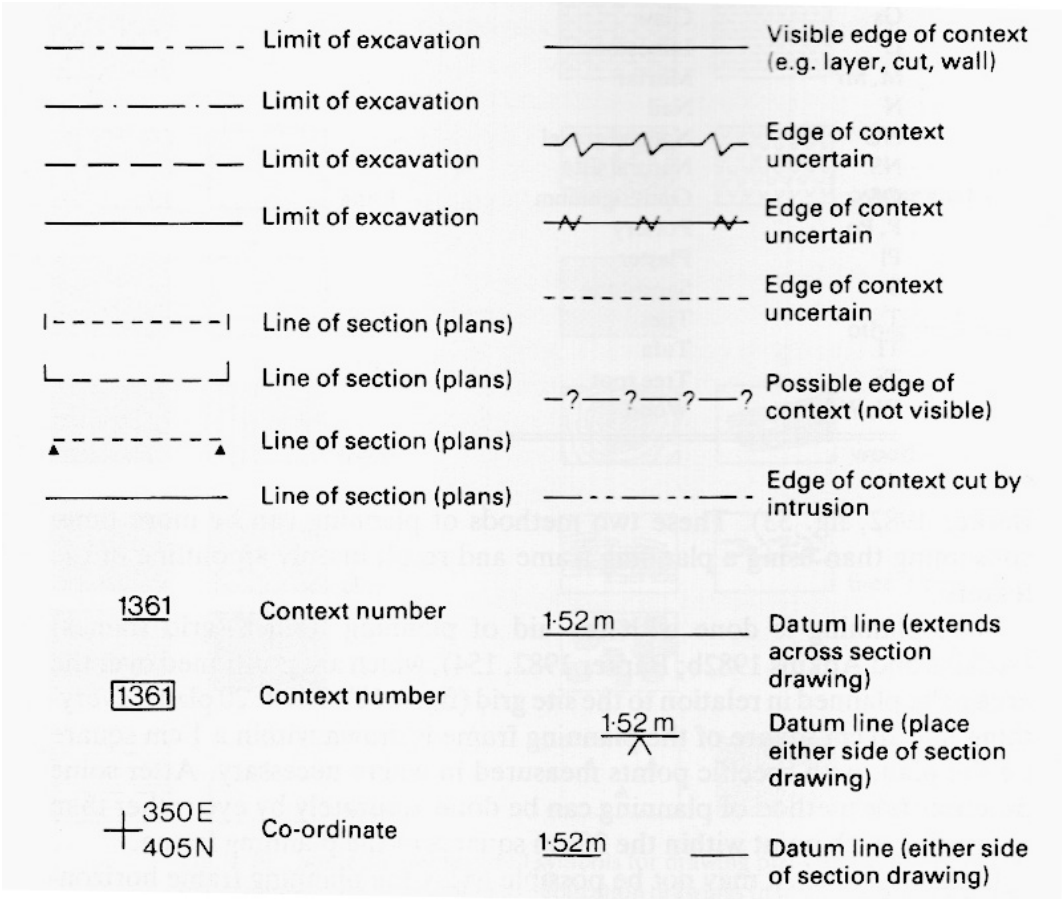


Figure 17: Examples of common drawing conventions used to illustrate the limit of an excavation, the edge of a context or the line of a section, including ways to set off context numbers, co-ordinates or a datum line. Reproduced from *Archaeological Illustration* by Lesley Adkins and Roy Adkins (1989, 76).

In order to show common materials and inclusions found within an archaeological unit, a drawing convention will usually employ a system of representative symbols. These symbols represent things like soil types, inclusions of charcoal or mortar, stones of a particular size, or distinct areas of multiple artifacts or ecofacts, such as shells or potsherds. In the field, the symbols used will likely be a simplified shorthand version of what might be seen in a post-excavation field drawing produced for publication, but the idea is the same. Because the representative symbols chosen are not the same for every project, a key is always necessary to explain the conventions used in any drawings prepared for dissemination (English Heritage 2007b, 31).

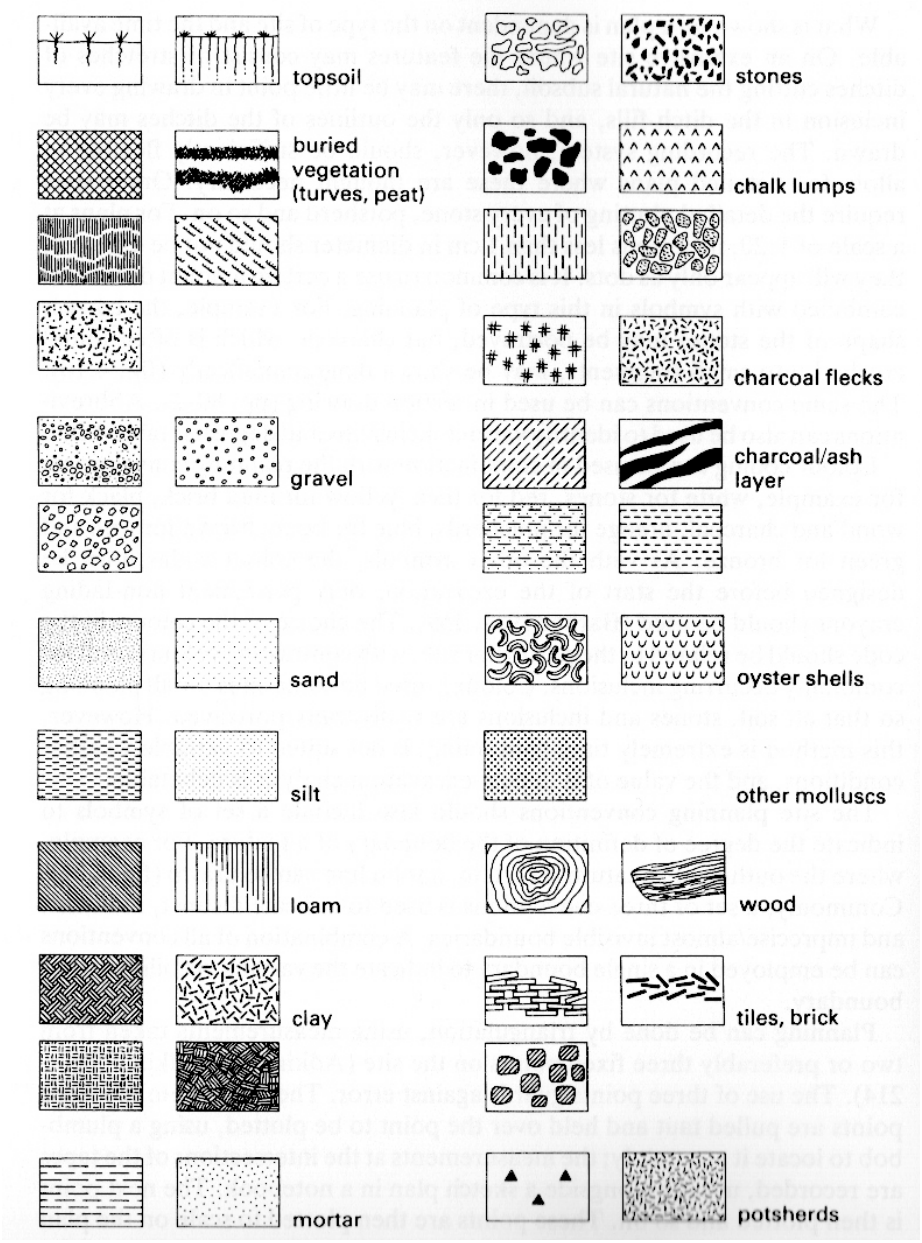


Figure 18: Examples of common symbols used to illustrate different materials found within archaeological units. A more simplified version would likely be used during field recording. Reproduced from *Archaeological Illustration* by Lesley Adkins and Roy A. Adkins (1989, 74).

In addition to drawing conventions using various symbols to communicate what is seen on the surface of a plan or section drawing, are the more formalised systems of showing slope in archaeological field drawings. These have been most commonly implemented using one of two different techniques; either hachures or contour lines. Both are cartographic techniques to show three-dimensional slope in two dimensions adapted for use in archaeology. Contour lines are still

used, but at the site or unit level, hachures show the direction of slope without ambiguity, and are therefore considered preferable by most archaeologists (Adkins and Adkins 1989, 79). The ambiguity of slope associated with contour lines can be seen in the the plan drawing by General Pitt-Rivers from his Excavations in Cranborne Chase (Figure 11), while the hachures incorporated into Stuart Piggott's drawing of the cairn at Cairnpapple Hill (Figure 15), leave no doubt as to the direction of slope in both the earthworks and sections.

The cartographic convention of hachures was developed in the 19th and 20th centuries, and originated from the practise of using shading to show slope (Imhof 2007, 10). Hachures are most commonly seen in plan drawing, but they are also used to show the slope of a section drawing, when illustrating it in plan view. Hachures show the direction of slope by placing a triangle at the top of the slope. The length of the hachure correlates to the length of the slope, and the steepness of the slope is indicated by how closely the hachures are grouped together (Hawker 2001, 18). While hachures are rarely used in traditional cartography today, they are still a staple in archaeological field recording. In their recent guide to good recording practice for archaeological landscapes, English Heritage state 'The hachured plan remains the most effective means of depicting earthworks, Even if plans are simplified for wider dissemination, the hachured earthwork plan is still the basis for the archival record.' (English Heritage 2007b, 15).

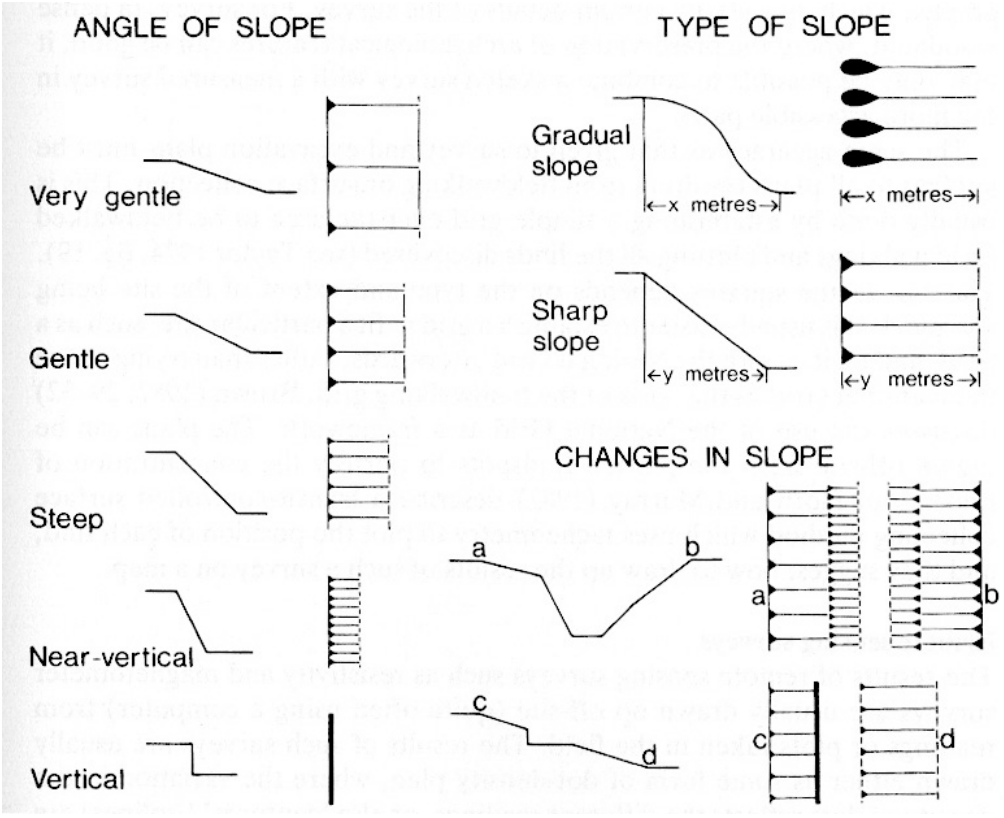


Figure 19: The hachure system for illustrating slope in two-dimensions, as commonly used in archaeology. Reproduced from *Archaeological Illustration* by Lesley Adkins and Roy A. Adkins (1989, 67).

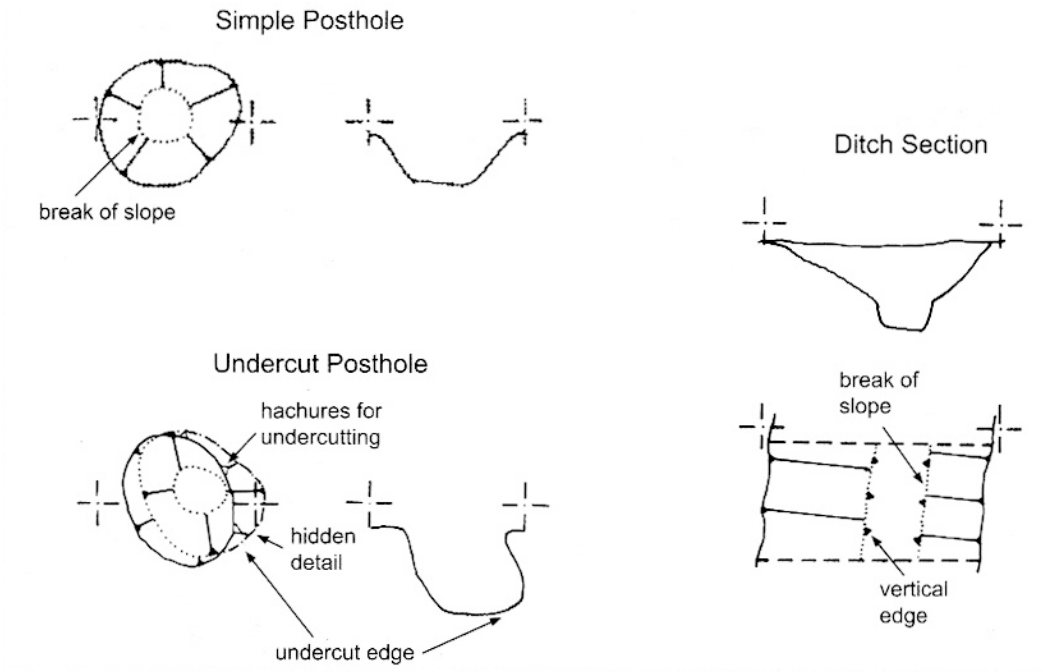


Figure 20: Examples of the hachure system showing its use in both plan and section. Reproduced from *A Manual of Archaeological Field Drawing* by J. M. Hawker (2001, 18).

3.3.2 Plan drawing

Field recording of the archaeological resource during excavation, in plan view, is done in a variety of ways. Steve Roskams lists three primary types of site planning in current use by archaeologists, including top planning (or single-level planning), phase planning and single-context planning (2001, 137). Top planning simply records everything seen in the unit at an arbitrary level (the unit is dug and then planned in five centimetre levels, as an example), irrespective of what the unit contains. It is rather like taking a snapshot of the unit, and then determining the relationships during post-excavation. Single-context planning records each context separately, so that information for that context is grouped together and the relationships established during post-excavation. Once a context is defined, it is then planned and excavated.

> The primary route to an understanding of the activity represented in the archaeological record is through the 'stratigraphic sequence'*...(Any single action, whether it leaves a positive or negative record within the sequence, is known as a 'context').* Within any such sequence the chronologically earliest context will always be found to be 'sealed' or 'cut' by a chronologically later context. Chronology in this sense refers to the relative date of activity between one context and another (Museum of London Archaeology Service 1994, 3).

What is particularly interesting about single context recording with regard to the Semantic Web, is that a context is meant to represent *an activity* which has formed the archaeological resource. By focussing on an action, a conceptual correlation can be made between the way a Semantic Web ontology and context recording is structured. For example, if context number four (an Anglo-Saxon hearth), is cut by context number three (a post-hole from a later structure) then the relationship can be easily expressed in a manner conceptually translatable into RDF triples:

Subject	Predicate	Object
context four	is cut by	context three
context three	cuts	context four
context three	is later than	context four
context four	is earlier than	context three

Both top and single-context planning are examples of field recording which create *primary data,* or data which attempts to simply record with as little interpretative input as possible. Primary data leaves interpretive decisions to be made during post-excavation, when the full scope of the information is available. Primary data is also important for future archaeologists, who may wish to pose different questions and create their own interpretations. In contrast, phase planning requires interpretive decisions be carried out during the recording process. Features within the unit thought to be part of the same time period are drawn together while the excavation is still in process, and thus does not create primary data. While there is disagreement about the extent to which phase planning is still in use, it is generally felt it should be an augmentory interpretive tool, and never a sole means of field recording (Adkins and Adkins 1989, 76-8; Roskams 2001, 139-40).

Advocates of top planning cite it as being less complicated, and that single-context information can be gleaned from the drawings later, if found to be necessary (Adkins and Adkins 1989, 77), but Edward Harris refutes this:

> By imposing the arbitrary strategy of excavation on sites with clear stratification, archaeologists destroy the primary data they seek, the very data they are supposedly best qualified to obtain. By using arbitrary levels, artefacts are removed from their natural context and mixed with objects from other strata, as the arbitrary level does not respect the natural divisions between the units of stratification on a site...There are some who reckon that the topography and character of stratification can be reconstructed from records made by arbitrary excavation [but],...

> the impossibility of such reconstructions is probably the rule, rather than the exception. Finally, the arbitrary strategy results in the creation of an arbitrary 'stratigraphic sequence' for a site (Harris 1989, 20).

In addition, single-context planning is seen as the only way to properly understand complex and deeply stratified sites; where understanding the spatial relationships between different contexts is nearly impossible when potentially unrelated information is lumped together in the same drawing (Roskams 2001, 140-1). As single-context recording is meant to document an archaeological *activity,* which is conceptually similar to the RDF triple, and is the form of planning most likely to produce primary data, sites recorded using this system seem ideal for incorporation into the Semantic Web.

3.3.3 Section drawing

The creation and recording of sections was the traditional means of establishing the stratigraphy of a site since they were first adapted from the discipline of Geology in the 19th century. Because sections are only placed in strategic places across a site, the information they give is never meant to be comprehensive. As such, they can lead to incorrect assumptions, especially for very complicated sites (Clark 1993, 281). The use of running sections could be used to help build the overall stratigraphic picture of a site, but until the introduction of the Harris Matrix in the 1970s, there was no systematic means of stratigraphic recording.

During the 1960s, Philip Barker's advocacy of the open-area system (which defines a site to be excavated from an open, horizontal view), meant archaeologists were no longer creating the large baulks associated with older forms of recording, and therefore large vertical planes. The horizontal nature of the open-area method made possible the creation of single-context recording, which in turn caused the usefulness of sections for defining stratigraphy to be called into question. Stratigraphic relationships can be established using single-

context recording in conjunction with the Harris Matrix alone. The single-context plans can then be overlaid to establish stratigraphic relationships, and then composite sections can be created during post-excavation analysis (Harris *et al.* 1993, 1-6).

For some, these new methods have sounded the death knell for sections as a tool for stratigraphic analysis, but the creator of the Harris Matrix disagrees:

> There are those who would advocate that sections are now obsolete, but sections have a purpose which cannot be met by any other means. Natural cross-sections give 'the third dimension of the land form, the other two being furnished by the map' (Grabau 1960, 1117). While there is little doubt that archaeological stratigraphy in the past has placed too much emphasis on sections, the reaction to this overbalance should not be to abolish sections. Their use should be brought into line with other stratigraphic methods, such as written records and plans (Harris 1989, 72).

Though on the forefront of implementing the single-context and Harris Matrix methods (Spence 1993, 23-46), the Archaeological Site Manual for the Museum of London Archaeology Service still contains guidance for making section/ elevation drawings (Museum of London Archaeology Service 1994, 12), as does the current Manual of Archaeological Field Drawing, published by RESCUE — The British Archaeological Trust (Hawker 2001, 31-6). Roskams advocates the use of sections in modern practice in two instances however:

> [to] give information on the internal configuration of a particular deposit, for example to throw light on formation and transformation processes within the silting in a ditch, or on the relationship between units, for instance by recording the

> character of the interface between successive layers [and] to solve specific stratigraphic problems on the site, for example the relationship between two inter-cutting pits, or between a trench-built wall and adjacent strata' (Roskams 2001, 144).

Despite the fact that under current practice, most section drawings are an analytical tool, and are therefore not typically considered primary data, provisions will still need to be made to incorporate them into the Semantic Web. The use of single context recording should allow spatial relationships to be defined within the structure of an RDF triple, whether it expresses a relationship which horizontal or vertical.

3.4 Field drawing goes digital

People have been applying digital technology to archaeological questions for nearly as long as long as digital technology has been available. The results of experiments in archaeological computing began to appear in publications in the early 1960s, both in traditional archaeology venues for and computer science, although Richards and Ryan cite a few even earlier examples from the 1950s (1985, 4). Robert Chenhall of the Department of Anthropology at Arizona State University, began publishing the Newsletter of Computer Archaeology as early as 1965 (Cowgill 1967, 335), and was an example of a researcher within the archaeological community turning their attention to computers as a tool for a more scientific archaeology (Chenhall 1968, 21).

The same cross-disciplinary interest was happening in the reverse as well. In the late 1960s, several individuals working in computer science within UK universities (primarily at Birmingham) including Sue Laflin, Jim Duran and John Wilcock, went on to found the Computer Applications in Archaeology (CAA) Association in 1973 (Julian Richards pers. comm. June 2010), and the annual CAA conference went on to become the most prominent international meeting for

archaeological computing (Lock and Brown 2000, 2). While initially held only in the UK and organised by British universities, the conference made its international debut at the University of Aarhus in Denmark in 1982 (CAA 2010). CAA stayed within Europe until 2006, when it was hosted for the first time in North America, in Fargo, North Dakota. Now becoming truly international, CAA held its first meeting in Asia with the 39th meeting in Beijing, China in 2011, and plans are underway for CAA 2013 in Perth, Australia. In addition, national associations have been formed in many countries. While not the only venue for the blending of computer science and archaeology (for example, the International Symposium on Virtual Reality, Archaeology and Cultural Heritage or VAST, has been held annually since 1999), the history of CAA is a useful way to illustrate the steady growth of the applications of digital technology to archaeology generally.

Whether due to the legacy of Processualism, or the interdisciplinary nature of archaeology, the development of digital technology has seen archaeological application soon after. Despite this consistently early adoption, archaeologists have always been forced to adapt technologies developed for other purposes to do their work (Richards 1998, 331). Digital technologies are no exception, and digital field drawing is a perfect example of the resourceful way archaeologists incorporate tools developed for other established industries into tools to do their work.

The adoption of digital technologies within the discipline of archaeology is not straightforward, and to regard it as simply a new set of tools represents the first of what Ezra Zubrow calls:

> ...two distinct and ultimately contradictory views. The first view is digital developments are essentially methodological. They provide a set of tools, similar to any other set of tools in the archaeological tool kit for solving problems that are generated by a variety of theoretical or narrative concerns... Many would see these techniques as being ‘a-theoretical’

> even 'anti-theoretical.' Although there may be underlying 'theoretical' assumptions, the techniques are universal and may be used by any theoretical position...The second view is digital developments create or at least influence the creation of theory in many ways. The digital domain emphasizes the very large and the very small and makes possible a re-emphasis on the individual as the primary actor. Indeed, if one believes that it reconstructs human mental processes it may be a proxy for theory itself (Zubrow 2006, 11).

Both views have implications, which are overarching and subtle, and should be considered when using any sort of digital technology. In the case of field drawing, these extend from the capture of primary field data using digital tools, to the processing of field data using digital technologies, and to the way data is communicated when created using digital means or presented in a digital format. Digital technologies may dip in and out of the process of excavation, post-excavation and publication/communication in varying degrees along the way, but its important to think about what impact they have on the conclusions being offered, and the ideas being communicated (Huggett 2004). As such, the translation of data derived from field drawing into Semantic Web technologies is meant to solve problems of interoperability, versatility and accessibility, but this will no doubt result in theoretical issues which will need to be considered.

3.4.1 Digital data capture

Archaeologists have been attempting to drag their computers into the field with them for as long as computers have been moveable (though not necessarily portable!) or accessible via remote connections (Richards and Ryan 1985, 41-2). While the reasons are varied, the motivation is typically a desire to automate, and thereby speed up the process; archaeology being a very time consuming and labour intensive endeavour, or to acquire data not typically available through analogue sources, such as the results of geophysical survey. The use of digital technology in

the field can be a recursive exercise as well, where its use on site informs and alters the way fieldwork is carried out during the course of the investigation (Gaines 1974, 454; Powlesland 1991, 156-7; Rains 2007, 1). True digital data capture, or the gathering of primary archaeological data in digital format, has been applied in all these areas, with varying levels of success and adoption.

The most common way in which digital technology has replaced analogue field drawing, and one of the first technologies to be successfully borrowed from another discipline, is digital survey. This includes Global Positioning Systems (GPS), Electronic Distance Measurement (EDM) and the Total Station Theodolite (TST), which Martin Carver calls 'the queen of the surveyors' toolkit' (2009, 67). While a GPS unit is meant to move and take readings based on the positioning data received from satellites, and is thus a new form of technology, the EDM and TST are digital replacements for analogue predecessors, namely various types of triangulation and measurement using hand-held tapes, the plane table, or the traditional theodolite (Adkins and Adkins 1989, 86-90; Collis 2001, 36-7).

Figure 21: Students working at the site of Burdale in the Wolds of North Yorkshire, UK. The TST is set up at the corner of the site to record small finds and the outlines of features as they appear. Photo by the author.

In all three instances, GPS (which takes readings while the user is in motion, and plots location using satellites), EDM (which measures distance electronically using a laser) and TST (which is has EDM capabilities combined with the theodolite's ability to record horizontal and vertical angles), are recording and storing a series of points. These points can be made singly to plot small finds, grouped linearly to mark boundaries, or grouped polygonally to describe closed features or contours. All create the same thing: vector-based spatial information forming a field drawing which is 'born digital', and therefore considered primary data. Where these technologies have largely replaced their analogue predecessors, digital vector data forms the backbone of the spatial record for an archaeological survey or excavation. While these technologies are common and proven, they are not yet universal.

Implementation of any form of digital data capture is still haphazard at best (at least in the realm of contract archaeology), according to Paul Backhouse, Manager of Graphics and Digital Media for Oxford Archaeology, which is one of the largest archaeology practices in Europe (2006, 52). The early adoption of digital data capture, may still be centred primarily in academia, but he also expresses his desire for the development of pocket computers which can be used for other types of on-site recording. He cites the main reason for the generally poor showing of other types of device, apart from survey equipment, as 'Archaeologists, it seems, cannot be trusted with equipment that use batteries without breaking something - electronic casualty rates are very high' (2006, 53). Attempts at fulfilling his wish have been made however, and they form a second, albeit even more academic type of digital data capture under exploration by archaeologists.

Collecting primary digital data in the field using hand-held computers was largely pioneered by Dominic Powlesland in the early 1980s for the Heslerton Parish Project (Powlesland 1986, 39). Powlesland began by using the Sharp PC-1500 in 1984 to record primary context and object data, and was able to keep them in use for object recording until 1996 (Powlesland *et al.* 2009), which is a testament

to their ruggedness and usability. Printouts were included in the site notebooks, alongside other traditional forms of recording via a battery powered docking printer (Powlesland 1991, 165).

Figure 22: First manufactured in 1982, the Sharp PC-1500 Pocket Computer, and its four-colour printer dock, was likely the first handheld computer used for digital field for recording in archaeology. Photos from the Pocket Computer Museum (Laroche 2010). Reproduced with permission.

Data about the spatial characteristics of the Heslerton Parish Project was recorded using pocket computers in coded text format, for use with the project's custom databases for context and object data, but the information was not spatial data *per se*. An EDM was used to record vector-based spatial data about the site, which was then incorporated into a 3D database, but traditional drawing methods were employed for field drawing. These drawings were later digitised in CAD during post excavation (Powlesland 1991, 164-7).

In the late 1990s, Nick Ryan at the University of Kent, also began exploring handheld devices for collecting primary archaeological field data, but from a different perspective. Ryan saw the potential of using GPS equipment, attached to a handheld device like the Apple Newton, as a way to gather more comprehensive

information to be incorporated into a Geographical Information System (GIS). He called the system Mobile Computing in a Fieldwork Environment (MCFE) using bespoke software called FieldNote. The MCFE project was meant to be a recording system which is *context-aware* or has 'the ability of the computer to sense and act upon information about its environment, such as location, time, temperature or user identity. This can be used not only to tag information as it is collected in the field, but also to enable selective responses such as triggering alarms or retrieving information relevant to the task at hand' (Ryan *et al.* 1998, 18). Ryan used the ability of a GPS within his MCFE system to start recording vector data as the user moves through their environment, and possibly vector-based sketches with a stylus, which combined with additional data entered into or captured by FieldNote, to quickly create detailed information about the environment. The information could then be downloaded into a desktop GIS for further use.

Figure 23: Left: Screenshot of the capture of vector-based spatial data within FieldNote on the Apple Newton. Right: Nick Ryan testing the MCFE system in the field, wearing a GPS device with a hat antenna, attached to an Apple Newton handheld computer. Reproduced from *FieldNote: extending a GIS into the field* (Ryan *et al.* 1999).

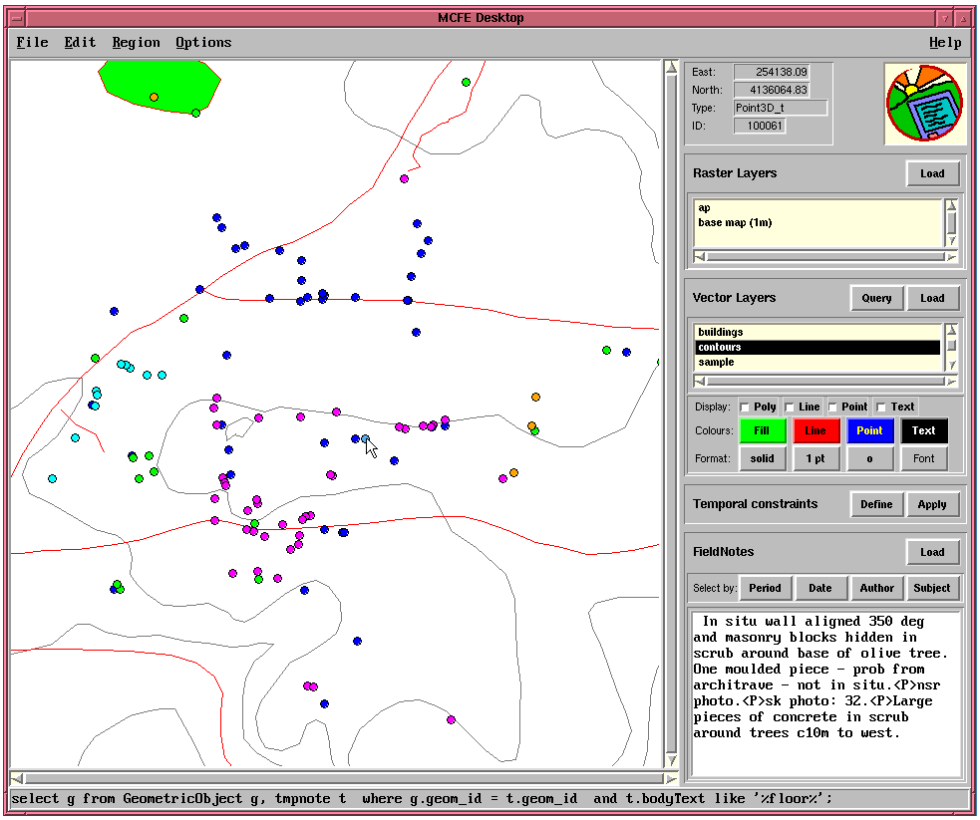


Figure 24: Screenshot of vector data in the FieldNote system downloaded from the Newton. Reproduced from *FieldNote: extending a GIS into the field* (Ryan *et al.* 1999).

Dominic Powlesland and his research group, now called the Landscape Research Centre (LRC) have continued to move forward with handheld devices as well. Undertaken in 2001 with Keith May of English Heritage, the *DigIt* project was both a traditional exploration and excavation of Anglo-Saxon settlement in North Yorkshire, and an opportunity to experiment with new technology for digital data capture. They chose to use the Handspring PDA made by Palm. It represented the first time the LRC used a consumer operating system rather than their own, but otherwise they were essentially using handheld devices for the same purpose as with previous projects. Importantly however, they did the first experimentation with vector-based field drawing, using a piece of plug-in hardware called the Seiko Smartpad. Unfortunately, they were disappointed by the technology on several levels, the most difficult to understand of which was a sonic pen and digitising surface which captured the data in vector format, but only allowed bitmap output, thereby negating much of its usefulness (Powlesland *et al.* 2009).

Figure 25: Excavation work carried out by the Landscape Research Centre in the area between the villages of Sherburn and East Heslerton in North Yorkshire, UK, showing the digital recording of vector-based primary spatial data. Tom Cromwell (English Heritage) is shown using a reflectorless TST, and Keith May (English Heritage) is using a Handspring PDA by Palm. Reproduced from *DigIT: Archaeological Summary Report and Experiments in Digital Recording in the Field* (Powlesland *et al.* 2009).

Nick Ryan has continued to develop his handheld system as well, with the MCFE project giving way to a new framework called MobiComp, and the FieldNote software becoming FieldMap. FieldMap now has the ability to create vector data both actively and passively. 'FieldMap allows existing notes to be edited and new ones to be created. These may be associated with a single point location, or attached to simple geometric shapes such as lines, circles and polygons. The shapes may be drawn manually on the displayed map, or collected automatically using the GPS data while the user walks over the area of interest' (Ryan and Ghosh 2005, 19-20).

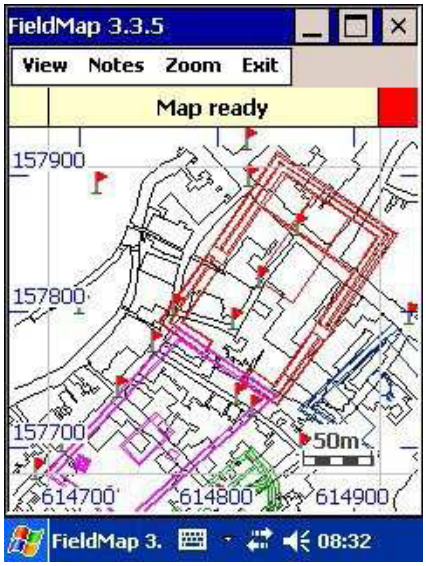


Figure 26: Screenshot of the handheld version of the FieldMap software, showing existing vector data which can be viewed and augmented within a *context-aware* GIS. Reproduced from *Ubiquitous Data Capture for Cultural Heritage Research* (Ryan and Ghosh 2005, 20).

Nick Ryan's work is primarily geared towards GIS applications, and therefore digital data capture at a landscape level, however Mike Rains of the York Archaeological Trust has been working at the site level. Developed through the ongoing collaboration between Rains and the Silchester Town Life Project, based at the University of Reading and headed by Michael Fulford and Amanda Clarke, this work began as a recursive experiment in simultaneous excavation, post-excavation and publication using digital technology (Clarke *et al.* 2002, 402). As the project expanded to accommodate the widely dispersed researchers involved, Rains began experimenting with both handheld devices and tablet PCs to facilitate instant recording of similar types of excavation data as Powlesland and Ryan's projects, which was then immediately available to researchers online.

The difference with Rains' work is he set out to see if it was possible to do real vector-based digital field drawing at the trench level. He created a drawing program using SVG which mimicked the grid of a planning frame, and allowed field drawing in plan view to be dynamically created, and then saved into the Integrated Archaeological Database (IADB) used at Silchester (Rains pers. comm. July 2003). While the tablet PCs were largely disappointing when tested (due to stylus issues and primarily the screens being far too dim even in partial shade), the methodology was sound (Rains 2007, 2).

Figure 27: Left: Mike Rains holding the ruggedised tablet PC tested in the field at Silchester. Though designed for outdoor military use and by far the most expensive of the models tested, the screen was so dim it was virtually useless (Rains pers. comm. July 2006). Right: Working in the Silchester field office with data drawn on a far less expensive, but better performing tablet PC. Photos by the author.

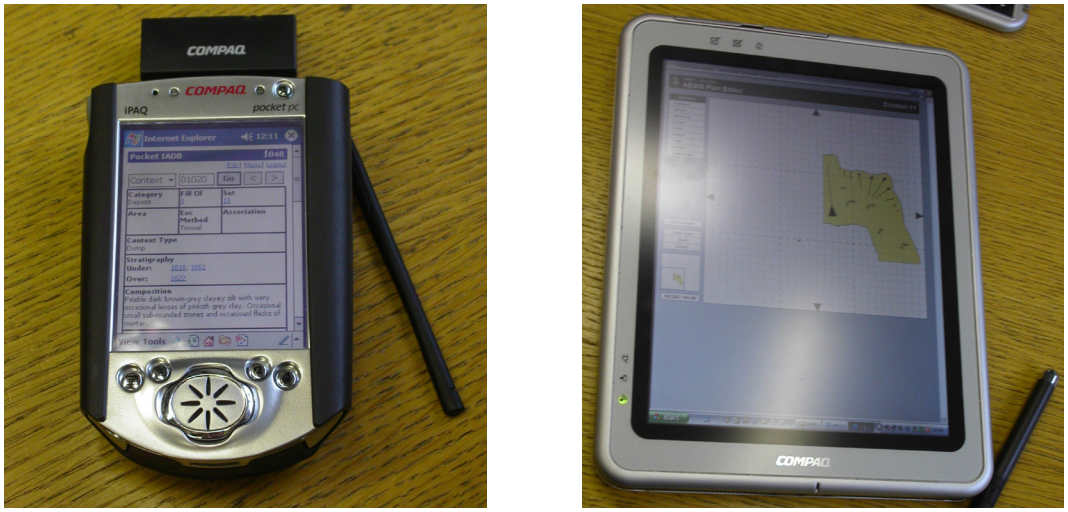


Figure 28: Left: The IADB on a handheld PC, which also has vector drawing capabilities. Right: A tablet PC showing the SVG-based vector drawing program which mimics the grid of a planning frame, and allows for drawings made in plan view to be instantly available within the database as soon as they are saved. Photos by the author.

Experimentation with new technologies is ongoing at Silchester, and most recently they have been trialling digital pens for context recording as part of the Virtual Environments for Research in Archaeology (VERA) project, with good success (Clarke and O'Riordan 2009). True digital field drawing will likely be stalled for the foreseeable future, as archaeology waits for digital display technology to develop sufficiently to allow for work in outdoor conditions. With the rate of

current innovation however, it is unlikely we will have to wait long, keeping in mind as we go forward Dominic Powlesland's adage 'Any ambition to capture all data digitally must be set against the quality, detail or intellectual depth of the data collected' (Powlesland *et al.* 2009).

Most of the history of digital data capture, especially digital field drawing, has been experimental projects using bespoke software and much experimentation with available hardware, but digital field drawing using entirely generic tools received international press recently, with the use of the iPad by a team led by Steven Ellis from the University of Cincinnati working at Pompeii. The Pompeii Archaeological Research Project: Porta Stabia started with six iPads at the site, and interest and uptake by the field crew was positive. Apple computer got wind of the innovative use of their product, and chose to publicise it, which brought massive amounts of attention to the project (Ellis and Wallrodt 2010). The Apple publicity cited the use of iDraw as the vector drawing program in use (Apple Inc. 2010), but during the most recent season the team was using the TouchDraw application made by Elevenworks. Touchdraw allowed the work to be carried out using native SVG, and then exported for use in AutoCAD. The SVG import/export limitations of CAD were compensated for by use of Inkscape and the DWG to SVG Converter by DWG Tools as a way to translate the SVG to DWG (Wallrodt 2011).

Figure 29: Use of the Apple iPad and the iDraw app for primary digital data capture of archaeological field drawing in native vector format. The experimentation was carried out as part of the Pompeii Archaeological Research Project: Porta Stabia. Reproduced from *iPads at Pompeii* (Ellis and Wallrodt 2010).

3D laser scanning is also being used for excavation recording in some instances. The ability to document and analyse a site in the same number of dimensions as the archaeological resource is very appealing, especially if it can be done quickly and easily, and generate native vector data which can be integrated with other types of field recording. Whether used from the air or terrestrially, laser scanning can potentially provide recording which is far more accurate than recording which is drawn by hand, with hundreds or thousands of points being recorded, and far more detailed than points 'shot in' individually with traditional survey equipment. Laser scanning is also considered far less subjective, as it omits the human translation of either the physical act of seeing and drawing, so for those interested in creating primary data which is as objective as possible, it has additional appeal (Pilides *et al.* 2010, 327; Shaw 2007, 40). As the technology becomes less cost-prohibitive, and more archaeologists are trained in using it, the demand will surely grow (English Heritage 2007a, 3).

It is not enough to create a three-dimensional point cloud of an excavation however, as information about colour and texture are vital parts of the recording process, but 3D laser scanning technology is attempting to include this information as well. Scanners can now collect intensity values that show the level of reflectivity, and therefore information about the texture of the surface under excavation. Scanners with the ability to capture RGB values can also gather colour information, which is necessary for seeing contexts which have no three dimensional characteristics (Payne 2011). Whether cost, training and technology will improve sufficiently for 3D laser scanning to become a viable alternative to more traditional forms of field recording remains to be seen, but the possibilities are certainly intriguing.

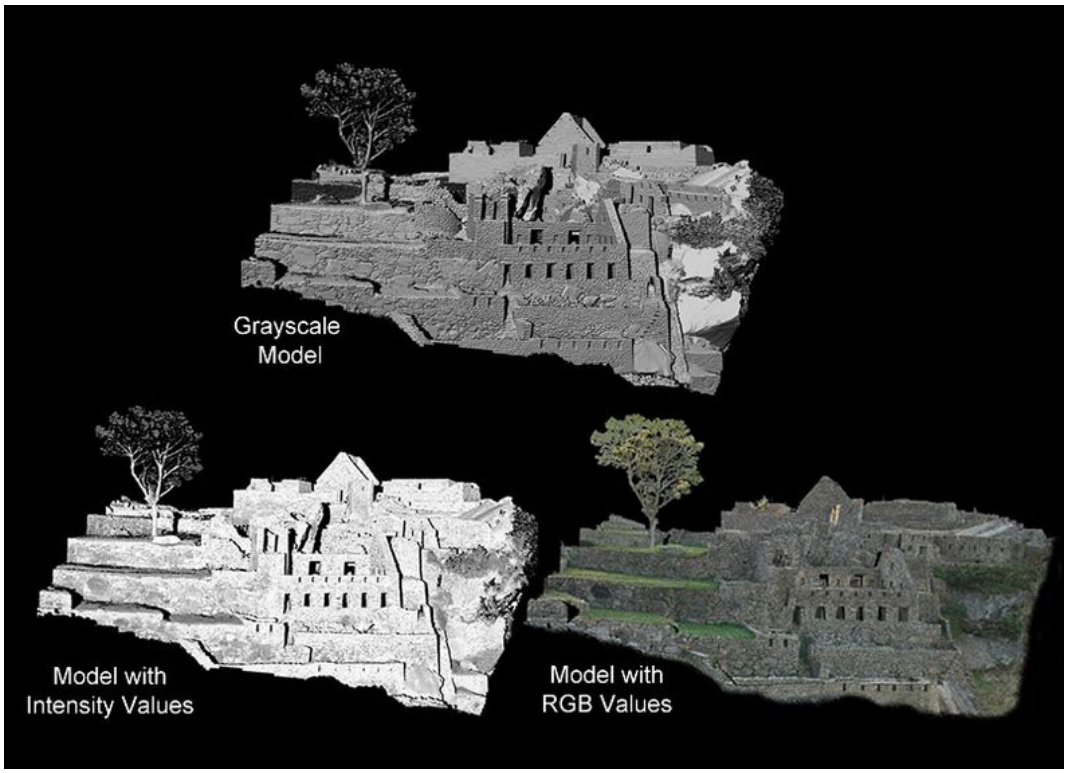


Figure 30: 3D laser scanning doesn't just produce a point cloud and a grayscale model, intensity values which give information about surface texture, and RGB values which reads colour information, can also be captured. Reproduced from *Laser Scanning for Archaeology: A Guide to Good Practice* (Payne 2011).

3.4.2 Retrospective conversion and 'heads-up' digitising

A term borrowed from the discipline of Library Science, 'retrospective conversion' in archaeology typically refers to the process of digitising existing hard-copy archaeological drawings, but could describe the digitisation of any primary visual data initially created in analogue format. It this instance, it refers to a digital version of an archaeological drawing not created with digitisation in mind, either because it was not a priority to those carrying out the research, or because the drawings were created before digitisation was an option. The term 'heads-up' digitising refers to the most typical process by which an archaeological drawing on paper is digitised, where the drawing is scanned and imported into a drawing program at the correct scale and orientation, and then traced using a mouse or a digitising tablet and stylus (Eiteljorg II *et al.* 2002). Whether a current project is planning to use 'heads-up' digitising to create a digital archive based on a selection of field drawings, or a site which has been dug previously is being re-examined and has drawings undergoing 'retrospective conversion' using 'heads-up' digitising, the result is the same. Both are generating secondary data in digital format from primary data, which means an additional translation process is taking place.

For field drawing, 'retrospective conversion' usually means the digitisation of plans and sections using a vector-based digital drawing program. These can include high precision Computer Aided Design (CAD) programs created for the architecture and engineering sector, like AutoDesk's AutoCAD, illustration programs like Adobe Illustrator and ACD System's Canvas, or a wide variety of specialist programs like Inkscape, which is an SVG editor. Like the forms of digital data capture used in survey and the handheld experiments mentioned previously, these drawing programs capture information using points, lines and polygons which are tied to *x* and *y* coordinates, with the addition of *z* coordinates for projects in three dimensions.

Digitising an archaeological drawing can be time consuming, but very useful for a number of reasons (Adkins and Adkins 1989, 233). The first is simply the ability to incorporate multiple scales, and see different levels of detail within a single drawing, (Reilly 1991, 134). Field drawings are created using a variety of scales, but vector drawing is based on mathematical calculations which can be re-calculated dynamically (unlike the static pixels used in raster images), so information created in different scales can be digitised in vector format on a 1:1 scale. The image is then viewed at whatever level of zoom is appropriate, and the vector image will simply re-calculate as necessary without a loss of information or resolution. Field drawings created at different scales can also be incorporated into the same drawing and viewed as a whole. Vector drawings can then become a single data source, from which individual views can be excerpted at an appropriate level of detail, to highlight particular information (Eiteljorg II and Limp 2008, 162). Field drawings are often created in unwieldy sizes and formats as well, in order to accommodate the necessary level of detail. Digitising such drawings preserves the detail in a way which makes them easier to work with, and allows the creation of simplified versions for analysis or publication in smaller or standardised formats (Hopkinson and Winters 2003).

Figure 31: Judith Winters, Editor of Internet Archaeology, with one of the very unwieldy large-format permatrace drawings from the 1975 Cricklade excavation. Photo by the author.

Another important reason to digitise a field drawing is so that information can be organised into layers. The archaeological record being set down in layers and the practice of archaeological excavation being concerned with removing those layers, it is not surprising this is particularly useful (Blomerus and Eiteljorg II 2009). In addition to CAD programs, most illustration programs allow for the use of layers, and in the case of programs like Adobe Illustrator, drawings created in CAD can be exported with their layers intact, manipulated, and then exported again into another format like SVG (Wright 2006). It is important to remember however, CAD programs were designed to facilitate the creation of things which need to be built, rather than document things which are being destroyed, so its really sheer luck they have functionality like layers, and other features which are useful for archaeology. Much searching on the Web only revealed one bespoke CAD program for archaeology called ArchaeoCAD. Developed in Germany by ArcTron, ArchaeoCAD, is designed to work with AutoDesk's AutoCAD, but with customisation for archaeological applications (ArcTron 2007), so it is still based on technology created for design rather than documentation.

Primarily because CAD programs are designed for use in industries like Architecture, Engineering and Manufacturing, where high levels of precision are vital, their use may create a false sense of accuracy when used with archaeological data (Eiteljorg II and Limp 2008). The person responsible for digitising a drawing, who may have had nothing to do with the field component of the project, has to make decisions about how to interpret the drawing. Attention must be paid to what, if anything, the digitisation process is imposing on the data. This does not mean the digitisation of field drawings creates erroneous information and should therefore be avoided. It simply means a level of translation is being added in order to create a more useful resource, and like any other step in the fieldwork process, the criteria and decisions used in that interpretation must be made transparent and explicit (Eiteljorg II *et al.* 2002).

Figure 32: Example of retrospective conversion. Upper image shows the original inked drawing, and the lower image shows a vector drawing which was created by first using 'heads-up' digitising in AutoCAD, and then importing the CAD file into Adobe Illustrator, with layers intact, for further refinement. Reproduced from *Problems with Permatrace: a note on digital image publication* by Guy Hopkinson and Judith Winters (2003).

CAD programs also allow the addition of information stored in tables to be linked to elements in the drawing from within the CAD program itself, or it can link to an external database. Digitisation is particularly useful for coping with drawings created using single context planning, especially for complex, deeply stratified sites. The ability to store the drawings of individual contexts separately and link them to relevant textual data for that context, and then group contexts together to form plans is very powerful (Lock 2003, 105). Contexts can be combined to create phase plans for analysis and publication (Reilly 1991, 134), and drawings of individual contexts which are part of a database allow composite drawings to be prepared quickly based on query results. Digital drawings can also be incorporated into a comprehensive data solution, which not only allows the data to be used flexibly and efficiently, but can also break down the traditional barriers between excavation and post-excavation (Lock 2003, 105-6).

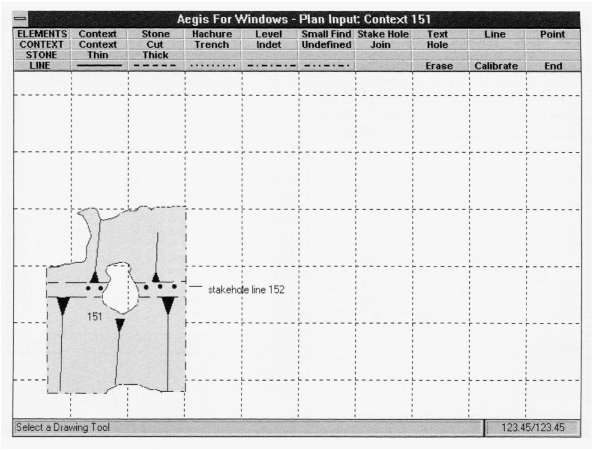


Figure 33: A digitised single context plan from an early version of the Integrated Archaeological Database, as developed by Mike Rains, when used as the recording system developed by the Scottish Urban Archaeology Trust (SUAT). Reproduced from *Using Computers in Archaeology* by Gary Lock (2003, 113).

One of the best examples of an 'integrated recording system' and one of the most long-lived is the Integrated Archaeological Database (IADB). This was recently evidenced at the 2010 British Archaeological Awards, where the IADB was Highly Commended in the Best Archaeological Innovation category. Originally developed in the late eighties for large projects being undertaken by the Scottish Urban Archaeological Trust (SUAT) by Pete Clark and Steve Stead, it was taken over by Mike Rains when he joined SUAT in 1989. When Rains moved to the York Archaeological Trust (YAT) in 1997, the IADB moved with him, and has continued to develop there ever since (Rains 2010). In use by a number of universities and commercial units, the IADB continues to innovate through its partnership with the Silchester Town Life project at the University of Reading.

The experimentation with handheld devices and tablet PCs for digital capture of primary archaeological data for use with the IADB is discussed in the previous section, but much of the innovative partnership between the IADB and the Silchester Town Life project has to do with breaking down the traditional distinction between excavation and post-excavation work. In particular, the creation of a Virtual Research Environment (VRE) for the VERA project was meant to allow excavation, post-excavation and publication to happen

simultaneously (Rains pers. comm. July 2006). This represents an example of a comprehensive data solution for archaeological fieldwork, moving toward a much more fluid paradigm, which would be impossible without digitised field drawings.

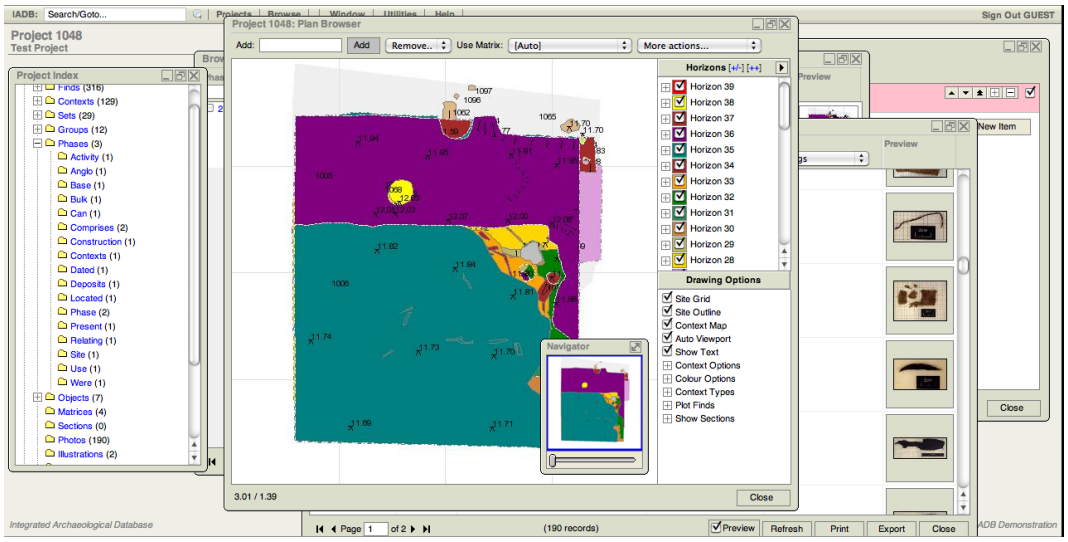


Figure 34: A screenshot of a dynamically rendered vector-based phase plan (and other windows) from the IADB. Taken from the IADB demo site. http://www.iadb.org.uk/demo/

Once a field drawing is in digital format, applications like the IADB show how powerful and useful they can be. At the same time, such a massive change in the way those drawings are created, viewed, manipulated and displayed has an impact on every part of what we understand about that data. Related to the issues of precision and its potential for influencing accuracy discussed in the last section, there are also issues surrounding the amount of certainty a digital image conveys. As machines, computers are perceived as being more accurate than humans, so digital drawings tend to carry more authority than images on paper. In order to create a digital image, authors will frequently need to combine data in which they have varying degrees of confidence in order to create something which communicates the information they are trying to convey (Miller and Richards 1995, 20), or so it can be used by other digital applications. Once data is in digital format, it becomes more fluid as it goes from one application to another, and the potential for creating misleading translations of that data increases (Gaffney and

Exon 1999). It will therefore be important to look carefully and document ways in which data derived from archaeological field drawings is transformed when using it with Semantic Web technologies, so users of that data understand how it has been processed, and can make informed decisions about the data when they use it.

3.5 Conclusion

The information communicated through field drawing is very significant to archaeologists, as shown by the results of *The Publication of Archaeological Projects: a user needs survey* (PUNS) report, published by the Council for British Archaeology. It places maps, plans and sections as third in importance, only behind the introduction and conclusion, in an archaeological report (Jones *et al.* 2001). Even photographic information is not rated as highly. This is even more significant as the results of the survey indicate very few people read a publication in its entirety.

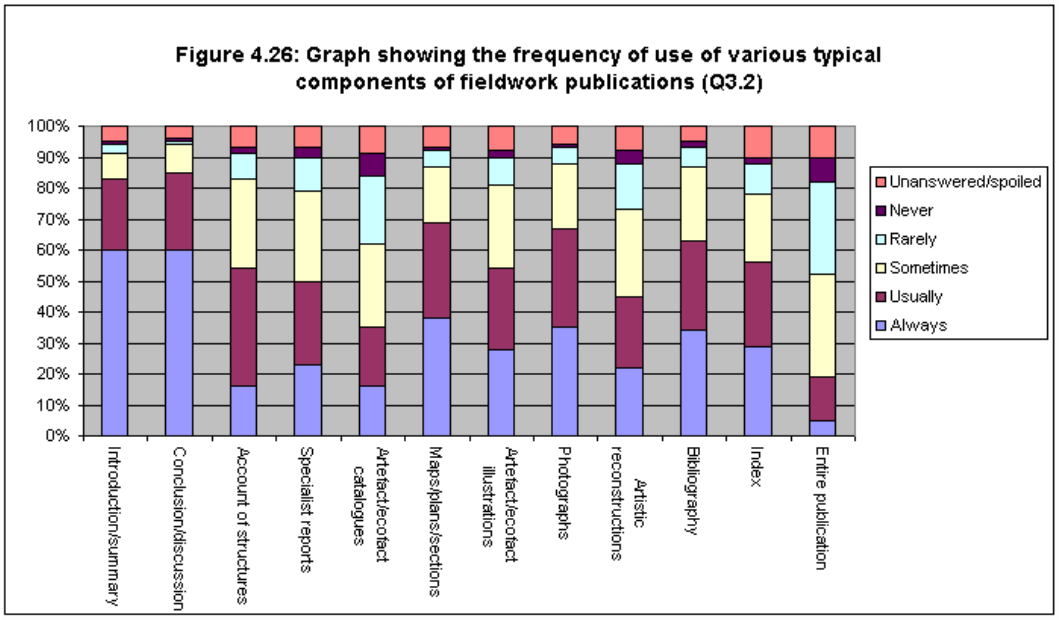


Figure 35: Graph showing the frequency of use of components of archaeological publication, reproduced from *From The Ground Up, The Publication of Archaeological Projects: a user needs survey.* Report and analysis undertaken by the Council for British Archaeology (Jones *et al.* 2001). http://www.britarch.ac.uk/pubs/puns/.

The PUNS survey was meant to evaluate the usefulness of archaeological project publications generally, and reflects the way project reporting and analysis has been traditionally presented. In contrast, the Historic Environment Information Resources Network (HEIRNET) User Survey was designed to assess the information needs of individuals and organisations, specifically using digital resources for archaeology and the historic environment (Brewer and Kilbride 2006). This survey produced some very interesting contrasts between what people in the Historical Environment sector find useful generally, and what is useful when it is presented in digital format.

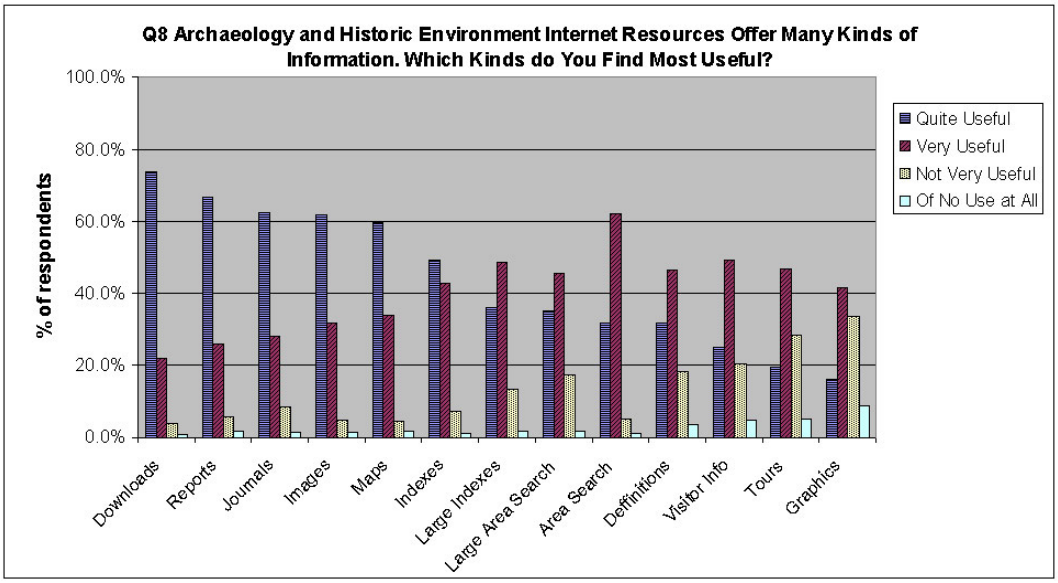


Figure 36: Graph showing the usefulness of Internet resources to practitioners working in archaeology and the Historic Environment, reproduced from *HEIRNET User Survey,* undertaken by the Archaeology Data Service. (Brewer and Kilbride 2006). http://www.britarch.ac.uk/HEIRNET/survey/section1.htm.

While the categories are not fully comparable, they are similar enough to show some significant differences that are of interest, especially with regards to archaeological field drawing. Maps rated extremely highly, as would be expected based on the PUNS report. In contrast, graphics, which is where elements like plans, sections and other types of vector-based spatial information would be included, received the lowest rating. In fact, of the 118 individuals who identified

themselves specifically as archaeologists, only five indicated online graphics were 'very useful.' Based on this information, one conclusion might be that a significant gap has developed between the type of resources archaeologists rely upon for their research, and the ability of digital technology to deliver those resources in a useful way. If this is the case, work needs to be done to address how the data from field drawings is presented on online.

Given the importance of vector-based field data to archaeology, and the challenges inherent in the technology with presenting it on the Web with full functionality, it is not surprising archaeologists feel they are not well served by the Web in this area. Perhaps this is set to change in the near future though. SVG, the W3C standard for vector data, is finally receiving native support across the majority of Web browsers, and proposals are currently underway to extend the current standard for use with online mapping and Web GIS (Li and Dailey 2011). SVG has been erroneously lumped in as an innovation forming part of the new HTML5 standard in some of the popular media, but its recent momentum is purely coincident. Whatever technology is used to make vector-based data available on the Web, the most important factor will be whether it is presented in a way that archaeologists will find useful, as they do within a traditional fieldwork publication.

The creation of a visual field record as used in archaeology has a history that both predates, and lags behind the creation of the discipline itself. Visual interpretations of information derived from what we now term the 'archaeological record' have been recorded as long as there were people to observe and draw it. Even as late as 1877 however, the publication of *British Barrows* by Canon Greenwell did not contain a single plan drawing of any of the over 300 barrows he excavated (Piggott 1965, 172). The transformation inherent in the destruction of the archaeological resource, in parallel with the creation of a comprehensive field record, is a relatively new concept.

What has traditionally been described as archaeological field drawing; the creation of plan and section drawings as a primary visual data record, has changed in form within the last several decades with the introduction of digital technologies, but its function remains the same: to translate the inherently visual and spatial information which makes up the archaeological record from Tufte's 'three-space world' into 'flatland' in a way which keeps sufficient meaning intact to allow understanding. Preservation of the relationships between pieces of information (and therefore meaning), being foundational to the Semantic Web, should theoretically allow the translation and transformation of data derived from archaeological field drawing with much of the original meaning intact. At the same time, the level of abstraction necessary to make data machine readable, and therefore usable within the Web of Data has implications for information meant to be visually understood.

How this is practically achieved has to do with the way field drawing is carried out, and what conventions and recording systems are used. The comprehensive nature of single context recording, and the way it creates an archaeological record which can be pulled apart and put together without losing its stratigraphic relationships, makes it well suited for Semantic Web applications. Individual contexts are both self-contained units of information, which derive their meaning from their stratigraphic connections with other contexts, much like the nodes and edges making up the graph data model used in RDF, and therefore the Semantic Web. How this might be practically demonstrated, is the subject of the next chapter.

Chapter Four

A Practical Application of Archaeological Field Drawing Data using Semantic Web Principles

What if you need to buy a new car? The obvious choice is to buy another car just like the one you have in the garage. The type of car that has been around for 30 years, works great on the road, perfect mileage/fuel consumption and every auto shop knows how to repair it. But there is also a new car on the market that is just as stable and fast, similar cost, but much more flexible because it also can fly and run underwater. Which car would you buy?

–Jans Aasman (2011)

Raw data now!

–Tim Berners-Lee (2009b)

4.1 Introduction

As with most new technologies, those who first take up the challenge of implementation are usually the practitioners who are most skilled; their enthusiasm driven by an understanding of the potential benefit on offer. Because the development of the Semantic Web is ongoing, any practical work undertaken now must still revert to the theoretical at times, but it is worth attempting to explore what is currently possible, to better understand what may be possible in the near future. Archaeologists are already partnering with computer scientists to explore how Semantic Web concepts and technologies might be useful, but at what point will non-specialist archaeologists be able to take advantage of the Semantic Web? What are the obstacles to be overcome, and what generic tools are available (or should be developed) which can be adapted for use with archaeological data? How best can the potential usefulness be demonstrated?

These are all important questions to consider if archaeology is going to move from experimental to practical use of the Semantic Web.

At the same time, archaeology is inherently spatial. Understanding the relationships created by the physical proximity of archaeological resources relative to one another is the foundation of archaeological research. These relationships are expressed in two and three dimensions, but throughout the history of the discipline, the most important visual key to understanding an archaeological resource has been the plan drawing (augmented by section drawings for understanding stratigraphic relationships). Whether a plan drawing begins its life on paper, *permatrace* or is 'born digital', today most are digitised using some form of vector-based drawing method, and are comprised of points, polylines and polygons, often with associated annotation and spatial geo-referencing. In the case of points organised into polylines and polygons, these become single objects made up of multiple pieces of information grouped together. The basic building block of the Semantic Web, the RDF triple, is designed to describe single subjects, predicates and objects, not subjects, predicates and groups of geo-referenced data points with relationships which must be preserved in order to be understood. If archaeologists want to include Semantic Web technologies and concepts in their toolbox, what is necessary to make sure data from visual and spatial sources is represented?

This chapter will explore these questions, by applying Semantic Web concepts to archaeological field data, using tools and technologies which are freely available, open source, and accessible to non-expert users whenever possible. It will also explore how spatial information might be included with textual data in future, thereby representing a more complete picture of the information typically generated by archaeological fieldwork (Karmacharya *et al.* 2009). To do this, spatial data from two different archaeological sites, derived from field drawings and their associated data, created by two distinct organisations with differing field methods, data collection techniques, and data manipulation practices will be used.

Both sites are located in Yorkshire, in the north-eastern part of England, in the United Kingdom.

The first dataset is from a rural Anglian and Anglo-Scandinavian site in the Yorkshire Wolds, near the village of Cottam, excavated by the University of York in 1994. This dataset was chosen because it represents a best practice exemplar of digital archiving, from a typical excavation as carried out in the UK. The archive is permanently held by the Archaeology Data Service, and is freely available for download (Richards 2001b). The second dataset is from a portion of the urban, multi-period, Hungate site; one of the largest excavations to ever be carried out in the York city centre. The York Archaeological Trust have been conducting fieldwork on the site since 2007, and the excavation is due to conclude in 2011. The dataset is from a portion of Area H, which is associated with the Anglo-Scandinavian occupation of the site. While the data generated by the Cottam excavation was initially created using common software like AutoDesk's AutoCAD and Microsoft Access and then converted to archival formats, the York Archaeological Trust has been using the bespoke Integrated Archaeological Database (IADB) for processing its archaeological data since 1997. The IADB is a complete data management system which handles the information from excavation recording and analysis through to eventual preparation for publication and archiving (Rains 2010). The IADB is open-source, and capable of exporting data in archival formats, and so conforms to best practice principles for bespoke data management for archaeology. The Hungate dataset was chosen because it is related to the Cottam dataset archaeologically, but differs technologically.

The Cottam and Hungate datasets fall into the same early medieval, eighth-tenth century AD, Anglian and Anglo-Scandinavian time period. They reflect the changing relationships of people living in what is now Yorkshire between the existing Anglo-Saxon Northumbrian kings, Anglo-Scandinavian York after its conquest by the Danes in 866, and the subsequent conquest and partitioning of the rest of the land of Northumbria in 876 into the Danelaw territories (Richards

2000, 27). While the Anglo-Scandinavian history of York has been the subject of considerable study (Hall *et al.* 2004; Smyth 1975-9), how the surrounding rural landscape interacted with the capital is less clearly understood, as these areas have received less attention from archaeologists. Addressing this deficit was the essential research aim behind the decision to excavate at Cottam, as part of what was initially titled the York Environs Project (Richards 1993).

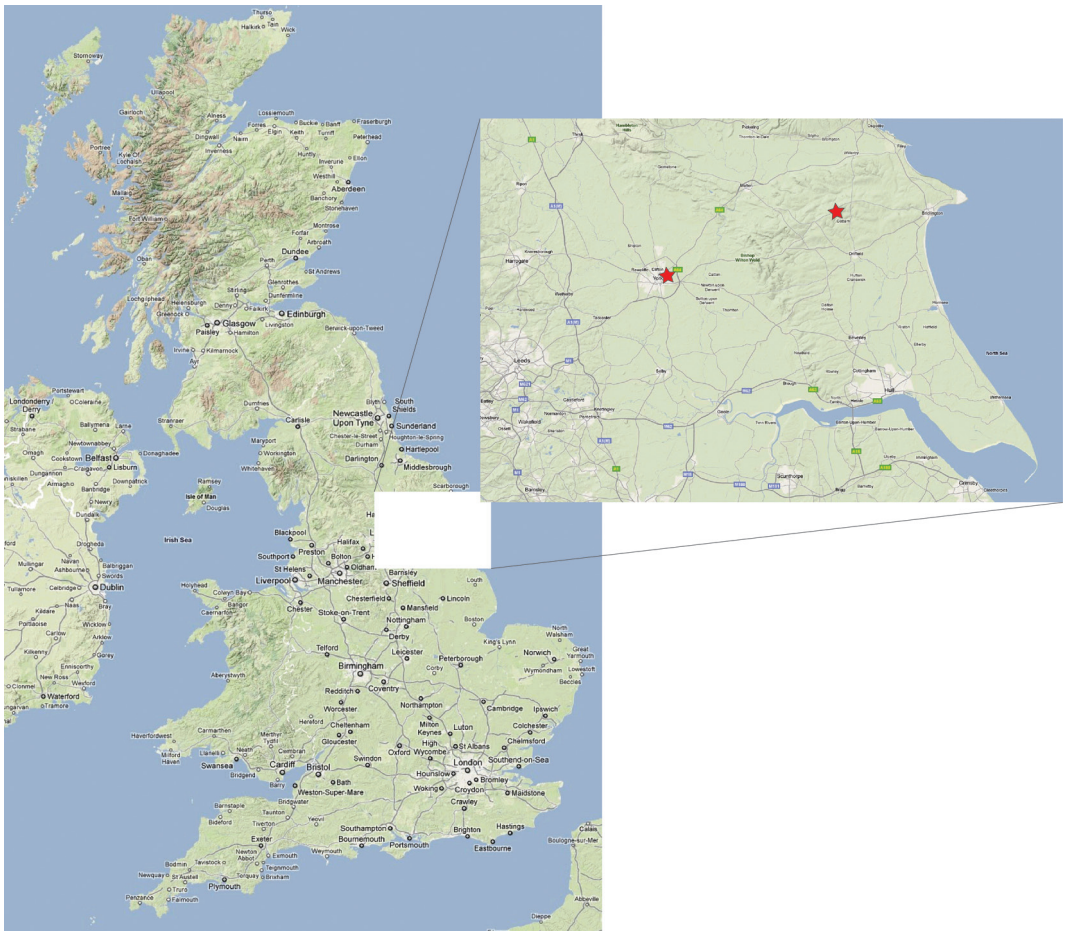


Figure 37: Locations of the Cottam and Hungate excavations, in the Yorkshire Wolds and the Vale of York, respectively. Base map compiled using AllAllSoft.com Google Maps Terrain Downloader. http://www.allallsoft.com/gmtd/.

This chapter will explore these two datasets; documenting the attempt to navigate them through Semantic Web processes, and into a theoretical discussion in areas where demonstration is not yet possible. It will work through several layers of the Semantic Web 'layer cake' as far as the current state of technology allows,

including implementations using best practices for publishing Linked Data. It will incorporate semantics using the domain ontology for the Cultural Heritage sector, the CIDOC-CRM, and the archaeology-specific extension to the CIDOC-CRM created by English Heritage, known as the CRM-EH. It will process data using newly available tools for aligning the data to the CRM-EH, which can then be queried using the Semantic Web Query Language (SPARQL), and discuss the particular issues relating to Semantic Web technologies and spatial data. It will attempt to construct new relationships from archaeological information generated by spatial data, make heterogeneous data interoperable, and explore ways to visualise the result.

Combining these two structurally heterogeneous, yet archaeologically related datasets using Semantic Web principles may facilitate new questions, and perhaps new comparisons between urban and rural settlement in York and the surrounding Wolds which might not have been possible (or considerably more difficult) by other means. It may allow the resulting data to be published in new ways, which can then be re-used by other archaeologists working in this area of research, who could then combine it with their own data and draw new conclusions as well. It is hoped that this chapter will demonstrate that use of the Semantic Web for archaeological field data has become well within reach for archaeologists, especially those using single context recording on data from sites within the UK, and that the incorporation of spatial data derived from fundamental visual sources like field drawings will be possible as well, if not now, then in the near future.

4.2 The sites

The City of York is located in the Vale of York, bordered by the Pennines, the North York Moors and the Yorkshire Wolds. It sits in the centre of the ridings of Yorkshire, in the north-eastern part of England. York lies roughly halfway between the capitals of London and Edinburgh, in a largely rural landscape. It has been a Roman, Northumbrian, and Viking capital, and it's many periods of historical and

political importance and decline means it has retained a different character to other, larger, more industrial cities in the region. York has generated a unique ebb and flow of influence on the surrounding region, and on the country as a whole. Belying its reach, York's changing fortunes are based on a very central focus; this is seen in the dense complexity of the city's archaeology, described by Patrick Nuttgens thus:

> Almost uncannily the city of York reveals its continuity. But it is not just a superficial or 'end-on' continuity, with periods following one another as if in a straight line; it is basic, central and fundamental. The same core dies and is reborn again and again (Nuttgens 2001, 6).

While York was a site of prehistoric activity, no evidence for a substantial pre-Roman settlement has ever been found. In the words of Patrick Ottaway 'York owes its existence to the Romans'. Looking for an auspicious site to build a fortress to house the Ninth Legion, the commander Petilius Cerialis chose the site which became known as Eboracum in AD71 (Ottaway 2007, 1). The area afforded advantages important to the Romans (and others) when choosing a place for settlement. York sits on a glacial moraine in the Vale of York, which is cut by the confluence of two rivers, creating a natural crossing of the River Ouse (still tidal at that time), which flows, nearly to the east coast before joining the River Humber, and then to the sea. The raised moraine in the relatively flat vale, with established trackways predating the Roman occupation by at least 2000 years, meant it was defensible, and allowed the establishment of good transportation and communication necessary for taking control of the area (Nuttgens 2001, 12-5; Ottaway 2007, 1-2). The Roman conquest of the north of England and the founding of Eboracum, their northernmost fortress, set in motion a unique set of circumstances which would make York the focus of a variety of important events, developments and historic relationships still visible in the city today. These circumstances influenced the lands surrounding York as well, including the distinct region known as the Yorkshire Wolds.

The Yorkshire Wolds have a long and rich archaeological history, with a greater density of prehistoric sites than any other region in the north of England (Fenton-Thomas 2005, 29). The term wold refers to the particular landscapes in England, which are part of a chalk layer formed during the Cretaceous period. There are other landscapes with similar morphology where the term wold is used (as in the Cotswolds in the southwest of England), but typically the term refers to the single chalkland region on the north-eastern coast, ranging from Spilsby in the south to Filey Bay in the north. The river Humber divides the chalkland into two regions. South of the Humber are the Lincolnshire Wolds, while the Yorkshire Wolds comprise the area to the north. The Yorkshire Wolds cover approximately 1,350 sq. km. (Stoertz 1997, 1).

During the early third century, villas become the preferred form of settlement, and the creation of rural estates over small indigenous settlements demonstrates the beginnings of a more hierarchical social system. There was an increase in the number of people living in the Wolds generally during the Roman period, and changes in how they were using the landscape. In addition to pottery production, they began more intensive agricultural practices with large areas under the control of the rural estates (Fenton-Thomas 2005, 75). During the post-Roman period, settlements like Cottam seem to have been focussed on nearby Driffield rather than York, and may have been under the control of one of the Northumbrian kings based at Driffield, but the situation is not fully understood (Richards 2001a).

With the conquest of York by the Danes in 866, focus seems to have returned to York, or Jorvik as it was called during the Anglo-Scandinavian period. The establishment of Jorvik as a Viking capital brought about a period of new activity in the city. Areas now considered the heart of the city are settled for the first time during this period (or resettled after a period of abandonment after the end of the Roman period), as evidenced by the excavations at Coppergate (Hall 2007, 53) and many other, smaller excavations throughout the city (Hall 2004), show the Anglo-Scandinavian period also ushered in a new period of mass-production

of goods for use within the city and beyond, with the quality of the craftwork showing the expertise of specialists was employed (Hall 2007, 55). This would have been impossible without the considerable new interaction with the area around the city:

> Most raw materials needed by York's manufacture/fabricators were available from the immediately surrounding area, as were the foodstuffs needed to feed a population which, most unusually for the time, was concentrating its productive efforts somewhere other than in farming. Iron ores, lead wool, flax, wood and timber, antler, animal products of all sorts including meat, milk, hides and bones, grain for bread and brewing, fish and shellfish - all could be brought into York from the estates, land-holdings and farms of the Yorkshire countryside. And back to this countryside might go the products of York's craftsmen... (Hall 2007, 56).

The Anglo-Scandinavian site at Burrow House Farm near Cottam, is one of the few places where excavated archaeological evidence of this interaction can be found outside of York. Subsequent work in this area has been undertaken using the larger corpus of information gathered using metal detection as part of the Viking and Anglo-Saxon Landscape and Economy (VASLE) project, but there is still considerable work to be done in order to better understand the complex connections going on during this time (Richards *et al.* 2009). Exploring new and creative ways to use the data generated by the excavations from the Anglo-Scandinavian period at Cottam and Hungate both may help contribute to this work.

4.2.1 The excavations at Cottam

The hamlet of Cottam lies near the village of Sledmere in the northern region of the Yorkshire Wolds, to the East of the Vale of York. Today, the Wolds are part of

a largely rural landscape, as they were during the period of Anglian and Anglo-Scandinavian settlement. To better understand the rural landscape surrounding York during this period, and its relationship to the city, excavations were carried out at an eighth-ninth century AD site at Burrow House Farm near Cottam. Cottam was identified as a 'productive site' or a site where metal detectorists were making significant finds of coins and other metal objects, and as these types of sites had not received significant attention from archaeologists in the past, it was chosen for excavation by the University of York. The excavations at Cottam demonstrated a possible change of allegiance (at least cultural, if not actual) from the Northumbrian royal family to the Viking kingdom based in York towards the end of the ninth century AD. The settlement prospered during this period, and also showed signs of more egalitarian trade practise, but was abandoned shortly thereafter in favour of the construction of a high-status farmstead at the nearby site of Cowlam (Richards 2000, 53; Richards 2001a; Richards *et al.* in prep).

Figure 38: Location of the excavation trenches near Burrow House Farm, near the hamlet of Cottam. Detail reproduced from *Anglian and Anglo-Scandinavian Cottam: linking digital publication and archive* (Richards 2001a).

Various types of fieldwork were carried out at the site from 1993 to 1996 by the University of York, building upon metal detecting data gathered from 1987 to 1996, and data from fieldwalking undertaken by the East Yorkshire

Archaeological Society in 1989. The excavations at Cottam were divided into two main areas of settlement, designated as Cottam A and Cottam B. Cottam A was located to the southeast of Burrow House farm, and was excavated in 1996 at the end of the project. While it yielded some Anglo-Scandinavian finds, it was determined to be primarily a Romano-British farmstead largely abandoned in later periods. Cottam B is located to the west of Burrow House Farm, and various forms of fieldwork were undertaken, including fieldwalking, aerial photography, metal-detecting and geophysics. Excavations were carried out in 1993 and 1995 (Richards 2001a).

Several archaeological phases were designated for Cottam B:

Period IIA: Anglian Phase A
Period IIB: Anglian Phase B
Period III: Anglo-Scandinavian
Period IV: Medieval and later

Three excavation trenches were opened in 1993, but only two were excavated. These are referred to as COT93.1 and COT93.3. Both trenches had features belonging to the earlier Period IIA, including traces of two timber structures, and several post-holes. COT93.1 also contained a shallow ditch with internal post-holes. During the later Period IIB, the previous buildings were demolished and a more substantial boundary ditch built. COT93.1 contained a series of pits, one of which contained a human skull, and COT93.3 contained a trench, a possible timber structure and several other features, including a corn drier. No period III occupation was found in the excavations carried out in 1993, and the site appears to have been abandoned after the Anglian period (Richards 2001a).

A further excavation trench was opened in 1995 and is referred to as COT95. The first evidence of occupation within the COT95 trench falls within period IIB, and included several shallow, truncated ditches running roughly east-west across the northern part of the site. One of the ditches included a series of stakeholes in an alignment suggesting a possible fence-line. The ditches contained a few finds,

including pottery sherds and a ceramic lamp base. COT95 is the only trench from the series of excavations at Cottam to have yielded structures dating to Period III; the Anglo-Scandinavian period. Several enclosures dating to this period were found using geophysical survey, and two fell partially within the excavated area. This included a large entryway in the south-western area of the excavation. Two parallel sections were placed on the east side, which revealed two large post-holes. The west side of the ditch was also sectioned in several places, revealing a series of depositions. Corresponding post-holes on the east side of the entrance were also found (Richards 1999, 41-3).

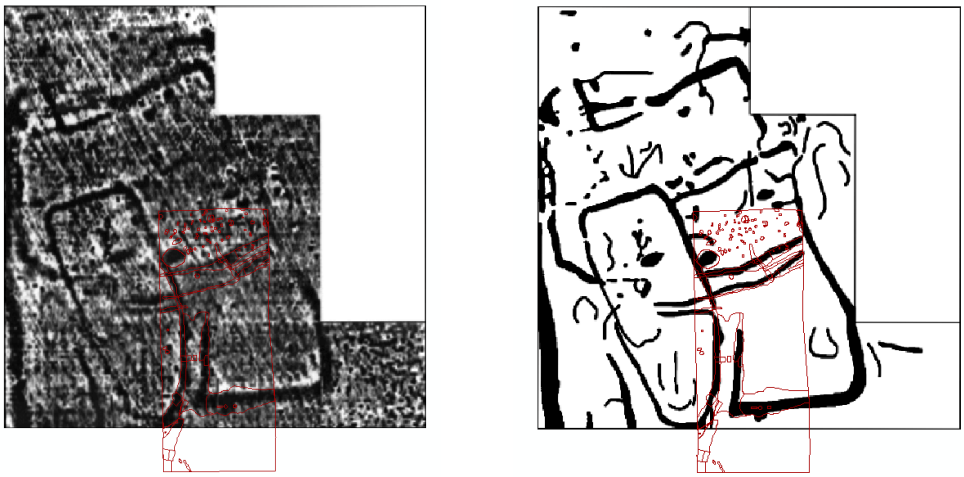


Figure 39: The COT95 excavation trench and contexts with relation to the geophysics. Magnatometry images reproduced from *Anglian and Anglo-Scandinavian Cottam: linking digital publication and archive* (Richards 2001a).

Figure 40: Left: The COT95 excavation trench from the north. Right: The COT95 excavation trench from the south. Images reproduced from *Anglian and Anglo-Scandinavian Cottam: linking digital publication and archive* (Richards 2001a).

The trench contained other gullies, slots and lesser trenches and a number of post-holes were found in the northern area. It was possible to see some alignments, but no obvious structures, as the building may have spanned several phases of occupation. The post-holes were classified as Class A, for those deemed probable post-holes, and Class B for those that could be post-holes, but it could not be known for certain. The largest post-holes were found in the northwest corner of the trench, along with several stakeholes. This area also contained a very large pit cut into the bedrock. While the purpose of the pit is unknown, it was deemed a likely chalk quarry hole (Richards 1999, 46-7). The finds from the Anglo-Scandinavian period of COT95 show the changes from the previous Anglian period. A far greater variety of ceramics were found, including York ware. A pewter disc brooch was found, which was of a similar style to those found at the Coppergate excavation. Other items associated with Scandinavian occupation were found, including Norwegian honestones and lead weights associated with a corresponding period of non-monetisation. The Anglo-Scandinavian settlement at Cottam is estimated to have lasted only around 50 years however, and the trend towards village nucleation may have influenced a move to the west to the larger village at Cowlam (Richards 1999, 97-8). To further the research aims which began at Cottam, excavations were carried out at Cowlam by the University of York during 2002-3, but only revealed evidence of Anglian occupation (Richards *et al.* in prep; Richards *et al.* 2009).

4.2.2 The excavation at Hungate

The Hungate excavation is the largest archaeological dig in York's city centre in the last 25 years. After seven years of preparation, excavation began in 2007 and is scheduled to finish by the end of 2011, with a further two years scheduled for post-excavation. The excavation is a mitigation scheme in advance of a major urban regeneration project (Connelly 2007, 1), which spans several streets and comprises a single area more than twice the footprint of York Minster. The York Archaeological Trust is not only responsible for carrying out the excavations in conjunction with the construction schedule, but has made it central to their many

public outreach efforts. This includes public tours and open days, community archaeology projects, participation of school groups and volunteers, as well as the annual *Archaeology Live* training school (Connelly 2007, 3).

Figure 41: Block H of the Hungate excavation facing southeast, with the Stonebow heading to the left in the foreground and the Hungate street frontage and Block G on the right. Reproduced from *Great Expectations for the Hungate Excavation* (Connelly 2007, 1).

The site is so extensive the area has been divided into eight blocks, corresponding to the multi-storey (primarily) residential buildings scheduled to be constructed. The area where the new construction will be most invasive is Area H, which is the only area undergoing complete excavation. Because of this, the work in Area H will span the entire five-year excavation, and will be the last area where building will take place. Area H has been divided into two sub-areas, H1 and H2. Work began in Area H1 first, with the goal of locating the medieval church St John's in-the-Marsh, in order to ensure the burial ground remained undisturbed (Kendall 2007, 6), and was found relatively quickly (Kendall 2009, 1). After 15 months, Area H1 was fully excavated and showed occupation from the Norman Conquest to the modern period (Connelly 2008, 1). Excavation also revealed occupation from the Roman period, including several burials, but no evidence was found from after the Roman period to the mid-10th century (Kendall 2009, 2).

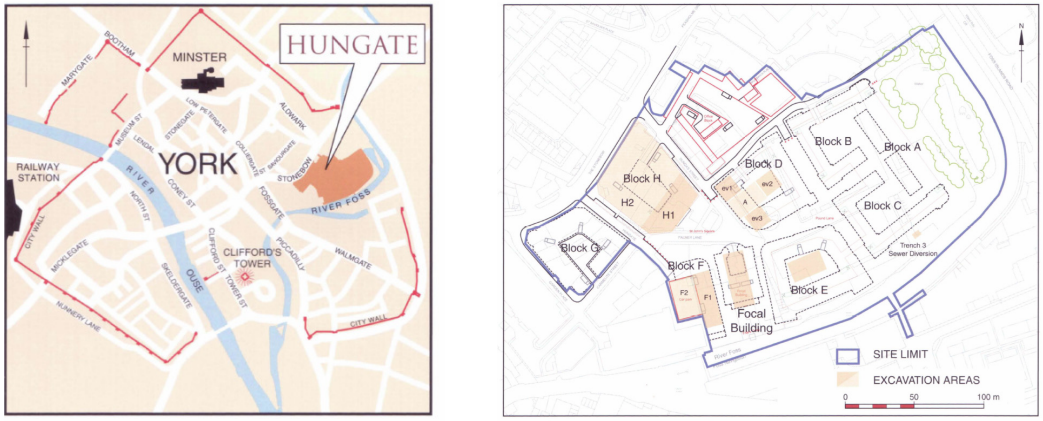


Figure 42: Left: Location of the Hungate excavation within the York city centre. Right: Location of Block H, and Areas H1 and H2. Reproduced from *Hungate Excavations: Season 2 Draws To An End* (Connelly 2008, 1).

In 2008, which was the final year of work in Area H1, a deep trial trench was dug in Area H2, which revealed a sunken floored building from the Anglo-Scandinavian period. As activity from this time period was not found in Area H1, nor from any of the other areas receiving less complete excavation, this was the first evidence of occupation during the time when York was a Viking capital (Hunter-Mann 2009, 4). At its widest, the trial trench spanned 9.0 metres and at its deepest, reached depths of 3.4 metres. Roman pottery was found, but there was no evidence of Anglian occupation. The earliest structure is a large, sunken floored building, measuring 4.3 metres long by 3.5 metres wide, with a depth of .8 metres. Unusually, one wall consisted of planks made from boat timbers. While the structure was similar to those found at the Coppergate excavation, the presence of a central hearth indicated it was likely a single story structure, more consistent with types found in Anglo-Scandinavian sites in London (Hunter-Mann 2009, 6). The trial trench also revealed a stony surface which aligns with Haver Lane, a residential street lost during the demolition in the 1930s, which suggests a possible Anglo-Scandinavian origin for the street (Hunter-Mann 2009, 7).

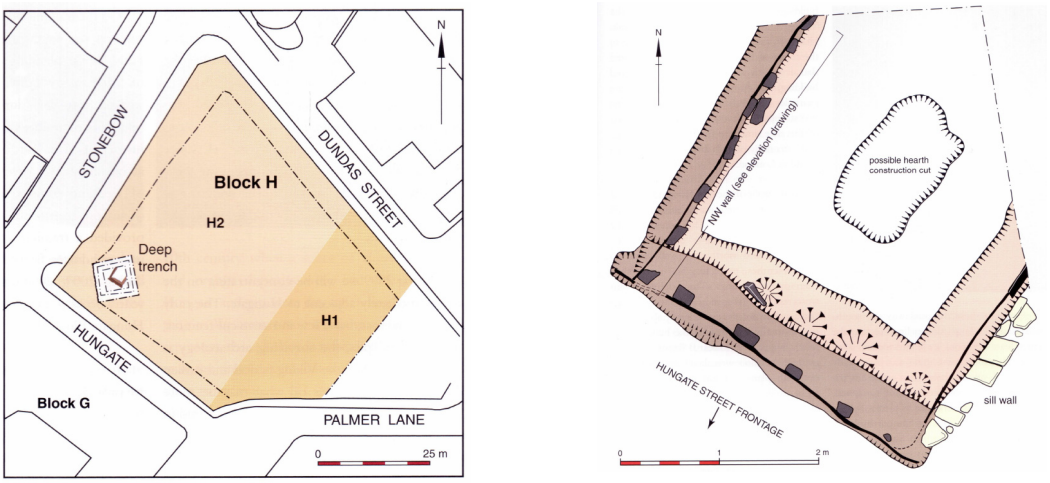


Figure 43: Left: Location of the deep trench in Area H2. Right: Plan drawing of the sunken floored building found in the deep trench in Area H2. Reproduced from *The Vikings Come to Hungate…* (Hunter-Mann 2009, 4-5).

Due to the good preservation of the wooden planks, further analysis of the boat timbers used to construct the sunken floored building was possible, including dendrochronological analysis. It was determined the trees from which the boat was built were cut down no earlier than 953, and broken up for use in the building within only 12 years. It was also determined that the wood was likely local; constructed using techniques originating in the southern coast of the Baltic Sea, and brought to the British Isles during the 5th-6th centuries. So while the structure was clearly Anglo-Scandinavian, the boat was not (Allen 2009, 9-11).

Figure 44: The deep trench facing southwest, showing the Anglo-Scandinavian features in the lower half. Reproduced from *The Vikings Come to Hungate…* (Hunter-Mann 2009, 4).

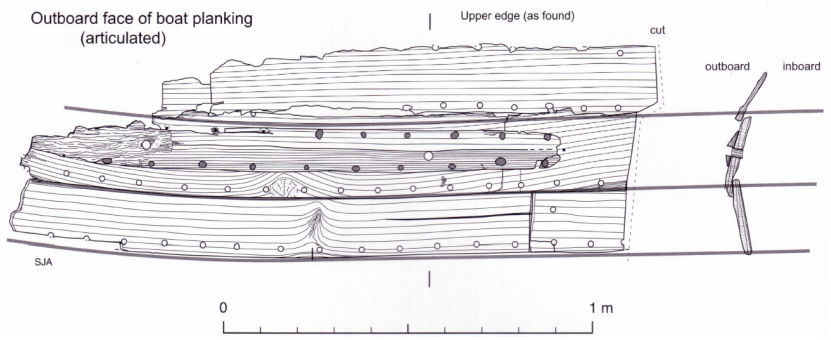


Figure 45: Plan and section of the boat timbers making up part of the sunken featured building found in the deep trench of Area H2. Reproduced from *Rocking the Boat* (Allen 2009, 10).

With the completion of the excavation in Area H1, full excavation of Area H2 commenced. The chronology of activity seemed to show a rising of the ground surface (by over a metre in some places) during the early to mid-10th century, which was probably a reflection of the site's proximity to flooding by the River Foss. These deposits had a distinct 'tiger striped' stratigraphy, which was not clearly understood. During the mid-10th century the area was partitioned with lanes made of river cobbles, which sit perpendicularly to the modern street of Hungate,

possibly showing the origin of its alignment. Sunken featured buildings appear in the late 10th century, but have been designated basements of unknown use. To the northeast, a line of cess, rubbish and wicker lined pits parallel to the modern street were also in use during this time. Later, during the 11th century, ditches were cut on similar alignments to the river cobble partitions (Connelly 2010, 1-3).

As work carried on in this area (H2), more Roman period burials were found in the south-eastern part of the area, and in the Anglo-Scandinavian section in the southwest, more rubbish and cess pits were found, and evidence for what was causing the 'tiger stripe' appearance of the raised deposits. Hearth structures were discovered which were used and then 'raked out' which contributed the black layers of ash, and destroyed oven-like structures were found, which were the source of the layers of burnt daub. In between the layers of burnt material, were layers of silt and clay, which produced the unusual striped deposits (Kendall 2010, 9-10).

Early results from the current and final year of excavation have revealed a total of six sunken featured buildings dating to the late 10th century. The most recent building to be found is of a new type. It was not cut as deeply as the others, and had a substantial stone-filled foundation and entrance. The building measured 7.4 metres by 4.1 metres, with a 4 metre long, 1.4 metre wide entrance passage. This was a large, substantially built structure, and unlike the other buildings, which sit between the cobbled/ditched partitions, sits across two plots, even though it is likely from a similar time period (Connelly 2011, 4-6). Excavation is due to be completed by the end of 2011 at Hungate, which will be followed by two years of post-excavation, and should result in further insight about the Anglo-Scandinavian activity in this part of York.

4.3 The data

The archaeological fieldwork carried out in the COT95 trench from the *Burrow House Farm, Cottam* excavations, and the H2 section of Area H from the Hungate

excavation, both produced a digital visual record of the spatial relationships of their archaeology (Richards 2001c). Both excavations were overseen by investigators trained in the most widely accepted excavation practice in the UK, so the records were created using single context recording techniques. This means each time a new context is encountered, it receives a unique number and an attempt is made to define and record its extent. As excavation proceeds, contexts may join and be found to be related, or during post-excavation analysis, grouped together to form phases which allow supposition about the nature of the activity found there, but the 'context' remains the individual unit used for defining the formation of the site.

Contexts are a convenient vehicle for defining spatial information digitally. They often form natural polygons, but in archaeology, defining a context as a polygon may not be straightforward. The edge of a context may be difficult to find, and during recording it is important to denote (with whatever drawing conventions are in use) 'edge uncertain' when necessary. Otherwise, the person recording the context may be making assumptions which do not truly reflect the nature of the archaeology in the ground. Contexts also frequently extend into unexcavated areas, so their full extent cannot be known. In such cases it is important to note the edge of the context as recorded, is not necessarily the edge of the context in the ground.

These issues can be problematic when attempting to define spatial data from digital archaeological archives in vector format. In order to create spatial entities ready for conversion into RDF, they must be rendered as closed polygons. Contexts extending into unexcavated areas must be truncated, and contexts where the edge is uncertain must be made artificially certain. While these issues are unavoidable, it is important to acknowledge the impact these changes may have on any archaeological interpretations made about the resulting data, and state what decisions have been made. It should be acceptable to make these compromises in order to understand the data in new ways, as long as assumptions are made explicit, so those using the data in future can decide with what level of certainty to interpret it.

Once archaeological contexts have been defined digitally, in vector format with closed polygons, software functionality can be used to augment the information about those contexts. Vector data can be brought into CAD or GIS programs (if not there already) to be georeferenced (specifically in GIS if the site is large enough to require projection), mitigating information about how decisions were made to close polygons can be added to attribute tables connected to the relevant contexts, and information calculated automatically about the contexts. Once the data has been processed and augmented as required in the CAD or GIS software, it is ready to be exported for eventual conversion into RDF.

In order to know how the data should be exported, it is necessary to understand how it will be used. For Semantic Web applications, the goal is to bring the data into a *knowledgebase*. A *knowledgebase* is the software component where the data within any data store actually lives. In order to facilitate easier use of and interaction with a *knowledgebase*, sets of tools have been created, which are referred to as *frameworks.* Most *frameworks* provide the storage structure for handling the RDF data, known as the RDF store, triplestore or graph store, along with a way of interacting with the data in the form of an access point or query processor, and a reasoning engine to allow inference (Hebeler *et al.* 2008, 142).

There are a variety of *frameworks* available for working with Semantic Web data in several common programming languages. The most frequently used by developers are Jena and Sesame. Jena is an open source project originating from the HP Labs Semantic Web Program (Jena 2011), and Sesame is freely available from OpenRDF.org; an open source project of Aduna Software (OpenRDF 2011). Both are written in Java, and their use requires a good understanding of Java programming. Both have an active user community and extensive documentation, but for an archaeologist without specialist knowledge of the Semantic Web and Java, their use would represent a very significant learning investment.

For most applications, it is not necessary to create a bespoke *knowledgebase* using a *framework*, and generic RDF store software will do. Most include similar functionality, in the form of access points and reasoning engines, and well-supported free versions are often available from software developers who are marketing more robust commercial variants. Use of these generic RDF stores usually requires a good understanding of UNIX, but represents a much lower learning curve than a *framework*. Before it can be brought into an RDF store, data needs to be aligned to an appropriate domain ontology and converted into RDF format. This can be another significant hurdle for anyone wishing to incorporate his or her own data into an RDF store (rather than using already existing Linked Data). As will be discussed in section 4.4, a tool is newly available for use by archaeologists using the single context recording technique, which provides this crucial step, and the data need only be exported as a simple table in standard CSV or SQL format.

The only process that cannot be handled using existing tools, and is still missing from this workflow is the step for taking data from the vector drawing program and converting it into the correct format in CSV. For this research, it will be handled by a custom loading program written in Java, but it could be done in other ways and/or using another programming language. Because the Cottam and Hungate datasets were created using differing techniques and technologies, and for somewhat different purposes, they followed separate paths to come to the point where their data was converted to CSV in preparation for inclusion in a common RDF store. The following recounts the journey of each dataset.

4.3.1 The data from Cottam

The data from Cottam was downloaded from the project archive for *Burrow House Farm, Cottam: an Anglian and Anglo-Scandinavian Settlement in East Yorkshire*, held by the Archaeology Data Service (doi:10.5284/1000339) (Richards 2001c). The Archaeology Data Service (ADS) is a UK national archive for primary archaeological data, which promotes standards, and creates best practice guidelines. Established in 1996, the ADS is based at the University of York, and

is made up of a consortium of Higher Education and related national institutions, with an advisory committee of individuals representing interests across the discipline (Archaeology Data Service 2011b). The ADS provides a wide range of services, but primarily maintains an archive of persistent, freely available archaeological data, which plays an important role in responsibly mitigating the destructive process inherent in much archaeological fieldwork.

The ADS archive contains all the primary data from the fieldwork undertaken at Cottam B. The HTML archive includes the overall research design, the Level III reports for the excavations carried out in two areas in 1993 (COT93, Area 1; COT93, Area 3), and in 1995 (COT95). It includes reports for the fieldwalking, geophysics and metal detection carried out over the site, and the relevant reports for the animal and plant evidence found during the excavation. Finds reports are also included for bone and antler, flint, iron, copper alloy and non-metallic objects, non-ferrous metal, post-Roman coins, pottery and stone. In addition to the reports available in the archive, the *Burrow House Farm, Cottam* archive is linked to the publication *Anglian and Anglo-Scandinavian Cottam: linking digital publication and archive* (Richards 2001a) published in the online journal, Internet Archaeology.

The article was originally published in the Journal of The Royal Archaeological Institute, under the title *Cottam: An Anglian and Anglo-Scandinavian settlement on the Yorkshire Wolds* (Richards 1999) and later the electronic version was created for Internet Archaeology as part of an experiment in electronic publication. The intention was to demonstrate how the interpretative synthesis of a journal article could be linked with the full corpus of digital data from which the synthesis was created. This would allow for analysis of the data as understood by the investigators to be published, while also providing access to the raw data for future use and interpretation, in line with best practices for the preservation of archaeological data as defined at the time (Austin *et al.* 2000). So both the archive and the interpretative information necessary to understand the archaeology at the site were easily accessed online for this research.

Many of the files and raw data used to create the reports are included in the archive in their original formats, including the resistivity and magnetometry data as DAT files, the geophysics plots with geo-referencing data as TIF files, the report illustrations as GIF files, the metal detector and excavation finds as JPG files and the vector drawings as DWG/DXF/DWF files. Metadata is also included to help the user understand how the files are structured. Database files are available in archival TXT format, and the entity relationship diagrams have also been included for anyone wishing to reconstruct the various databases. The raw dataset has been published in its entirety, using non-proprietary formats whenever possible for long-term accessibility.

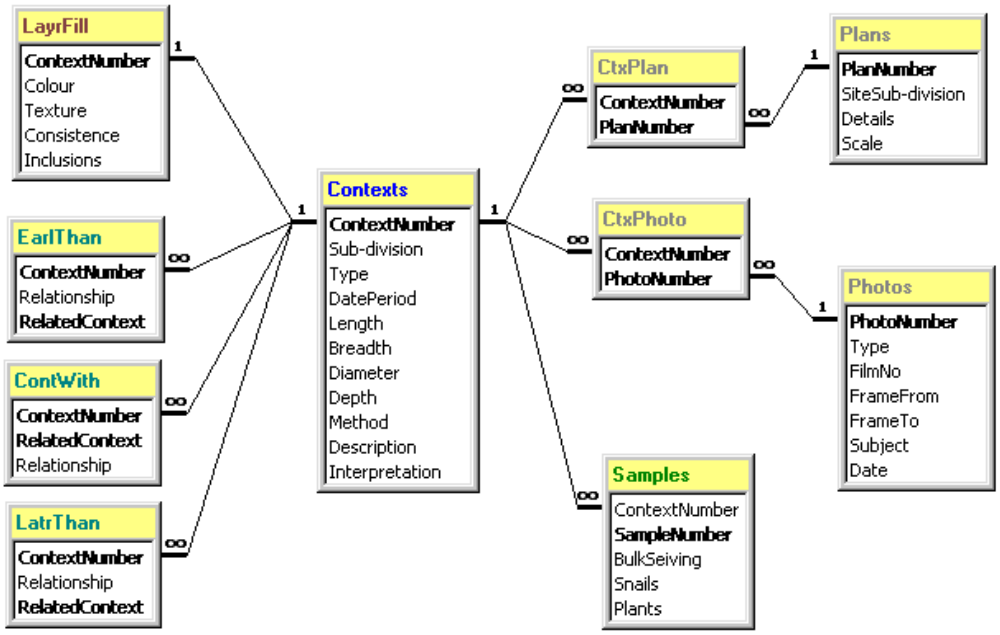


Figure 46: The entity relationship diagram for the context database from the *Burrow House Farm, Cottam: an Anglian and Anglo-Scandinavian Settlement in East Yorkshire* archive held by the Archaeology Data Service. (doi:10.5284/1000339) (Richards 2001c).

This research focussed primarily on the TXT files and DWG files. Vector plans are available for COT93.1, COT93.3, COT95, the Cottam B study area, and a plot of the cropmarks found at the site. The database files consist of the context data for the COT95 excavation only, but the content of the finds database includes all the work carried out at Cottam B. After evaluating all of the data available in the

Burrow House Farm, Cottam archive, the DWG file from the COT95 excavation was chosen to carry forward for use with this research, as it was the only trench containing evidence for Anglo-Scandinavian activity. The availability of the additional context data, which could be incorporated with the data from the DWG file, also made it the richer potential dataset for experimentation when compared to the COT93 datasets. The context data for COT95 includes a table with data and descriptions for each of the contexts, including tables defining the 'later than', 'contemporary with' and 'earlier than' spatial relationships between the contexts, and a layer/fill table. There are also tables associating sampling, photos and plans with their relevant contexts.

The COT95 plan drawing consists of vector polylines of the cuts of the major excavation contexts, with separate layers for Phase IIb and Phase III. The drawing was created for illustrative purposes only, so no context or annotation information is associated with the polylines and polygons directly. Contexts are simply labelled with their context numbers drawn either on top of, or next to them. This makes the COT95 drawing typical of vector plans created for publication rather than analysis.

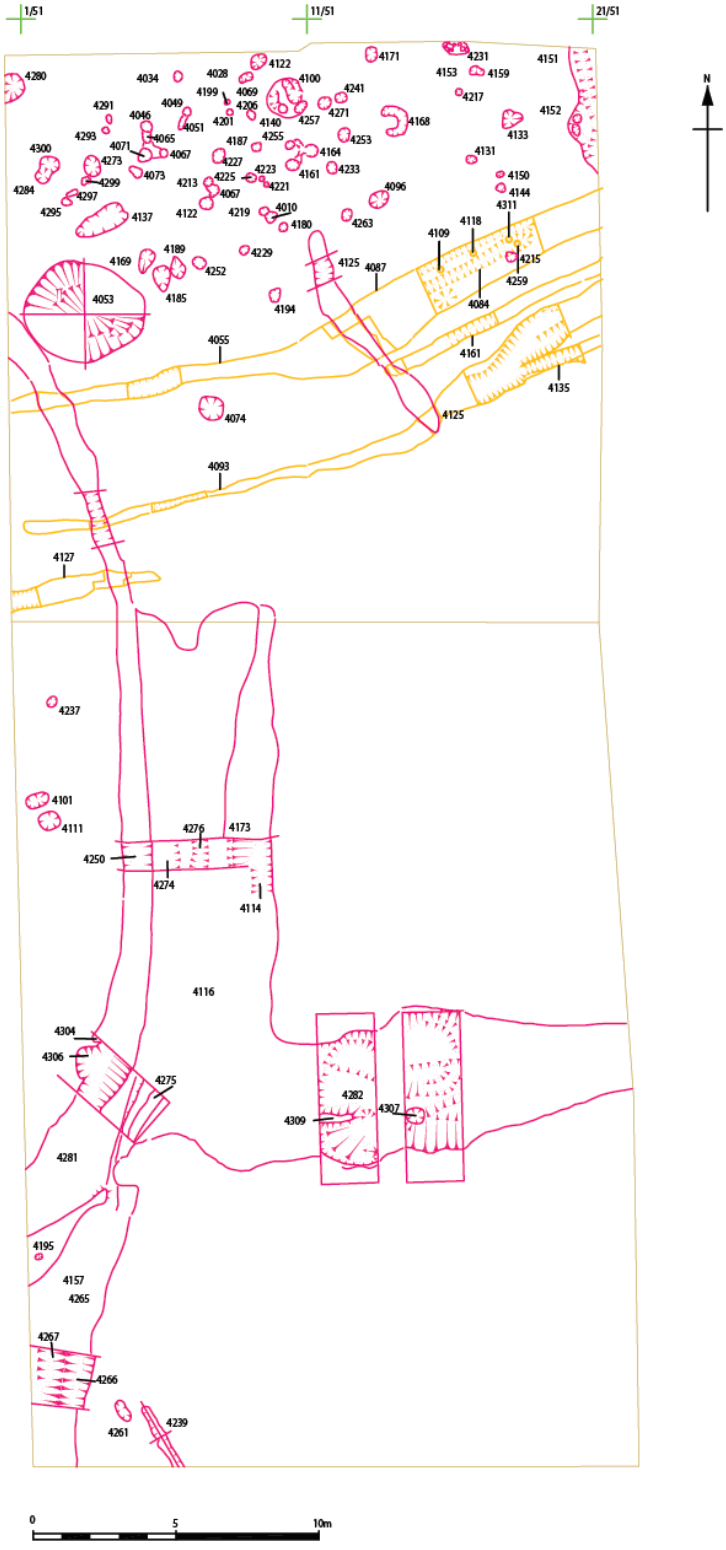


Figure 47: Plan drawing from the COT95 excavation trench. Contexts from the Period IIB: Anglian Phase B are shown in yellow, and contexts from the Period III: Anglo-Scandinavian Phase are shown in pink. From the *Burrow House Farm, Cottam: an Anglian and Anglo-Scandinavian Settlement in East Yorkshire* archive held by the Archaeology Data Service. (doi:10.5284/1000339) (Richards 2001c).

To make it ready for Semantic Web use, the drawing had to be cleaned and prepared. Using AutoDesk's AutoCAD 2008, all extraneous information was removed, including the hachures and context number labels. Then all the contexts had to be identified and converted into closed polygons where necessary. Polygons with dashed lines indicating 'edge uncertain' were converted to solid lines, and polylines representing contexts extending beyond the excavated area were truncated at the excavation wall to form closed polygons. Notes were made about these changes for annotation later in the process.

Although the size of the excavation trench probably did not require assigning a projection to the data, the CAD drawing was then brought into GIS using ESRI's ArcGIS 9:ArcMap 9.3.1 to georeference and project the data. New fields were added to the annotation table for context numbers, drawing notes, the x and y coordinates for the centre point of the context, along with calculations for the area and perimeter. As no attribute data was initially part of the drawing, each context was identified by hand and its context number added to the table. Any context containing 'edge uncertain' data lost by making a closed polygon, or was truncated where the context extended past the edge of the excavated area (or both) was noted in the table as well. Then the relevant data from the GIS attribute table was then exported using the Geography Markup Language (GML) format using FWTools.

Figure 48: Plan drawing (in red) from the COT95 excavation trench, georeferenced and projected in ArcMap 9.3.1, showing its position relative to the Burrow House Farm buildings.

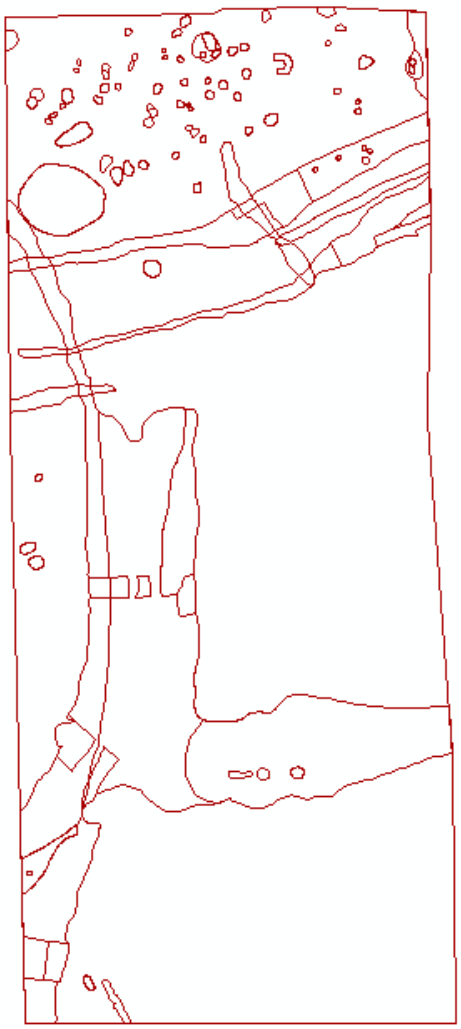


Figure 49: Plan drawing from the COT95 excavation trench as created within the GIS.

FWTools is a set of open source tools for working with GIS data created by Frank Warmerdam. It consists of several kits, including the Geospatial Data Abstraction Library (GDAL) within which is a translation library for GIS vector data called the OGR Simple Feature Library (OGR) (Warmerdam 2011). As the data from Cottam and Hungate is already in vector format, the *ogr2ogr* translation tool was used. Once in GML, the data was translated into CSV using a small, bespoke program written in Java by Michael Charno, called the STELLARPreloader. The STELLARPreloader converts the GML into CSV, ready for processing by the STELLAR tool (see page 21). Specifically it converts the geospatial information about each context from the GML file into the Well-Known Text

(WKT) format. The STELLARPreloader requires the context number field to be declared explicitly for the extraction, but the choice of other fields drawn from the attribute table is customisable. In the case of Cottam, the fields chosen were: Area, Perimeter, CentroidX, CentroidY, Centroid (a comma delimited concatenation of CentroidX and CentroidY), DrawNote (where changes made to the context polygons by the author were noted) and Phase. The STELLARPreloader then converted the data into CSV format for import into the next phase of transformation into RDF.

4.3.2 The data from Hungate

The data from Hungate was exported from the Integrated Archaeological Database (IADB), which is the bespoke data management system developed by Michael Rains at the York Archaeological Trust (YAT). The data is currently unpublished, and permission was kindly given by YAT for its experimental use as part of this research. The IADB is the in-house system used at YAT, but is also made available for download without cost, and is used by several other academic and commercial archaeological field units (Rains 2011). Much of the development of the IADB has been through Rains' partnership with the Silchester Town Life project based at the University of Reading. This collaboration has resulted in original work in several areas, including digital data capture (Fisher *et al.* 2009), virtual research environments (Rains 2007) and experiments in concurrent excavation, analysis and publication (Clarke *et al.* 2002).

Whereas the Cottam archive represents a complete archaeological dataset conforming to best practice principles using traditional software and methods (and specifically commercial software designed for other disciplines, but adapted for use by archaeologists), the IADB represents a best practice exemplar for a complete data management system, incorporating new technologies and ideas to specifically improve the archaeological research process, while maintaining focus on everyday practical use. Steve Stead and Pete Clark at the Scottish Urban Archaeological Trust (SUAT) developed the initial concept for an integrated

database for archaeology in the late 1980s. When Rains replaced Stead in 1989 he implemented the concept based on other projects he developed for Durham University and (what became) Historic Scotland, and created the IADB. At first the IADB was meant to be a framework for post-excavation analysis, but has become a full virtual research environment. In development for over 20 years, the IADB has been implemented with different programming languages and commercial software over the years, but is currently based entirely on open-source, Web server based solutions, in PHP with MySQL, and SVG for vector graphics (Rains 2011).

As the fundamental design principle of the IADB is that it is integrated, in order to access the data for use in this research, it had to be split into its constituent parts for export. Export options from the IADB for vector drawings include SVG and DXF, and for data held in tables, CSV and SQL. One of the most distinctive features of the IADB is its use of native SVG for all vector drawing. SVG being the W3C XML standard for vector data on the Web, it is also a non-proprietary vector data format now available as an export option across most popular vector-based drawing and spatial programs. With the advent of Internet Explorer 9, SVG is finally supported across all major browsers, which means it also displays natively on the Web in most cases. One of the original development goals for the SVG standard was to create an exchange format for vector data, and as such it seems an ideal archival format for Web and non-Web use alike; AutoDesk's DXF format being the *de facto* standard in absence of a true non-proprietary format. While the major CAD, GIS and vector drawing programs now have support for the export of data in SVG format, import support is still lacking, and until this is remedied it cannot be considered archival (Meng 2008, 1019). As momentum around SVG builds however, the IADB will be perfectly placed to take advantage of it.

Because SVG cannot be imported natively into AutoCAD 2008 or ArcMap 9.3.1, the vector plan drawing exported from the IADB in DXF format was

used. Additional data about the contexts was also exported in both CSV and SQL format, including a table containing the stratigraphic relationships between the contexts (limited to 'later than'). The data in CSV format was used for this research. As only the on-going excavation in Area H2 has yielded data from the Anglo-Scandinavian period of occupation at Hungate, this was the dataset exported by Rains. The dataset should not be considered complete however, as only the data from the deep trench and a small buffer zone surrounding it has been processed and input into the IADB as of 2011. More is waiting to be input, and still more Anglo-Scandinavian material will certainly be found before the excavation is finished. While beyond the scope of this research, far more information is available about the contexts from within the IADB, including finds data, images, and additional documents relating to the site and bibliographic references. The IADB also allows contexts to be grouped together as sets, and sets grouped together within phases for post-excavation analysis (Rains 2011), so the IADB provides rich potential for easy incorporation into the Semantic Web.

To make the Hungate drawing ready for conversion into RDF, it was also tidied and prepared. Using AutoDesk's AutoCAD 2008, all extraneous information was removed, including the hachures and spot height measurements. Unlike the Cottam drawing, the Hungate drawing consisted entirely of closed polygons, so decisions about how to interpret areas where edges were uncertain or truncated by the limits of the excavation, were made by staff at YAT before the data was exported.

Figure 50: The location of the data from the deep trench from Hungate (in red), as exported from the IADB and projected in GIS.

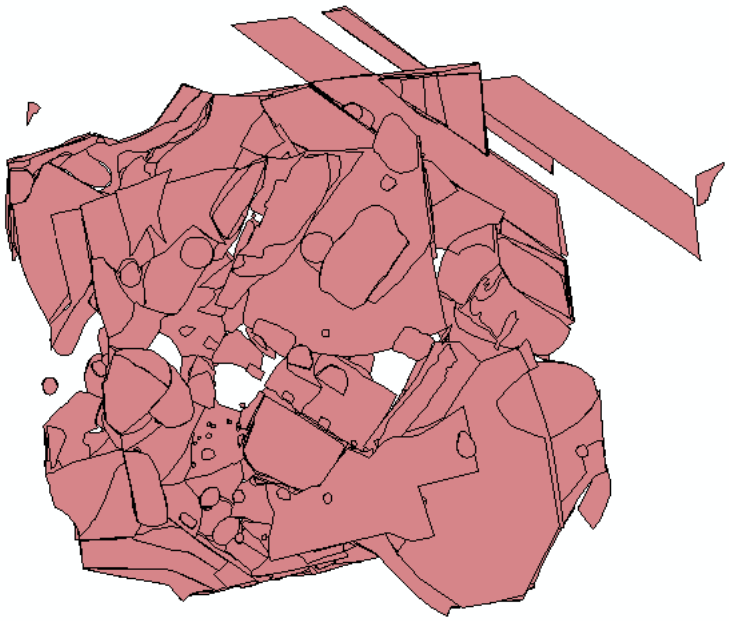


Figure 51: The data from the deep trench from Hungate, as exported from the IADB, showing each context as a closed polygon.

As with the Cottam drawing, the size of the excavation trench probably did not require assigning a projection to the data, but the CAD drawing was brought into GIS using ESRI's ArcGIS 9:ArcMap 9.3.1 to georeference and project it. Context numbers were already included in the annotation table as exported from the IADB, but two columns were added to use the 'calculate geometry' function to capture coordinates for the x and y centroid of each context, along with calculations for area and perimeter. From this point the process was the same as with the Cottam drawing, with the relevant data from the GIS attribute table being exported as GML, and processed through the STELLARPreloader Java application. The information about each context was then extracted from the GML file, and the additional WKT information for Hungate included: Area, Perimeter, CentroidX, CentroidY and Centroid. The STELLARPreloader then converted the data into CSV format for export into the next phase of transformation into RDF, bringing both drawings to the same point in the workflow process.

4.4 The domain ontology

In order to create data ready for Semantic Web use, two major components must come together, though the order in which they are implemented varies. First there is the need to bring all the data into a *common data model*, which for the Semantic Web is always RDF. This is not to be confused with terms like RDF/XML, N-Triples, N3 or Turtle, which are different serialisation formats for RDF (different ways to convert RDF data into a structured and storable format) (Hebeler *et al.* 2008, 74). Data derived from a variety of RDF serialisations can all be combined in a single RDF-store, as long as they conform to the RDF data model. Then there is the need to bring all data into a *common knowledge model*, which is typically referred to as an ontology (Allemang and Hendler 2008, 1). Within the Semantic Web, the term ontology has become a general term for a spectrum of classifications including taxonomies, thesauri, and conceptual models (also known as conceptual reference models), and has a range of semantic strength (listed weak to strong in this case). Ontological strength refers to how rich the relationships between the

data can be. The stronger the ontology, the more complex the relationships, and when an ontology refers to a particular area of knowledge, like archaeology, it is called a domain ontology (Daconta *et al.* 2003, 156-67).

As this project falls within the broad category of Cultural Heritage, the ISO standard domain ontology (ISO 21127:2006) most appropriate for archaeology is known as the CIDOC-CRM. The CIDOC-CRM was developed by The International Committee for Documentation of the International Council of Museums (ICOM-CIDOC) Documentation Standards Group (CIDOC CRM 2010). CRM refers to the term Conceptual Reference Model, meaning it is an ontology with strong semantics, and can express rich and complex relationships between the data. An extension of the CIDOC-CRM for archaeology, called the CIDOC CRM-EH (or just the CRM-EH), was recently developed as part of a collaboration between the University of Glamorgan Hypermedia Research Group and English Heritage. It is compatible with the single context recording standard and fieldwork as typically carried out in the UK, and as such represents an appropriate domain ontology for use with the datasets in this research.

There are a variety of ways to create data that conforms to the RDF data model. Data is represented using a wide variety of formats, but most commonly within relational databases, XML documents or tab/comma delimited data in text or ASCII format, and this data must be translated with its existing semantics intact (as much as possible) into RDF (Hebeler *et al.* 2008, 301-7). This is not a straightforward process, and is dependent on the nature of the original data, how the RDF data will be used, and the translation tools available. Typically, this has been done using the tools within a Semantic Web *framework*, like Jena or Sesame, and often requires a higher level of computing expertise than most archaeologists are comfortable with.

Several automatic conversion tools have been developed for general use, which allow translation to be carried out without the need to set up a *framework* (Byrne

2010, 75), but a new project called Semantic Technologies Enhancing Links and Linked data for Archaeological Resources (STELLAR) has recently developed tools which allow archaeological data, created using single context recording, to be both translated into the RDF data model, and aligned to the CRM-EH knowledge model in one step (STAR 2011). It also facilitates the assignment of Uniform Resource Identifiers (URIs) for publication as Linked Data, meaning it has gathered some of the most challenging aspects of making archaeological data ready for Semantic Web use, and made the process accessible for archaeologists with a far greater range of computing expertise. In addition to this obvious boon, there are compelling interoperability reasons for using tools like STELLAR which go beyond making the Semantic Web more accessible to non-specialists. This section will outline the process of taking the data from the Cottam and Hungate drawings through translation into the RDF data model, aligned to the CRM-EH ontology, with Linked Data URIs, using the STELLAR tool; thereby making it ready for the next step in its Semantic Web journey.

4.4.1 The CIDOC-CRM

While the CIDOC-CRM is appropriate for use with archaeological data, it is a very broad brush meant to paint across the common domains of the Cultural Heritage sector, and as such is not meant to include terminology or relationships specific to the disciplines found within it. In other words, the CIDOC-CRM is meant to be a lowest common denominator domain ontology for Cultural Heritage, so any ontology built on top of it for a specific discipline, like archaeology, will retain a basic level of interoperability across the Cultural Heritage sector at its core. This conforms to ontological re-use best practices, and should ensure good interoperability between disciplines related to archaeology as the Semantic Web develops. In practical terms this means:

> The CRM is a formal ontology which can be expressed in terms of logic or a suitable knowledge representation language. Its concepts can be instantiated as sets of statements that provide a

> model of reality. We call any encoding of such CRM instances in a formal language that preserves the relations between the CRM classes, properties and inheritance rules a "CRM-compatible form". Hence data expressed in any CRM-compatible form can be automatically transformed into any other CRM compatible form without loss of meaning. Classes and properties of the CRM are identified by their initial codes, such as "E55" or "P12" (Crofts *et al.* 2010, iv).

The CIDOC-CRM is made up of 86 classes (designated by the letter E, as they were known previously as entities), 137 properties (designated by the letter P), and the inheritance rules providing the structure of the ontology. As the full CIDOC-CRM is too broad for most purposes, subsets can be designated, as long as they meet a minimum criterion called a 'reduced CRM-compatible form' (Crofts *et al.* 2010, iv). Minimum standards are also set for import and export compatibility, so data can be reliably considered interoperable without loss of meaning. The data provider is meant to make their level of compatibility explicit by stating a 'compatibility claim declaration' for those wishing to consume the data as well. (Crofts *et al.* 2010, vi-ii).

The CIDOC-CRM is an object-oriented semantic model, so it is readable and comprehensible to humans, but is also designed for conversion into machine-readable syntaxes for encoding semantic metadata like RDFS, DAML+OIL, OWL, etc. It is also designed to be compatible with traditional relational or object-oriented schemas, and instances can be used with encodings like RDF, XML, DAML+OIL and OWL (Crofts *et al.* 2010, viii). As such, implementations like the Erlangen CRM/OWL (ECRM) represent interpretations of the CIDOC-CRM as used in practice. The ECRM was created using OWL-DL, which is subset of the full OWL language specification (the restrictions fundamental to OWL-DL makes the logic of OWL decidable, and thereby allows complete reasoning - see Chapter 2) (Hebeler *et al.* 2008, 159). Implementation of the ECRM has attempted to be as

close as possible to CIDOC-CRM specification (ECRM 2011), and was therefore chosen as the basis for STELLAR implementation of the CRM-EH.

4.4.2 The CRM-EH

The CRM-EH grew from a data modelling project called Revelation, led by Keith May for the Centre for Archaeology (CfA) at English Heritage. Revelation was meant to help bring cohesion to the many disparate data systems which had been designed for use by the English Heritage field unit over 25 years, and initially consisted of a series of data flow diagrams and entity relationship models to assess what sorts of data the CfA was using and producing. As patterns began to emerge, attention was given to some of the new Semantic Web ideas and technologies in development in the early 2000s, like domain ontologies. The decision was made early on not to attempt to create an ontology for use by all archaeological systems, but rather to model the data as created and used by the CfA with a reasonable and consistent level of granularity. The first application of the Revelation modelling project was as a planning tool for future systems design at the CfA, but the project moved forward always with the thought of making the work available to wider group of users outside the CfA (Cripps *et al.* 2004; Cripps and May 2010, 57-60; May 2006; May and Cross 2004).

Seeing the potential of a domain ontology as a way to provide a relational language which could be used across all CfA data, it was first necessary to determine whether such an ontology already existed. While no such ontology was available for the archaeological domain, the development of the CIDOC-CRM as a standard for Cultural Heritage was well known; information about it having been presented at the CAA conference for several years by that time. Upon examination, the high level concepts were found to be compatible with the archaeological domain, and after consultation with members of the CIDOC-CRM Special Interest Group, it was determined the best course of action was to create extensions to a ‘reduced CRM-compatible form’ of the CIDOC-CRM specific to archaeology, rather than start from scratch. This group of extensions was dubbed

the CRM-EH. Once created, the decision was made to further test its applicability and potential across a greater range of archaeological datasets, beyond those generated by the CfA. UK research council funding was secured in 2007, and English Heritage partnered with the Hypermedia Research Unit at the University of Glamorgan on the Semantic Technologies for Archaeological Resources (STAR) project (May *et al.* 2008).

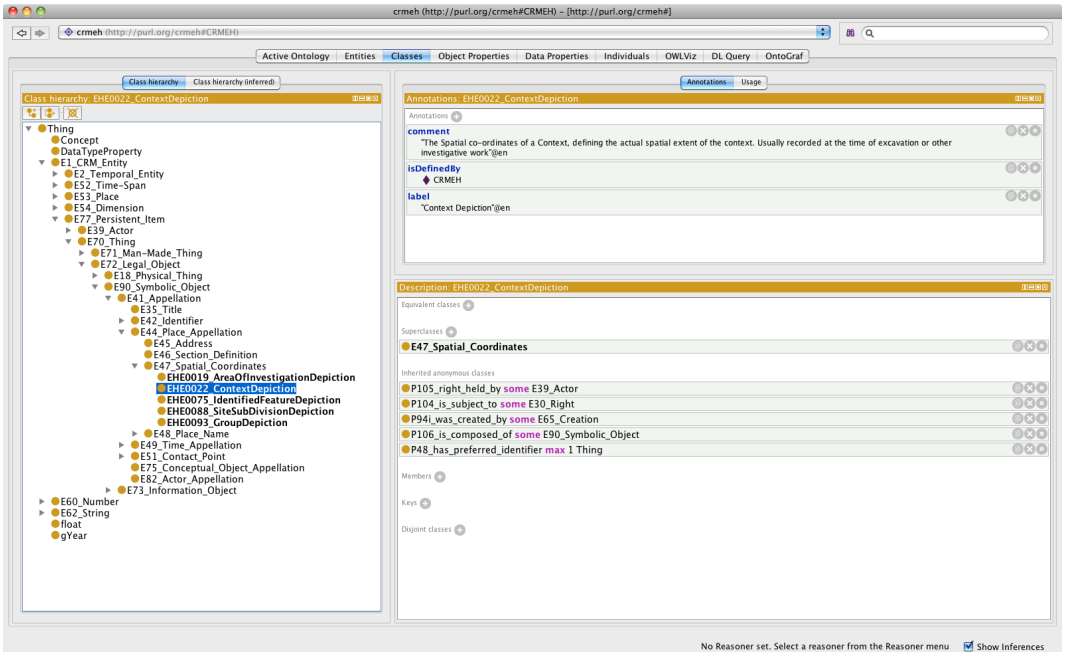


Figure 52: Screenshot of the CRM-EH in the Protegé ontology editor, showing the where EHE0022 Context Depiction, the class assigned to the spatial coordinates for a context fits into the structure of the ontology. It falls with E47 Spatial Coordinates. EHE refers to the CRM-EH, whereas E refers to the Erlangen implementation of the CIDOC-CRM. This illustrates the way the CRM-EH conforms to best practices by extending the most appropriate existing ontology for maximum interoperability, rather than developing a new ontology.

The STAR project had several objectives to explore the usefulness of the CRM-EH in a variety of ways. It set out to test whether the domain ontology could be used to make grey literature more accessible to broader research, whether datasets representing disparate types of software usage, stages of archaeological project management and archaeological time periods could be made interoperable, and if the increased access to grey literature could be incorporated. Once the data was combined, the value of the newly interoperable data would be shown through an

online demonstrator, using the kinds of multi-concept querying that is typically beyond the scope of relational databases. The findings were then meant to undergo evaluation and information about the project outcomes disseminated (STAR 2011; May *et al.* 2008; May *et al.* 2010).

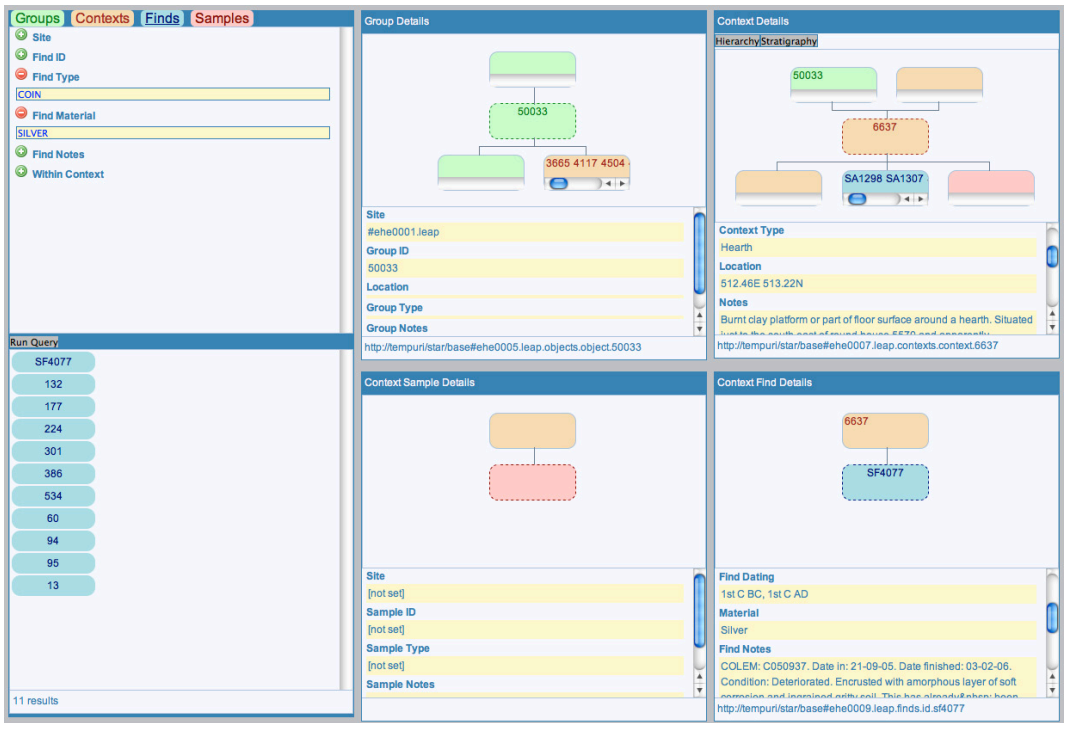


Figure 53: Screenshot of the STAR research demonstrator, showing the results of a search across all the excavation datasets for silver coins. The query returned 11 results. The first result, small find 4077 (SF4077) from Silchester, is identified within the Context Find Details, which is shown as being found within context 6637. Clicking on 6637 populates the Context Details box above and shows the relevant stratigraphic relationships, including the context's Group Details.

In addition, the STAR project incorporated Simple Knowledge Organization System (SKOS) concepts to work with the CRM-EH. SKOS was developed by the W3C to allow relationships between existing domain glossaries, taxonomies and thesauri to be defined in a consistent way, and its incorporation broadened the usability of data mapped to the CRM-EH (May 2009). Upon the successful completion of the STAR project, one of the main outcomes was an appreciation of the degree of specialist knowledge required to carry out these practical objectives, including mapping data to the CRM-EH and translating it to RDF. As a result, further UK research council funding was secured to develop tools to bring

this functionality to non-specialist users. The new project was called Semantic Technologies Enhancing Links and Linked data for Archaeological Resources (STELLAR) (STAR 2011).

4.4.3 Using STELLAR

Several methods of experimentation were tried initially for extracting, translating, mapping and manipulating the data created from the Cottam and Hungate GIS drawings. Initially, FWTools was used to extract and translate the shapefiles and attribute data from the GIS files into GML, allowing it to be brought into a PostgreSQL database, where mapping to the CRM-EH could be automated using SQL, and then called by a Java class from within a Jena *framework* to create an RDF file. This work was challenging, interesting and informative, but ultimately resulted in an appreciation of the real difficulty inherent in mapping data to a domain ontology, and translating it into RDF for a non-specialist user. Not only was much of the work done on the command line using UNIX, working with Jena required a solid understanding of Java, and working with PostgreSQL required learning about how non-WYSIWIG database structures and querying with SQL.

This, coupled with understanding the nature and requirements of the Semantic Web is well beyond what most non-specialists archaeologists would wish to attempt. In addition, it illustrated areas for real concern with regard to true interoperability, and attempts at mapping to the CRM-EH could not be done with confidence. How closely does one person's interpretation of the CRM-EH match the interpretations of others? In addition to the logistical challenges of mapping data and translating it to RDF, does someone wishing to consume the data have to vet the mapping before they feel confident it is interoperable with their other data, despite being mapped to the exact same domain ontology? What level of specialist knowledge would that require? These were worrying questions that arose during the experimentation process with the Cottam and Hungate data. Fortunately, the STELLAR project was well underway at the same time these experiments were being carried out, and a prototype version was available for trial with this research.

STELLAR was developed by the Hypermedia Research Unit at the University of Glamorgan with the ADS and English Heritage. The objectives of STELLAR were to create best practice guidelines for mapping and extraction of archaeological data into RDF, aligned to the CRM-EH, and develop an application for non-specialist users. In addition, STELLAR was designed to take this work a step further, by allowing URIs to be designated as part of the translation and mapping process, thereby making the result ready for publication as Linked Data. The project was also required to be evaluated, and the findings disseminated to the wider community (STAR 2011), which was how it was able to be included in this research prior to the completion of the STELLAR project itself.

Two versions of the application were created. STELLAR.Console is a freely downloadable command line application, which can be used for data import originating in SQL or CSV format for mapping to the CRM-EH and conversion into RDF/XML syntax. STELLAR.Console allows more advanced users customisation, control and the ability to set up batch processing for large amounts of data and custom templates if necessary, while ensuring the result still conforms to a consistent mapping of the CRM-EH. STELLAR.Web is a browser-based application that provides a subset of the functionality of STELLAR.Console.

STELLAR.Web requires the data already be in CSV format, and column headings within the CSV file must be changed to correspond with the matching terms as set out in the STELLAR guidelines, but no specialist knowledge of the CRM-EH (or ontologies in general), or RDF is required. CSV data is uploaded from a local file, the template appropriate to the type of archaeological data chosen, the file submitted for conversion, and the resulting RDF/XML file becomes available for download. If the data is also being prepared for publication as Linked Data using a predefined system of URIs, these can also be defined during the process, but are not required (Binding 2011; Tudhope *et al.* 2011a; Tudhope *et al.* 2011b).

Figure 54: Screenshot of the STELLAR.Web browser-based application, showing the context data from the Hungate excavation about to be converted using the CRMEH_CONTEXTS template.

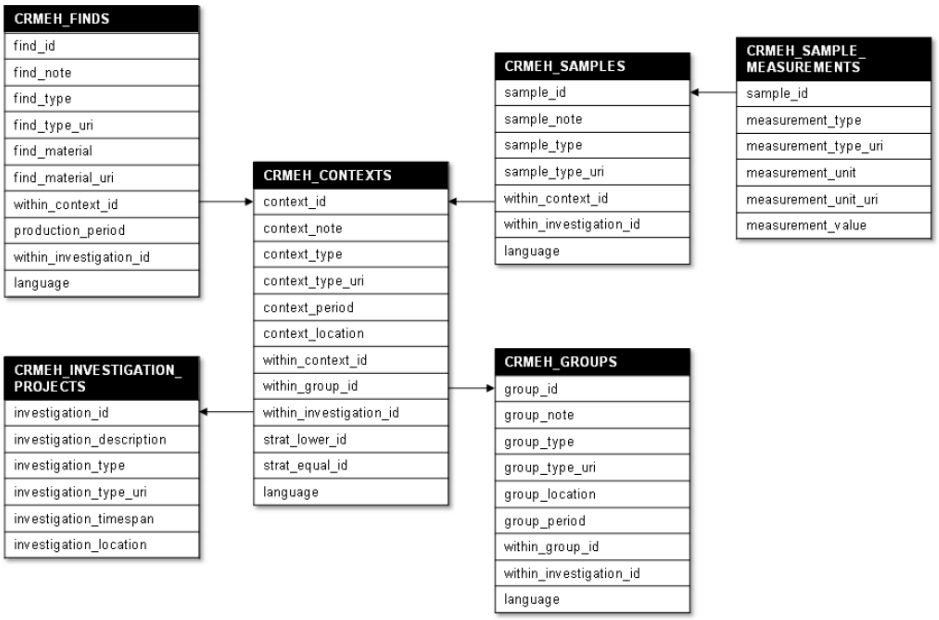


Figure 55: The structure and relationships of the STELLAR templates. The CRMEH_CONTEXTS and CRMEH_INVESTIGATION_PROJECTS templates were used with the Cottam and Hungate data. Reproduced from the STELLAR mapping and extraction guidelines (Binding 2011).

After using the STELLARPreloader Java application to convert the Cottam and Hungate data to CSV format, they were ready for conversion into RDF using STELLAR. As the STELLAR templates CRMEH_CONTEXTS and CRMEH_INVESTIGATION_PROJECTS were sufficient to map the Cottam and Hungate data, and the data was already in CSV format, STELLAR.Web was all that was necessary to do the conversion. In addition, the ADS recently developed the URI set they will use for their Linked Data publication as part of their participation in STELLAR. As the Cottam dataset is already fully published and held as an ADS archive, Linked Data URIs were assigned to allow the results to be added to the existing Cottam archive, which would make available a further type of download option in the future. Use of STELLAR not only removes the necessity for a huge amount of specialist technical knowledge needed to prepare archaeological data for the Semantic Web, it also takes the guesswork out of aligning data to an ontology. It would be simple enough for anyone publishing their archaeological data to include their STELLAR mapping file, and a user who knows nothing about the Semantic Web could check it and determine whether they agree that the column headings chosen for the data by the publisher are a good match for the fields in the appropriate STELLAR template, and therefore their own data.

4.4.4 Assigning URIs

While assigning URIs is not a requirement for using the STELLAR templates, creating appropriate URIs for Linked Data publication should be done whenever possible. If data is created for publication where URI naming conventions are already established, then an archaeologist using STELLAR need only find out how their data fits into the larger convention and designate it during the process of setting up the template. If a convention has not yet been established, then it becomes important to understand how and why URIs are assigned before proceeding. URIs themselves have been a source of confusion over the years. So much so that the W3C has actually changed the way they are defined to reflect the popular way people think about them, rather than their original, formal definition

(World Wide Web Consortium 2001). Today, the importance of the term URI lies in conveying the idea that something is meant to be persistent, rather than to define its place at the top of a structural hierarchy of Web concepts. A URL is meant to say where something is living at the moment, but a URI should be used to define the characteristics of a resource which shouldn't change (Berners-Lee 2000, 68).

Because the ADS was designed specifically as an archive to hold archaeological data persistently, it is appropriate that any Semantic Web data formally published from the archive use the ADS URI naming convention. The naming convention chosen by the ADS conforms to best practice standards as set out internationally by the W3C (World Wide Web Consortium 2008b), and being a UK archive, by the UK Cabinet Office (UK Chief Technology Officer Council 2009). As the ADS publishes URIs as Linked Data which need to be fully resolvable, a 'data' subdomain has been established within the ADS DNS, designating all Semantic Web data will be published under http://data.archaeologydataservice.ac.uk. Any data formally archived with the ADS will use its Digital Object Identifier (DOI) number. As the already established DOI for the Cottam archive is 10.5284/1000339, the base URI for the Cottam data input into STELLAR was:

http://data.archaeologydataservice.ac.uk/10.5284/1000339/

The Cottam data generated by this research has been deposited back with the ADS using their URI naming conventions, and published as Linked Data with the rest of the STELLAR archive. The URIs are live and fully resolvable, but as STELLAR was a development project of the ADS, the contents of the STELLAR Linked Data RDF store has not been incorporated into the general ADS archive. If the ADS decides to publish the Linked Data from this research directly as part of the existing Cottam archive, because the URIs are persistent, it can be moved at any time without needing to change them.

As the Hungate archive is not being formally published at this time, URIs for use with this research were created purely for demonstration purposes. As such, they have been given the URI http://www.diggingitall.co.uk:8080/data/ and housed in an RDF store set up temporarily for this research. The York Archaeological Trust maintains a persistent archive of their digital archaeological data, and may create their own URI set at some point in the future that would house the data from Hungate, but there are no plans to do so at this time. Once the data from Cottam and Hungate were processed using STELLAR.Web and output in RDF format, mapped to the CRM-EH with Linked Data URIs, it was time to work with it.

4.5 Working with the data in RDF

In order to make the Cottam and Hungate data interoperable, they must be available for querying across both datasets simultaneously. While the future promise of the Semantic Web is the ability to query across data held in myriad places, here they will be combined into a single RDF store for demonstration. As this research is meant to explore whether archaeologists can use Semantic Web technology that is now freely available and technically accessible to non-specialists, generic RDF store software called AllegroGraph was used. AllegroGraph is Linux-based, and requires a knowledge of command-line UNIX for the initial setup of the server, but once created, the addition of the AGWebView interface provides a WYSIWYG environment to interact with the data.

Through AGWebView, users can easily create a named repository (grouped data within the RDF store), and populate it with the data created using STELLAR by simply uploading a local file using the built-in interface. AllegroGraph comes preloaded with the most common namespaces, like RDFS and SKOS, but it is simple to add the relevant namespaces specific to the STELLAR data, like the ECRM and CRM-EH. Once the data and the appropriate namespaces to define the relationships between the data are loaded, it is possible to browse the data within AGWebView. AGWebView contains several preloaded SPARQL queries, which bring up lists of the classes and properties in the repository. Data can be easily

indexed to allow free text searching, it is possible to save SPARQL queries that can be loaded and executed, and to download the results of those queries as CSV or XML for inclusion back into a relational database. The entire repository can also be exported in a variety of RDF syntaxes.

For those able to write SPARQL (or Prolog), AGWebView acts as a traditional SPARQL endpoint for querying the data. For those able to script with JavaScript (or Lisp) the AllegroGraph server can also be controlled using the AGWebView interface, and existing scripts can be uploaded and executed, or written directly. This level of access would be sufficient for most expert users (unless they are happier programming in one of the many other languages with which AllegroGraph has client compatibility), and use of AGWebView means there is little need to interact with AllegroGraph directly once it is running. AllegroGraph also allows reasoning using RDFS++. RDFS++ includes reasoning capability for all the predicates found within RDFS and several key predicates from OWL (Franz Inc. 2011b). In addition, the fact that AGWebView is a Web interface means, not only does it have functionality sufficient for a range of user needs, it can be accessed over the Web at any time.

The primary function of AGWebView is to act as a SPARQL endpoint. SPARQL syntax feels familiar for those already using SQL for querying data in relational databases, but because SPARQL is querying against data in graph format, the nature of the queries can be quite different. While relational data is structured to conform to an Entity Relationship Model (ERM) and querying relies on this structure, querying the semi-structured nature of graph data has been described as querying the 'RDF haystack' (McCarthy 2005). If the 'Web of Data' is going to become as ubiquitous as the current 'Web of Documents', then WYSIWYG interfaces which execute SPARQL queries are going to have to be the norm, but for now, to work with RDF data, it will be useful for archaeologists to learn a bit of SPARQL. This effort should be rewarded with the ability to ask questions of the data which either had not been previously envisioned, or are more complex than would be possible with relational data.

Being able to visualise Semantic Web data is fundamentally important. The decentralised, semi-structured nature of graph data allows interpretation through the visual patterns it creates. The Semantic Web allows humans to *literally* see data differently. AGWebView provides three basic ways to visualise data. In node view, data is displayed within its subject-predicate- object structure, allowing users to click through the data and move in any direction, but it typically shows the data in disparate pieces. In graph view, dots (or nodes) are used to represent subjects and objects, and lines (or edges) are used to represent the predicates connecting them. This creates an overall picture of the data and relationships, which are communicated and understood visually. AGWebView also has a Google Maps interface, which allows georeferenced data to be visualised as a single point. Franz Inc., the developers of AllegroGraph and AGWebView, has also developed a more visually sophisticated RDF data querying and visualisation tool called Gruff for desktop use. The Cottam and Hungate data will be visualised using both AGWebView and Gruff.

AGWebView and Gruff allow the data to be visualised, and therefore analysed, but as the Cottam and Hungate datasets were created with Linked Data URIs, the data should be publishable as well. Once data aligned to the CRM-EH has been created in RDF/XML format with appropriate Linked Data URIs using STELLAR, and simply made available for download as static files, they can be considered properly published Linked Data. This is significant, because it demonstrates it is now possible for even non-specialist users to not only create and use Semantic Web data, but to publish it as well. Use of static files typically means downloading entire datasets, or predefined subsets of the dataset. If users want to query data to access only the subset they are interested in, it may be necessary to provide a Linked Data interface. AGWebView is not designed to be a Linked Data server, but freely available frontend software like Pubby, which was developed by Chris Bizer and Richard Cyganiak at the Free University of Berlin, can be used to add this service on top of AllegroGraph.

4.5.1 Creating and populating the RDF store

There are a number of generic RDF stores currently available, which can be used as standalone implementations, or plugged into a framework like Jena or Sesame if customised interaction is required (Hebeler *et al.* 2008, 155). Each has different strengths in functionality, usability, cost, choice of programming environment and support, and several were explored for use with this research. AllegroGraph by Franz, Inc. was eventually chosen, as it is available without cost for use with up to 50 million triples, has a good range of features, requires a limited amount of specialist knowledge to install and run, has a Web-based SPARQL endpoint and visualisation tool, and also has some interesting spatial and temporal features. If it had been more desirable to have both traditional RDMS and RDF operations in one data store, then Virtuoso (the freely available product created by OpenLink) would have been a good choice as well, or if money were no object, the Oracle 11g enterprise database with the Oracle Spatial extension and its native Semantic Web functionality would have been an even better choice. The features of Oracle Spatial will be explored further in section 4.6.

As this research was carried out using a Macintosh server running OSX, and AllegroGraph 4 is only available for the Linux operating system at this time, a 64-bit Linux Ubuntu virtual machine was created using Parallels software, and the AllegroGraph 4.3 version of the server software installed. This was fairly straightforward, even for someone with limited familiarity of command-line UNIX. More difficult was setting up the AGWebView interface on the virtual machine for use with the OSX Web server. AGWebView listens on default port 10035, and it was challenging to set up the correct port forwarding between the Ubuntu virtual machine, the Parallels interface and the OSX Web server, but it was certainly possible. While not robust enough for anything other than demonstration, the configuration has proved itself to be perfectly serviceable. The virtual machine had 4GB of dedicated RAM and the OSX server had a further 4GB of RAM available, and was able to run AllegroGraph easily, but obviously if it were being run on a dedicated Linux system, it would have been faster.

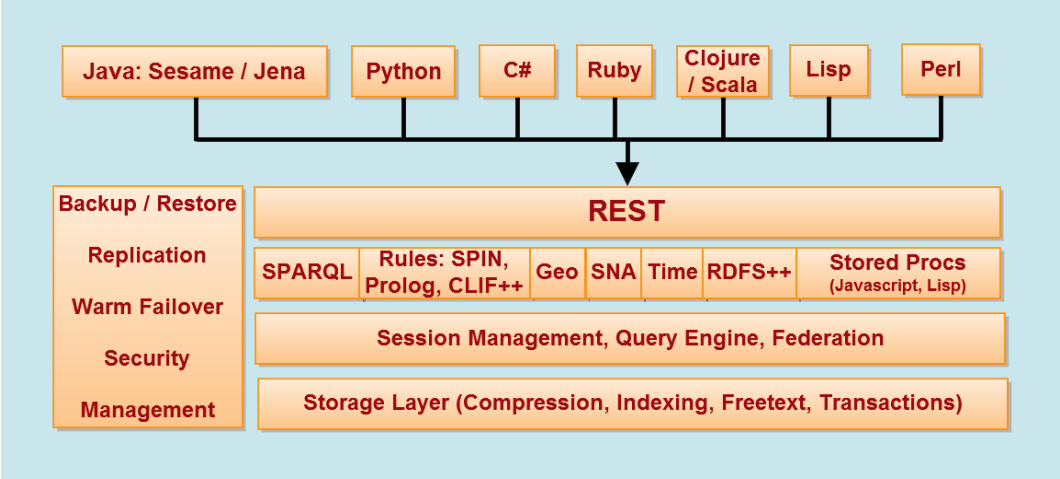


Figure 56: The underlying architecture of the AllegroGraph 4.3 RDF store. Reproduced from the AllegroGraph website. http://franz.com/agraph/allegrograph/.

Once AllegroGraph was installed with the AGWebView interface available live on the Web, a superuser account was created for the author and a read-only account setup for access by the thesis readers. Once logged into AllegroGraph, there are four menu choices: Catalog, Scripts, Admin and User. Catalog is where Repositories are created and accessed and Site Settings can be input. A single, empty repository called *Thesis* was created in Catalog, and a Google Maps key code input so the AGWebView mapping functionality could be used. Once inside a repository a new menu becomes available, including Overview, Queries, Scripts, Namespaces, Admin and User, along with checkboxes for Reasoning (to turn on the RDFS++ reasoning for queries), Long Parts (to display fully-expanded URIs) and Contexts (which displays a graph URI as the fourth element of any triple returned from a query) (Franz Inc. 2011c).

Inside the *Thesis* repository, the overview allows the user to see how many triples are in the RDF store, to explore the RDF store using some pre-defined SPARQL queries, to add or delete single sets of RDF triples, or to import groups of triples from local files, server-side files or to bulk-load them from online sources. Users can also access several other functions, including viewing and creating free-text indices, and viewing and editing the active standard indices.

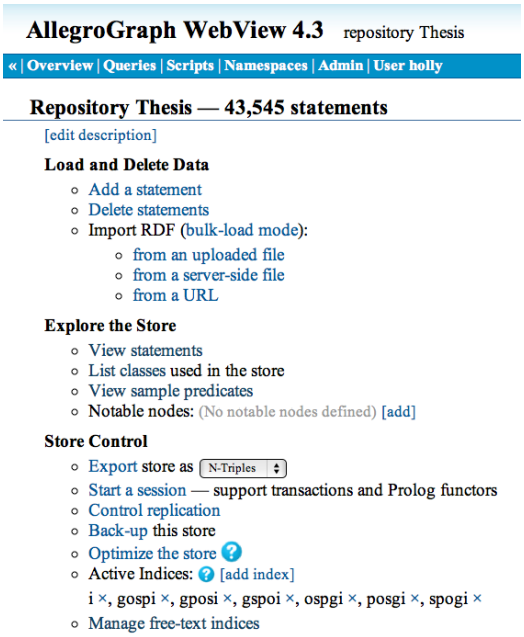


Figure 57: Screenshot of the superuser read/write access view of the Thesis repository within AllegroGraph, showing the 'Load and Delete Data' functions, a few preloaded SPARQL queries to 'Explore the Store', and other control functions.

As STELLAR generates files in RDF/XML syntax for download as local files, this was the option used for populating the *Thesis* repository. AllegroGraph 4.3 supports the upload of RDF data in N-Triple, N-Quad, RDF/XML, TriX and Turtle serialisation formats, and the CRMEH_CONTEXTS and CRMEH_INVESTIGATION_PROJECTS for Cottam and Hungate were each uploaded. Because each of the files has been structured exactly according to the CRM-EH using STELLAR, the data automatically 'stitches' itself together within the RDF store using the built in relationship between the CRMEH_CONTEXTS and CRMEH_INVESTIGATION_PROJECTS templates.

CRMEH_CONTEXTS
within_investigation_id

links to:

CRMEH_INVESTIGATION_PROJECTS
investigation_id

This illustrates how data from other templates could also be incorporated at any time. For example, if finds data were to be added, the template:

CRMEH_FINDS
within_context_id

would automatically link to the existing context data via:

CRMEH_CONTEXTS
context_id

This ability to grow and change in whatever direction is deemed necessary shows one of the real strengths of the Semantic Web, and why it could be particularly useful to archaeologists. The STELLAR templates could be expanded at any time and in any direction to accommodate the diversity of information with which archaeologists might like to work. As an example, a whole subset of finds templates could be created for faunal, human, or environmental remains which could be incorporated whenever a more specific level of finds detail was necessary. The current templates in STELLAR are meant to be a starting point, but finer grained templates could be developed by specialist groups and added at any time, as long as they conform to the same CRM-EH configuration, and can create a relationship with the existing templates. Thus, the data held within the *Thesis* repository is only the tip of the iceberg of what could be seamlessly added at a later date, either using more of the existing STELLAR templates, or new ones as they are created in future.

Once the data was brought into the *Thesis* repository it was necessary to add the namespaces for the various domain ontologies referenced by the Cottam and Hungate data, so it was properly disambiguated and resolvable. Defining namespaces also allows simplified views when working with the data through the use of prefixes. AllegroGraph comes with the most commonly used namespace references pre-loaded, including RDF (rdf), RDFS (rdfs), OWL (owl), Dublin

Core elements (dc), Dublin Core terms (dcterms) and SKOS (skos). Under the Namespace menu, it is possible to define additional namespaces, and the namespaces for the Elangen implementation of the CIDOC-CRM (ecrm) and the CRM-EH (crmeh) were duly added. Once the repository was populated, it was fully indexed for faster querying, and to allow free text queries. With the data and the appropriate namespaces and indexing in place, the *Thesis* repository was complete and ready for use, which for an RDF store typically means querying with SPARQL.

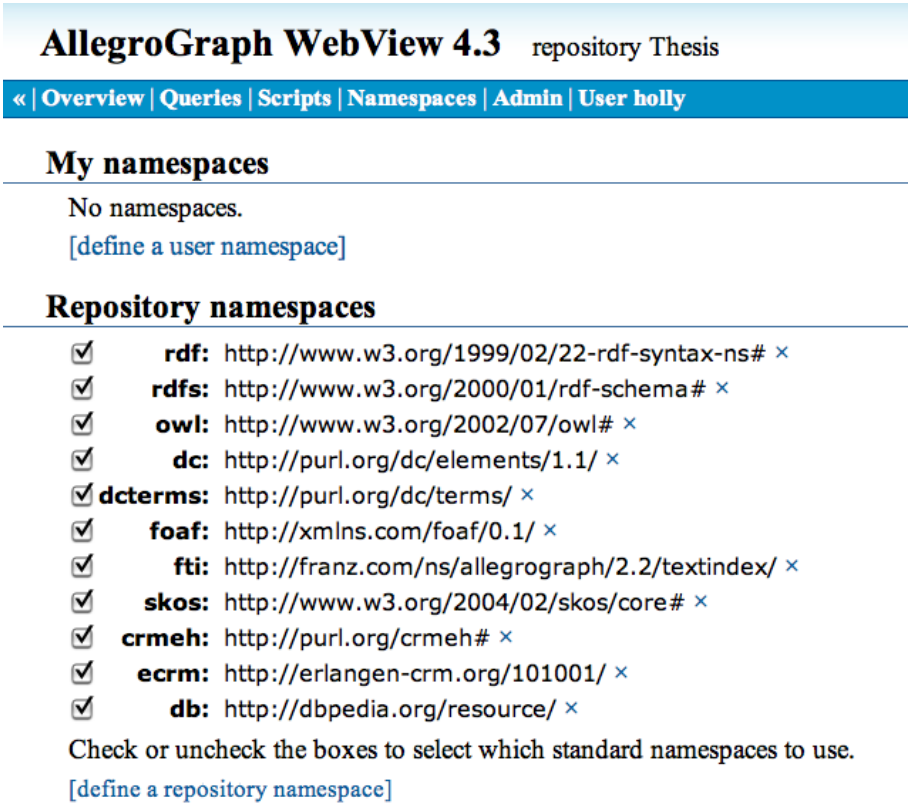


Figure 58: Screenshot of the Repository Namespaces, showing the CRM-EH and ECRM, along with the most commonly used namespaces.

4.5.2 Querying the data

The first W3C recommendation for the RDF data model and syntax specification was completed back in 1999 (World Wide Web Consortium 1999), but no formal way of querying the model was immediately defined. In 2004 the W3C created the RDF Data Access Working Group (RDF-DAWG) to consolidate the several

attempts to create a query language for RDF into a W3C recommendation. The first working draft specification was released later that year, which resulted in completion of an official recommendation in 2008, and was named using the recursive acronym SPARQL, for SPARQL Protocol and RDF Query Language. Once the initial recommendation was complete, a new SPARQL Working Group was formed in 2009 to develop SPARQL 1.1 (DuCharme 2011, 43), which at the time of this writing is in its final call for comments as a working draft, and nearing completion. Now that an official W3C recommendation for SPARQL has been developed, query interfaces, which are known as SPARQL endpoints, are becoming more and more common, either as a way to provide access to data held within a specific RDF store, or as a generic endpoint for querying data published from an external Linked Data store. Full implementation of SPARQL 1.1 will mean SPARQL will function not only as a way to query data and return desired results, but also manage and update the data in the RDF store.

AllegroGraph 4.3 currently only supports the new SPARQL 1.1 *Update* and *Subquery* functions, but is working to add new functions in successive releases (Franz Inc. 2011d). AGWebView provides a Web-based SPARQL endpoint for AllegroGraph's SPARQL engine, to allow querying of data within repositories held within its RDF store. Under the Queries menu, a user can create a new query, execute a saved query, and execute any recent queries created in the current session, or run a free-text query if any free-text indices have been defined. Queries can be written and saved in either SPARQL or Prolog with pre-defined limits on the number of results returned (Franz Inc. 2011c). Defining the structure of a SPARQL query is known as creating a 'triple pattern', and data returned from the query is said to have been 'matched' to the triple pattern, and multiple triple patterns within a single SPARQL query are known as a 'graph pattern' (DuCharme 2011, 3, 9). For those already conversant with SQL, SPARQL will feel a bit familiar, but the main query forms return quite different things. The workhorses of SPARQL are the *Select* and *Construct* queries. The *Select* query returns the variables and values of a query in table view (Feigenbaum and

Prud'hommeaux 2011). As an example, within the list of saved SPARQL queries in the *Thesis* repository is a very basic *Select* query called 'Select by Context Type', which looks like:

```
SELECT *
WHERE {
 ?ContextUID ecrm:P2_has_type ?ContextType
}
```

Running this query results in a list of all the contexts in the repository that have an archaeological feature type, and displays them in table format. Here is a snippet of the results:

ContextUID	**ContextType**
EHE0007_50808	E55_EHE0007_deposit
EHE0007_50809	E55_EHE0007_cut
EHE0007_4190	E55_EHE0007_layer
EHE0007_4007	E55_EHE0007_cut

This snippet of data illustrates a problem with the interoperability of the two datasets. While both projects use the word 'cut' to describe the negative features within the site's stratigraphy, at Hungate (context numbers 5xxxx) positive features are termed 'deposit', while at Cottam (context numbers 4xxx) are termed 'layer'. These terms could be made interoperable however, using the W3C Simple Knowledge Organization System (SKOS) namespace terms skos:closeMatch, skos:exactMatch or skos:altLabel, depending on the opinion of the archaeologist using the data (World Wide Web Consortium 2009c).

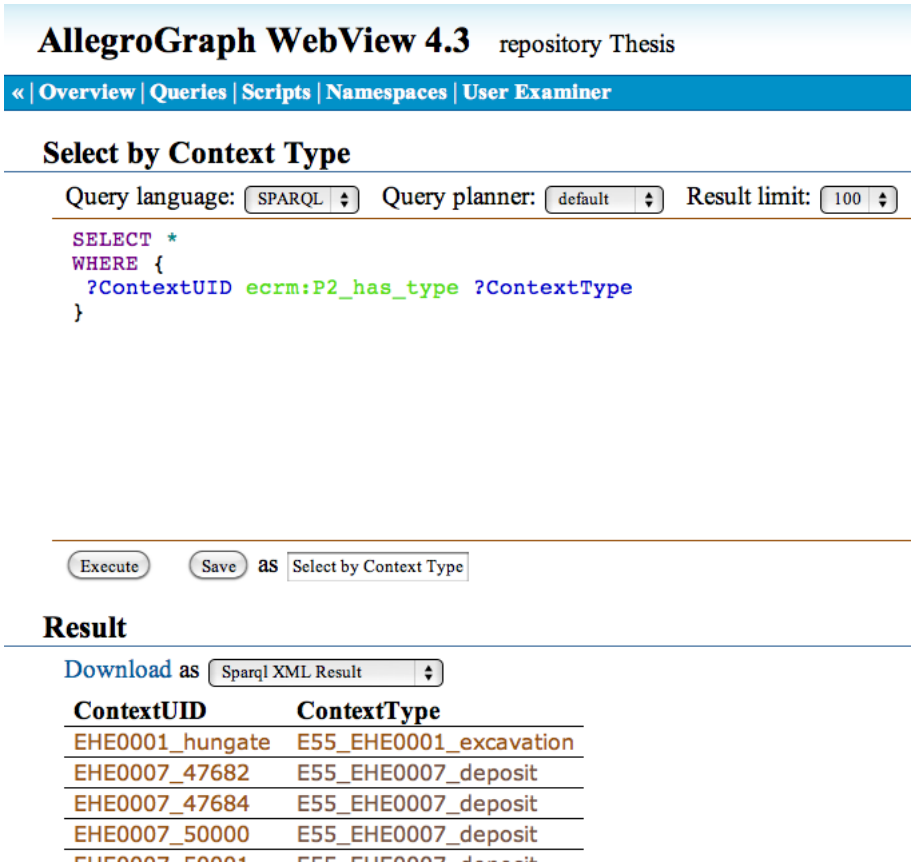


Figure 59: Screenshot of the 'Select by Context Type' query, showing a snippet of the result.

Because a *Select* query simply produces a table, the download options in AGWebView reflect the data's characteristics, and the only options are either a SPARQL XML or CSV file. These could easily be incorporated back into a relational database or any other traditional data structure if desired. In contrast, the *Construct* query doesn't just produce a result in a table, it creates a new subset graph of the selected data, which means it returns results as triples. The saved query 'Construct by Context Type' in the *Thesis* repository, asks for the 'Context Type' from the same data, but uses the *Construct* query:

```
CONSTRUCT {
?ContextUID ecrm:P2_has_type ?ContextType
}
WHERE {
?ContextUID ecrm:P2_has_type ?ContextType
}
```

The results look similar, but because *Construct* always returns full sets of triples, the Erlangen CIDOC-CRM predicate P2_has_type is also returned, and the same snippet of data looks like:

Subject	**Predicate**	**Object**
EHE0007_50808	ecrm:P2_has_type	E55_EHE0007_deposit
EHE0007_50809	ecrm:P2_has_type	E55_EHE0007_cut
EHE0007_4190	ecrm:P2_has_type	E55_EHE0007_layer
EHE0007_4007	ecrm:P2_has_type	E55_EHE0007_cut

While a *Construct* query may seem like a *Select* query that returns extra pieces of information that aren't really necessary, *Construct* is actually a much more powerful query form than *Select.* Because *Construct* keeps the relationships between the data intact as a subset graph of the full graph contained in the repository, the data can be downloaded in a variety of RDF serialisations, and it can be visualised as a graph using AGWebView (or any other Semantic Web graph data visualisation tool, like Gruff). The real strength of the *Construct* query however, is its ability take implicit relationships existing within the data, and make them explicit across multiple data sources. It is certainly possible to dynamically generate explicit data from implicit relationships within a relational database, but it would be very difficult to achieve across more than one relational database at once if their structures were not exactly the same, which is exactly what Semantic Web data is designed to do (DuCharme 2011, 112).

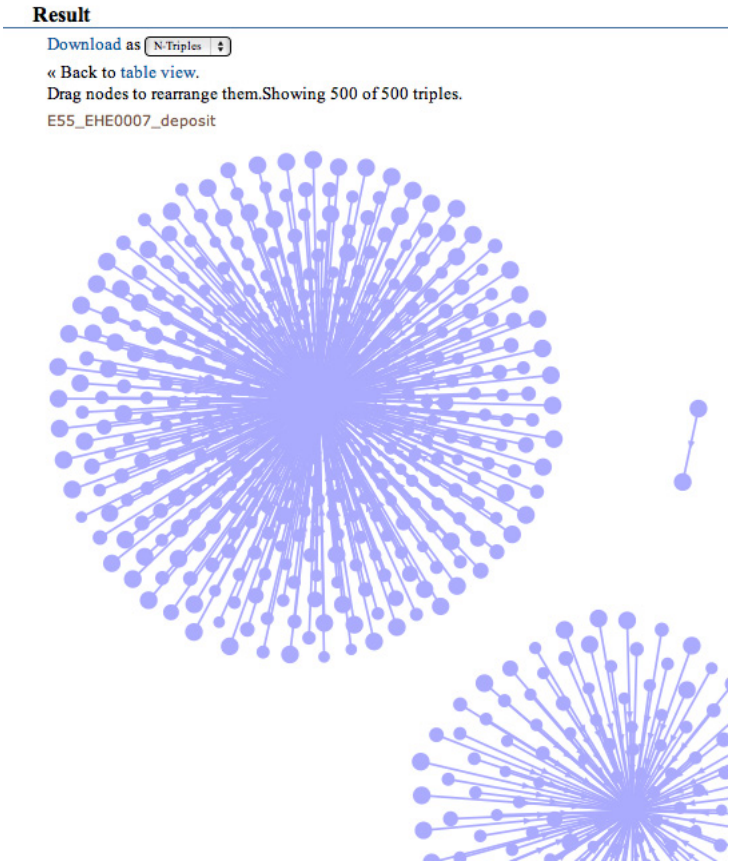


Figure 60: Screenshot of the 'Construct by Context Type' query, showing the first 500 triples in graph view. The larger cluster is E55_EHE0007_deposit, the smaller cluster is a subset of E55_EHE0007_cut, and the single triple is <EHE001_hungate> <ecrm:P2_has_type> <E55_EHE0001_excavation>.

As an example, the IADB holds the stratigraphic relationships between its context numbers within a separate table. The table consists of two columns, CON1 and CON2, with CON1 being stratigraphically above CON2 for the Hungate data. Within the data from the Cottam Context database, there is a table holding an 'earlier than' relationship between the contexts. These two sets of data in heterogonous formats were aligned to the CRM-EH using the CRMEH_CONTEXTS: strat_lower_id designation, and can now be easily queried together, building the foundation for more complex kinds of queries. At first glance, a query asking what context number is above another across two different excavations might not seem very useful, but because it can be combined with other queries that may not have been envisioned before, its use becomes apparent. For example, if a user were to ask a general query about when a context containing finds is

located directly below a context containing the base of a residential structure across multiple sites, hitherto unseen patterns might emerge about items lost through the floor of a house, or intentionally buried prior to construction. This is the sort of question archaeologists will find the Semantic Web can answer, but would be difficult for a relational database, if the query was not envisioned during its construction, and certainly not over multiple, heterogeneous databases at once.

In addition, because we explicitly know the 'what is below what' relationship between every context, we also implicitly know 'what is above what', and this could be designated using a SPARQL query. The designers of STELLAR have anticipated this however, and have built this functionality right in, so anything with the strat_lower_id relationship automatically has any inverse relationship defined. So the designation 'ecrm:P120_occurs_before', automatically has the inverse relationships 'ecrm:P120i_occurs_after', which looks like:

EHE1001_50413	ecrm:P120_occurs_before	EHE1001_50390
EHE1001_50390	ecrm:P120i_occurs_after	EHE1001_50413

So stratigraphic relationships are automatically built from the relationships existing within the triples. Within the thesis repository, the above would be:

EHE1001_50390
|
EHE1001_50413

Executing the saved query 'Construct by Stratigraphic Matrix' in the *Thesis* repository will list the following snippet in table view:

Subject	**Predicate**	**Object**
EHE1001_50557	ecrm:P120_occurs_before	EHE1001_50552
EHE1001_50573	ecrm:P120_occurs_before	EHE1001_50552
EHE1001_50565	ecrm:P120_occurs_before	EHE1001_50553

Because it is a *Construct* query, it can also be visualised as a graph. Visualising stratigraphic data as a graph is certainly possible using AGWebView, but more sophisticated tools like Gruff can do a better job of showing relationships that actually look like a stratigraphic matrix.

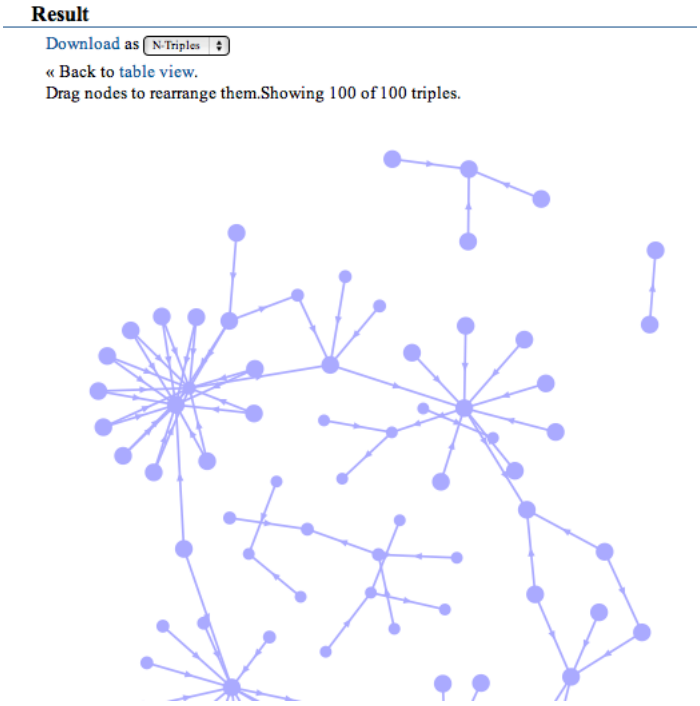


Figure 61: Screenshot of the 'Construct by Stratigraphic Matrix' query, showing the first 100 triples in graph view. The relationships are defined by mousing over the nodes and edges, and the direction of the arrow on each edge shows the direction of the relationship. While interesting, AGWebView is not sufficiently sophisticated to do a good job of communicating stratigraphic relationships.

AGWebView also allows geospatial data points to be visualised using a Google Map interface. To do this, AGWebView uses an RDF geospatial typed literal to reference the correct data type. This takes the form of a suffix added to the actual geospatial coordinates, and is a function of how RDF handles typed literals, rather than something specific to AllegroGraph or AGWebView (Powers 2003, 53).
The literal suffix for displaying the geospatial coordinates on the Google Map in AGWebView is:

<http://franz.com/ns/allegrograph/3.0/geospatial/spherical/degrees/-180.0/180.0/-90.0/90.0/4.0>

When a SPARQL query returns data with the correctly typed geospatial coordinates, the result includes the option to 'Display geospatial data in this result on a map', which then displays the points in their proper geolocation. As the Cottam and Hungate data come from two different URI sets, the query requires use of the UNION function to return both sites. The query is saved in the *Thesis* repository as 'View Sites on a Map' and looks like:

```
SELECT *
WHERE {

{ <http://www.diggingitall.co.uk/hungate/EHE0019_
hungate> rdf:value ?Geolocation . }

UNION

{ <http://data.archaeologydataservice.
ac.uk/10.5284/1000339/EHE0019_cottam> rdf:value
?Geolocation .}

}
```

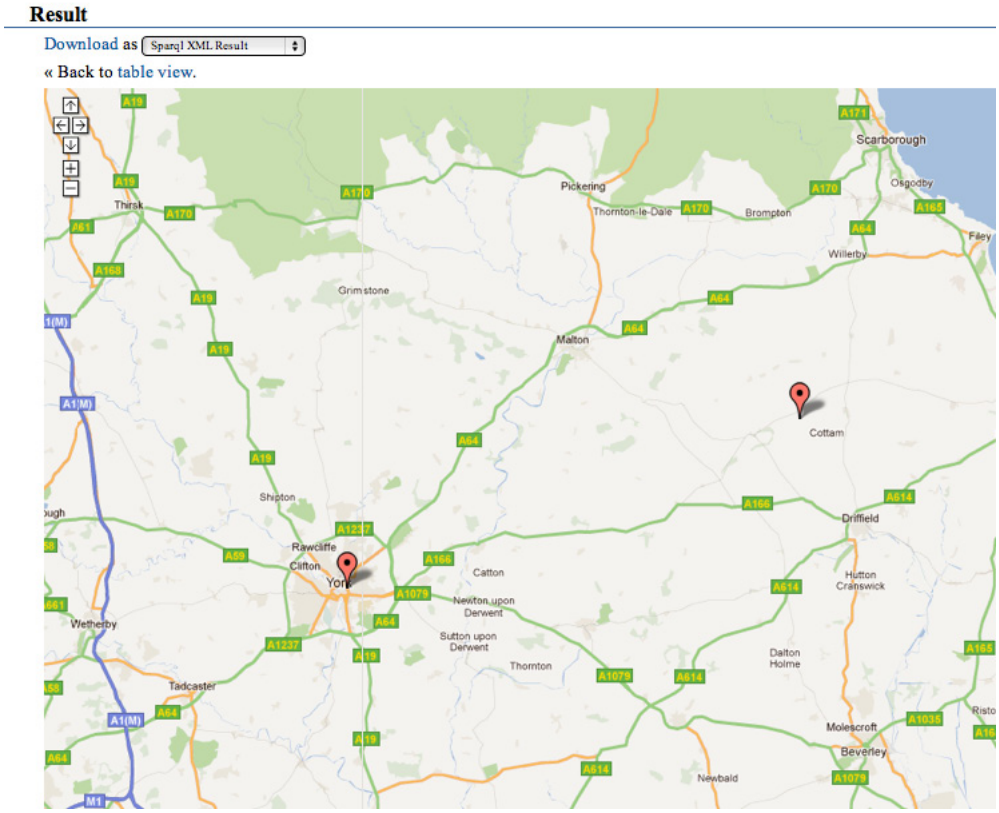


Figure 62: Screenshot of the 'View Sites on a Map' query, showing the locations of the Cottam and Hungate excavations. AGWebView has built in functionality to display any values with the properly formatted typed literals on Google Maps.

As the generic Web interface for AllegroGraph, AGWebView is naturally the easiest place to visualise data within the *Thesis* repository. Any valid *Construct* query will activate the option to view the data as a graph, up to a limit of 500 triples. AGWebView automatically displays the first 50 triples, with the option to add further triples in groups of 10 or 25. Each group of additional triples adds to the complexity of the graph and as it grows and changes, patterns begin to emerge. Hovering over a node will show its subject/object value, hovering over an edge will show its connecting predicate. While sufficient to show basic clustering and relationships within the data, the visualisation capabilities of AGWebView are quite basic, with no way of differentiating types of subjects, predicates and objects using colour or placement.

This section has just touched on what SPARQL queries can do, but most archaeologists will not want to learn SPARQL in order to use Semantic Web data. This means moving beyond SPARQL endpoints for specialist users into Web applications with Graphical User Interfaces which use SPARQL as a query language for internal system processes, with either a RDF store or a relational database (or both) as the backend (DuCharme 2011, 208). The advantage of graph data is its lack of fixed structure, but it can be challenging to understand the nature of the data and its relationships without a way to visually understand it. Graph data is meant to be *seen* to be understood, and this is done more robustly in applications like Gruff.

4.5.3 Visualising the data with Gruff

Also created by Franz.com as a way to work with data held within an AllegroGraph RDF store, Gruff is a desktop application, which can either access data in a repository by listening on a local port, or on a remote port via the Web. Testing remote access to the *Thesis* repository proved unsatisfactorily slow however, which was likely due to being run from a virtual Linux server, and wouldn't be a problem on a more robust system. Gruff allows Read/Write access, so it can be used to manage the RDF store as well. Gruff's primary purpose is

a visualisation tool however, and as such allows four different ways of viewing the data, including graph view, table view, query view and the recently added graphical query view. Graph view provides the familiar node and edge visual format seen when creating a *Construct* query in AGWebView, but with much more sophisticated graphics and functionality. Clicking on a resource in the graph takes the user directly to the relevant information in table view, which is also similar to that found in AGWebView, but with a more polished interface. Query view allows SPARQL and Prolog queries as in AGWebView as well, but again with a more intuitive and robust interface. AGWebView does not have a graphical query view however, and being a desktop application, Gruff does not have the ability to display geolocated data on a map, so the two interfaces are complimentary, rather than Gruff being a more full-featured version of AGWebView.

The heart of Gruff is graph view, where RDF triples can be loaded, explored, expanded, connected and the structure of the data understood. The interface is clean, well designed and intuitive, and allows for considerable customisation of the visual elements. Nodes are displayed as text boxes that are colour coded to correspond with their class type, and predicates are displayed as colour coded lines, described in a legend that sits on the left side of the screen. Creating a visualisation in graph view is begun by selecting a single node or group of nodes, and building the graph up to reflect the desired information. Essentially this is done by right-clicking directly on a node, and then choosing to display a linked node from either drop-down menus, or a tree. Displaying a linked node from a tree brings up a list of all the available predicates and nodes linked to the primary node, either as subjects or objects in the triple. Choosing one or more predicates and linked nodes then builds out the graph in the chosen direction and the visualisation expands and automatically redraws to accommodate the new information. In this way, it is easy to see what the connections are to any chosen link quickly, and start building meaningful connections. The legend is created automatically as the graph is built, and nodes can be moved individually to give the user full visual control over the result.

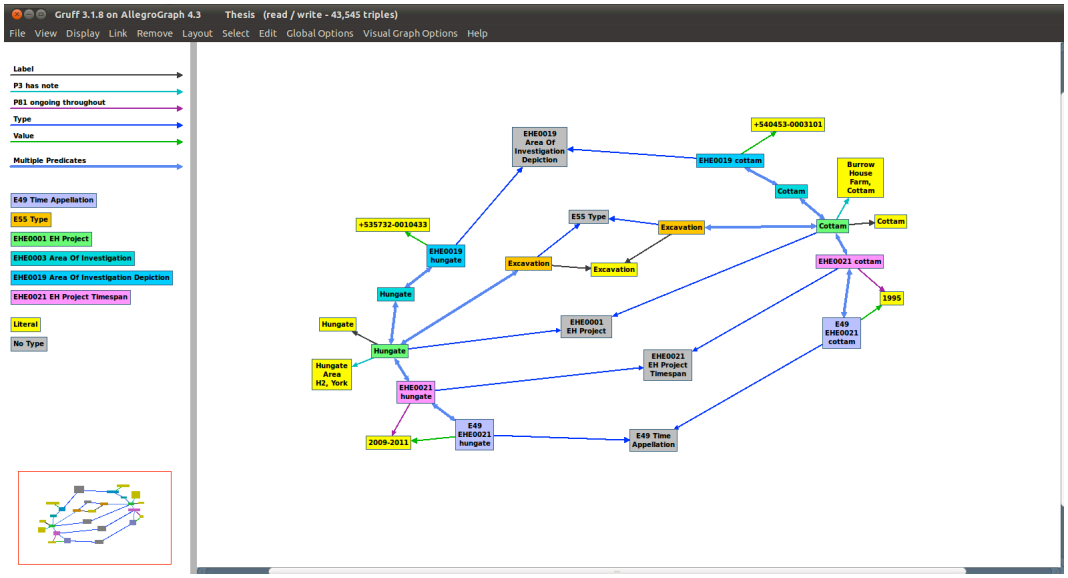


Figure 63: Screenshot of 'graph view' in Gruff, showing the relationships between the information about the Cottam and Hungate investigations, which corresponds to the CRMEH_ INVESTIGATION_PROJECTS template in STELLAR.

Of particular interest to archaeologists, graph view in Gruff allows control of the direction of predicate arrows in order to show hierarchy. This means context numbers can be built into stratigraphic relationships automatically. In the case of the CRM-EH domain ontology, this means specifying 'P120 occurs before' will always be placed in a downward direction, and the inversive 'P120i occurs after' will always be placed in an upward direction. While overall stratigraphic graphs of a single site are likely far too complex to visualise in Gruff, the process of building a graph in Gruff is extremely useful for understanding the stratigraphy of a site. The relationships between the contexts become apparent, even without clicking through the node into table view to see all the information about the context. Additional information can also be added to the visualisation in the places where it might be useful wherever it is present, such as the type of context, or the archaeological time period to which the context belongs.

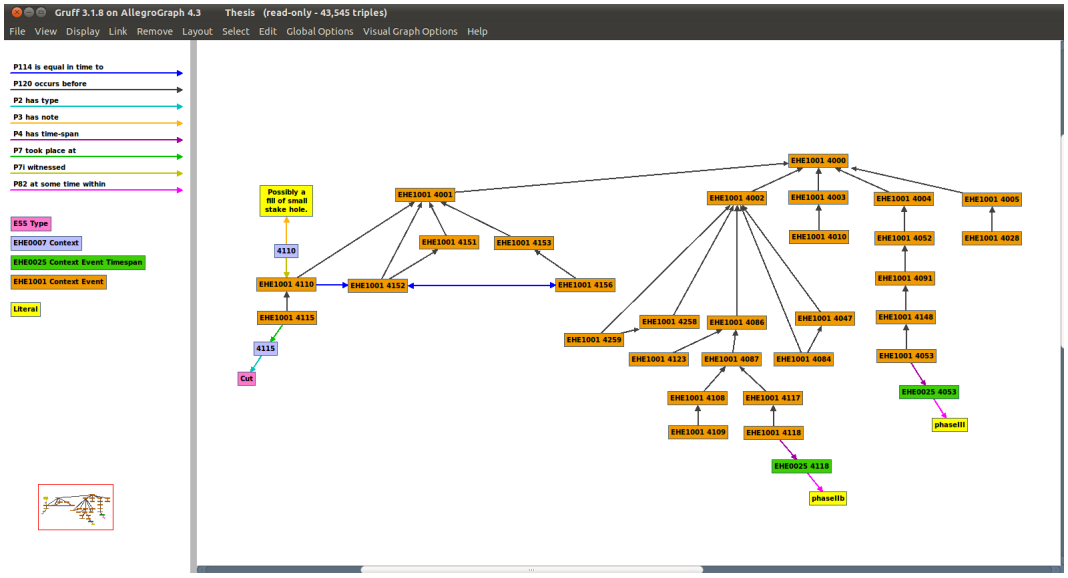


Figure 64: Screenshot of 'graph view' in Gruff, showing the stratigraphic relationships between the contexts within a subset of the data from Cottam. Setting the direction of the predicate arrows between context numbers to correspond with their stratigraphic relationships allows the building of matrices, which can be augmented with other useful information about the context, such as context type and phase.

Table view in Gruff displays all the properties and values (predicates and nodes) associated with every piece of data in the RDF store, so the user can click through and navigate the data as they would in AGWebView, but it can also be used to populate the graph view. Right-clicking on any of the properties listed linked to the selected node will add all the values associated with the property to the graph so they can be visualised as well.

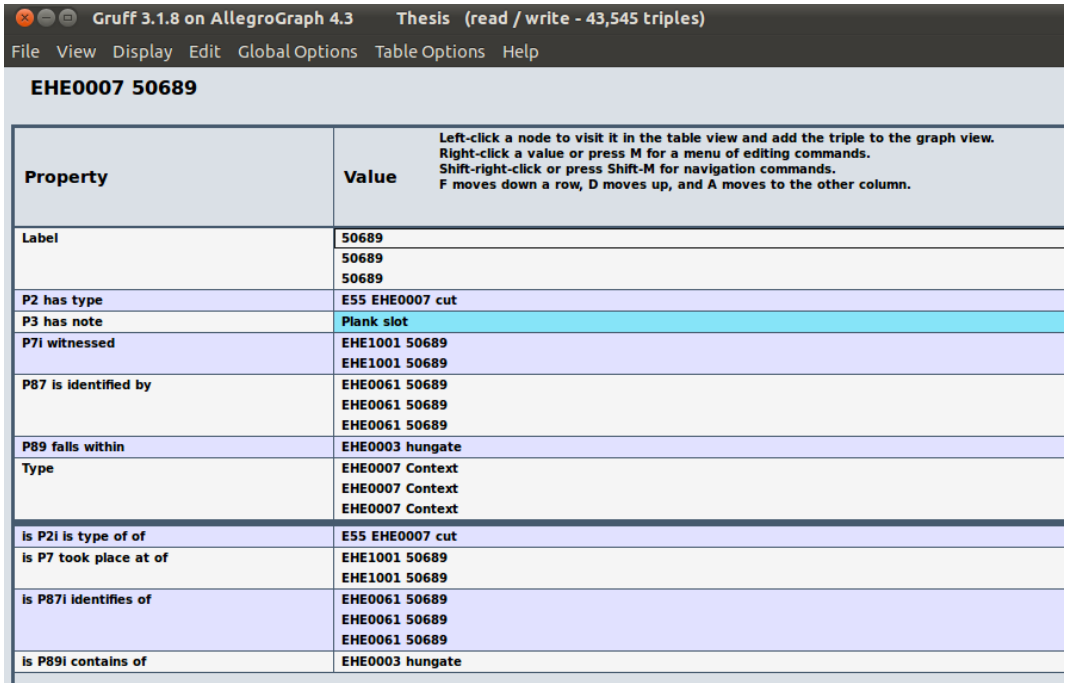


Figure 65: Screenshot of the properties and values associated with a single context from Hungate, shown in Gruff 'table view'.

Query view in Gruff is similar to query view in AGWebView, and most other SPARQL endpoints. It has a field to enter a query, which can then be saved and/or viewed in another format, shows the data returned from the query in table format and displays the explicit nodes and predicates from within the query. Interestingly, query view in Gruff does not support *Construct* queries, so the only export option for the resulting data is CSV format, rather than a true graph subset that could be exported in one of the RDF serialisations. This illustrates that Gruff is meant to be a visualisation tool more than a data management tool, and from a data migration standpoint, users are better off using AGWebView or working in AllegroGraph directly for full access to output formats.

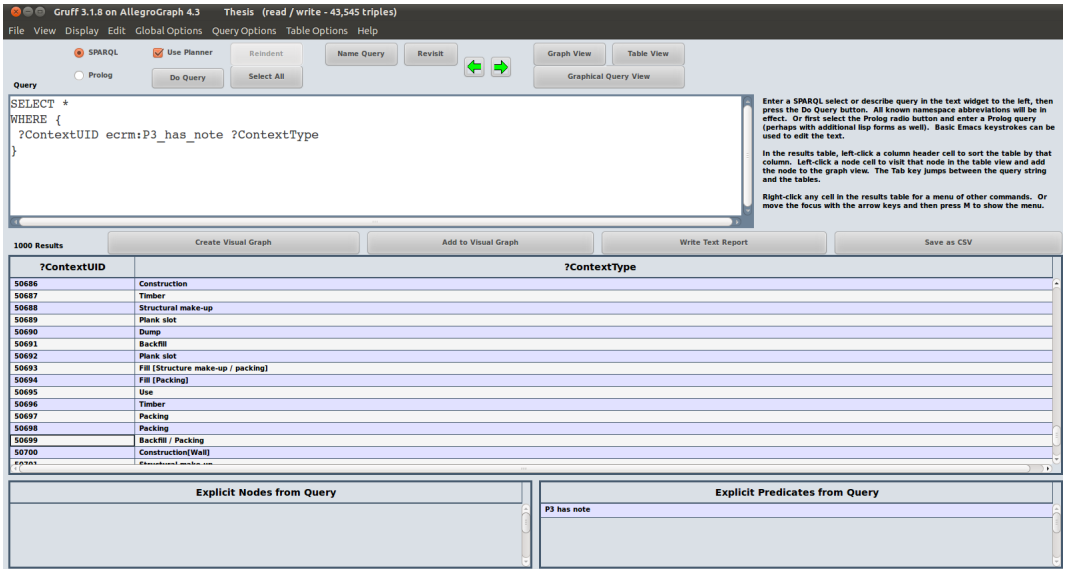


Figure 66: Screenshot of ‘query view’ in Gruff, showing a list of descriptive notes associated with the contexts in Hungate and Cottam.

Gruff also offers a graphical query view, which allows users to build queries and view their results without having to use SPARQL at all. A good understanding of the data to be queried is still necessary however, along with knowledge of how these types of queries are structured, so it would still pose a challenge for someone new to working with RDF. Just because it is possible to describe something in the form of an image, doesn’t mean it automatically translates into a valid query, and there is still a considerable learning curve to construct the nodes and edges and their properties in such a way that they represent the user’s question, and return a valid result. The real advantage of graphical query view is not that it gets around a need to understand how queries work in SPARQL, but it speeds up the process of writing queries, and makes it easier to write more complex queries. In addition to the basic query forms like *Select,* users can add an array of filters to return quite specific results. These can be challenging to write by hand in SPARQL, but can be added within graphical query view.

Queries in graphical query view are built up in much the same way data is viewed in graph view. Right-clicking on the layout screen brings up choices for a initial node to start off the query. The node can either be a variable or non-

variable, and non-variable nodes can be chosen in a variety of automated ways, including alphabetised menus of nodes within the loaded repository, or using freetext queries for matching nodes. Predicate or predicate variable links are then drawn between nodes to build the queries. These can also be quickly chosen from existing lists from the repository. Filters can then be added to the nodes themselves, and/or the predicate links can include filter functions as well. Node filters include filter matching for both types of data and values within the data, including subject or object type (ie only include EHE0061Context UID), whether a node is blank, whether it is a specific URI or whether it is a literal. It also has the same reverse functions to exclude data (ie only exclude EHE0061 Context UID). Node filters for data values include =, not=, <, >, <=, >=, along with the ability to enter text for using regular expressions (regex) or not-regex expressions.

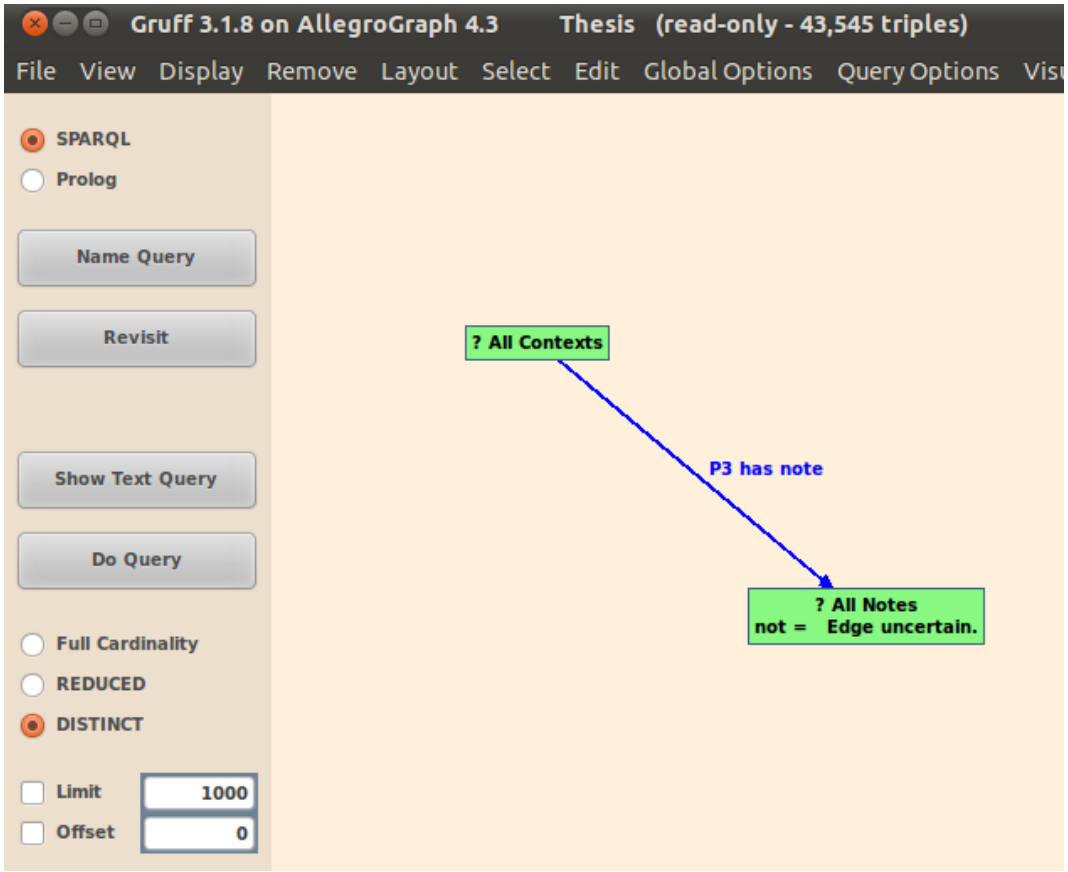


Figure 67: Screenshot of graphical query view in Gruff, showing a query which returns all the contexts and the notes describing the contexts, but excluding any which are designated as 'Edge uncertain' (not = edge uncertain). This allows data about which the archaeologist does not have confidence, to be easily excluded.

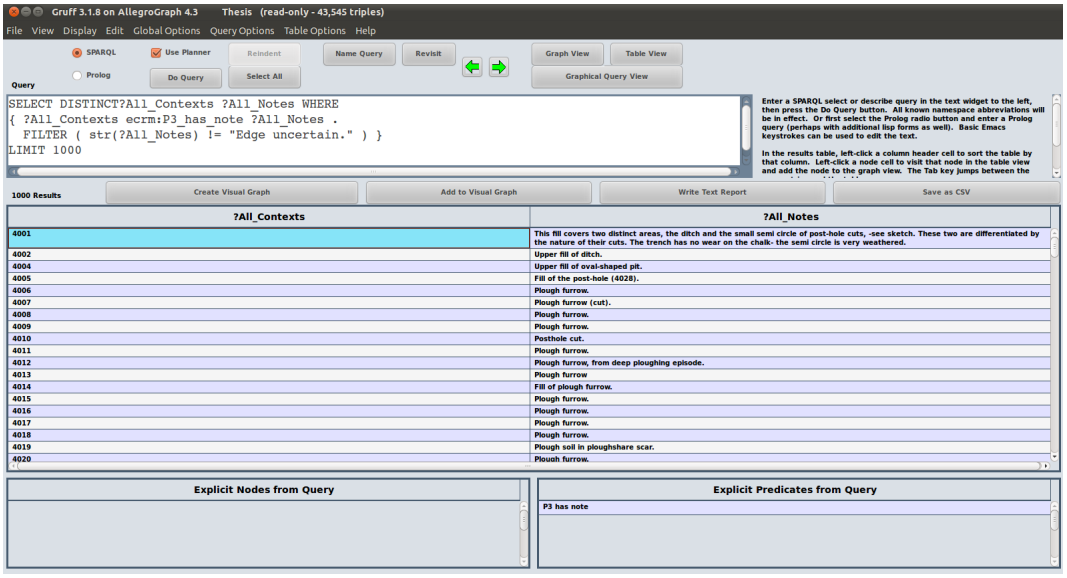


Figure 68: Screenshot of query view in Gruff, showing the tabular results and the SPARQL code created in graphical query view in the preceding image. Creating queries in graphical query view and then viewing the code generated in SPARQL can be very instructive.

Once the query is built in graphical query view, the resulting table and automatically generated SPARQL code can be viewed in standard query view and edited further if desired. Being able to view the SPARQL code from queries created in graphical query view can be very helpful when trying to learn SPARQL, though unfortunately changes made in standard query view are not reflected in graphical query view, which would also be helpful for learning. Gruff and other RDF store browsers are useful tools, and visualising graph data will continue to be the best way see, understand and use it. These browsers still require significant understanding of the nature and structure of Semantic Web data however, and more Graphical User Interfaces for non-specialist users need to be developed before most archaeologists will be prepared to make use of it.

4.5.4 Publishing the data with D2R and Pubby

As discussed in Chapter Two, from its introduction in 2006 much of the energy driving the Semantic Web has come from the Linked Data movement; the idea that in order to have a Web of Data, whatever data we hold should be made

(appropriately) available on the Web in a format that enables optimal use and re-use. Two ways of doing this were explored with the datasets from Cottam and Hungate, using applications developed at the Freie Universität Berlin by the Web-based Systems Group (who also brought us DBPedia). The intention behind both projects was to develop generic tools that could facilitate the publication of existing data as Linked Data, thereby broadening its availability as quickly as possible. The first is called D2R Server, which is meant to make legacy relational datasets publishable as Linked Data without having to convert them to RDF, and the second is called Pubby, which is meant to provide a Linked Data frontend to an already existing SPARQL endpoint.

D2R is part of the D2RQ platform, which also includes the D2RQ engine and the D2RQ Mapping Language. The D2RQ Engine is a plug-in for use with the Semantic Web *frameworks* Jena and Sesame, which creates a virtual, read-only RDF graph of the relational data. The data can then be accessed and manipulated from within the *framework* as though it were in RDF format. The D2RQ engine also functions as a way to dump the entire contents of a database into a single, static RDF file, if that is preferable to working with the data through a *framework.* In either case, a mapping file created with the D2RQ Mapping Language is the key to the translation. The D2RQ mapping generator creates a default mapping of the database schema, which can then be customized as needed to conform to a desired vocabulary. If data is added or changed, but the desired format in RDF remains the same, the mapping can be re-used, allowing for fast and efficient translation once the system is set up. Once the data is ready and a mapping file created, it can be passed to D2R Server for Web publication. D2R Server publishes the data for use in an HTML or RDF browser, and most current browsers also support access to the SPARQL endpoint. The SPARQL endpoint allows users to browse the data via the classes and properties within the RDF store, or create a subset of the data using a SPARQL query. The data can then be downloaded in JSON or RDF/XML format, if desired (Bizer 2010).

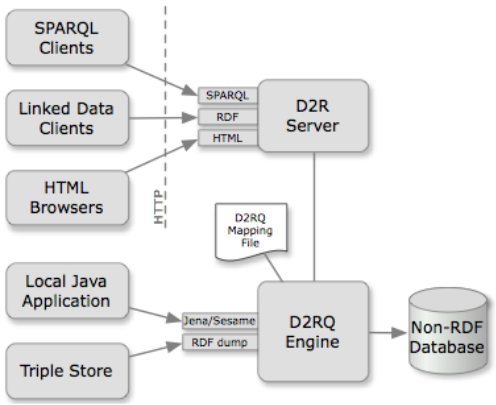


Figure 69: The design of the full D2RQ Platform architecture. Reproduced from the D2RQ website. http://www4.wiwiss.fu-berlin.de/bizer/d2rq/spec/#architecture.

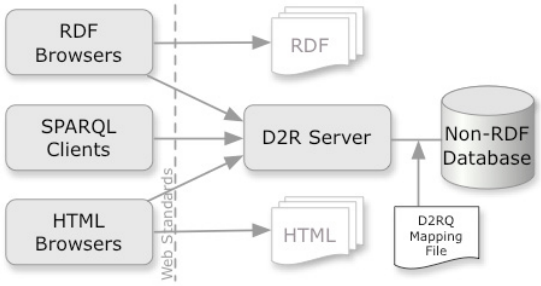


Figure 70: The design of the D2R Server architecture. Reproduced from the D2R website. http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/.

It is not necessary to use the entire D2RQ platform in order to publish Linked Data with D2R Server. D2R Server can be directly connected to a compatible database (such as Oracle, MySQL, PostgreSQL, etc.), along with a D2RQ mapping file in place to translate between the two (Bizer and Cyganiak 2010). To explore this, a PostgreSQL database was created, populated with data from Hungate in CSV format. Only Hungate was chosen, as the Cottam data has already been published as part of the ADS STELLAR Linked Data demonstrator with the appropriate base URI (http://data.archaeologydataservice.ac.uk/10.5284/1000339/). The D2RQ mapping file was generated from the CSV files prepared for the STELLAR translation, which resulted in sets of attributes corresponding with the column headings necessary for mapping to STELLAR. These attributes were then customized to map to the ECRM and CRM-EH as appropriate, which were added to the base URI of http://www.diggingitall.co.uk/

data/hungate. The result is data that should be fully compatible with other datasets using the STELLAR mapping of the CRM-EH. This demonstration shows it should be possible for archaeologists who are not interested in RDF *per se*, to still make their relational data available, mapped to an appropriate, publicly available ontology for interoperability, with minimal effort. Specialist knowledge may be necessary for the initial setup, but once the database connection is made, the D2RQ mapping file created and customized, and the D2R Server established it should be relatively easy to add more databases as time goes on.

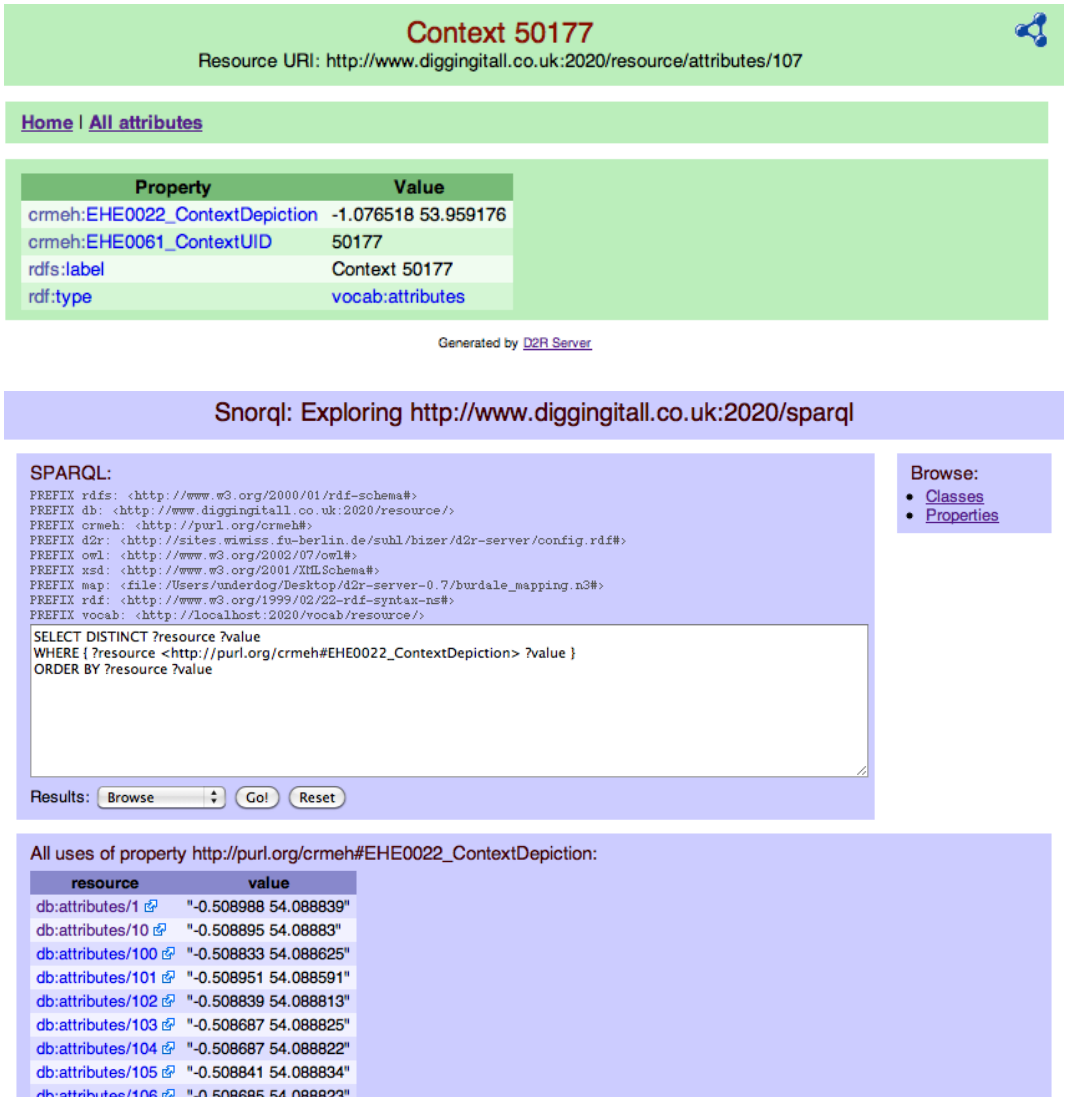


Figure 71: Screenshots of the D2R Server interface. Upper image shows the context number and geospatial location of a context within the Hungate dataset, mapped to the CRM-EH. Lower image is the D2R Server SPARQL endpoint interface, showing a list of the geospatial coordinates for all the contexts in the database. The data can be queried, browsed or downloaded from here.

Pubby provides a Linked Data frontend to RDF data for client applications using the SPARQL protocol, and turns a SPARQL endpoint into a Linked Data server (Cyganiak and Bizer 2010). Pubby takes a different approach from D2R, in that it only shows data in table view, and relies on external RDF browsers like OpenLink Data Explorer (created by OpenLink Software), Marbles (created by the Web-based Systems Group at the Freie Universität Berlin) or Information Workbench (created by Fluidops) to work with the data. RDF browers have varying levels of functionality. Marbles is the most basic, with similar functionality to Pubby in that it shows the data in table view, Data Explorer has the type of functionality associated with a SPARQL endpoint like AGWebView, with the ability to browse, filter, query and visualise the data as a graph. Information Workbench has the most viewing options, with a wiki view and a pivot view (which shows all the results, including images, as a clickable pivot table), along with the traditional graph and table view. Pubby also allows data to be downloaded in Turtle and RDF/XML serialisations, so it is up to the user whether they wish to browse the Pubby dataset live via URIs, or with a file downloaded from Pubby.

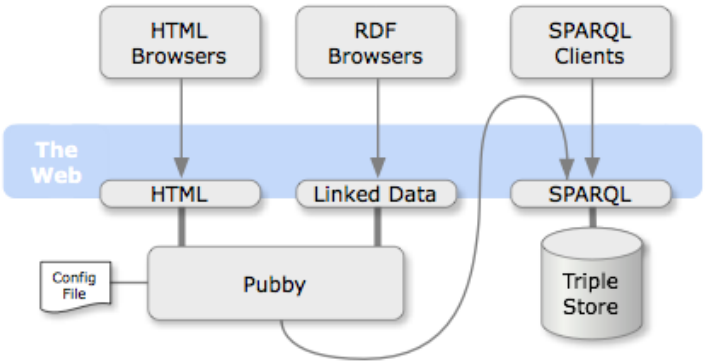


Figure 72: The design of the Pubby Server architecture. Reproduced from the Pubby website. http://www4.wiwiss.fu-berlin.de/pubby/.

The *Thesis* repository already has a SPARQL endpoint within AGWebView, and with the addition of anonymous user access, it was possible to add access through Pubby to the data within the AllegroGraph RDF store. Once again, Cottam having been published directly from the ADS (using a Pubby server as well), Pubby

was pointed to a separate repository containing the Hungate data at http://www.diggingitall.co.uk:8080/data/. Pubby is a Web application designed to run within a servlet, and was set up using Apache Tomcat 6. The link between the SPARQL endpoint and Pubby is controlled with a simple configuration file in Turtle syntax. Because Pubby is looking directly at the RDF store in AllegroGraph, the data is already mapped to the CRM-EH with fully resolvable URIs and requires no further processing to be fully interoperable with other data processed using STELLAR. By publishing the data using tools like D2R Server and Pubby, as well as making the data available to selected users via a SPARQL endpoint, the *Thesis* data has now been brought through a complete workflow. This workflow has been outlined in Appendix B.

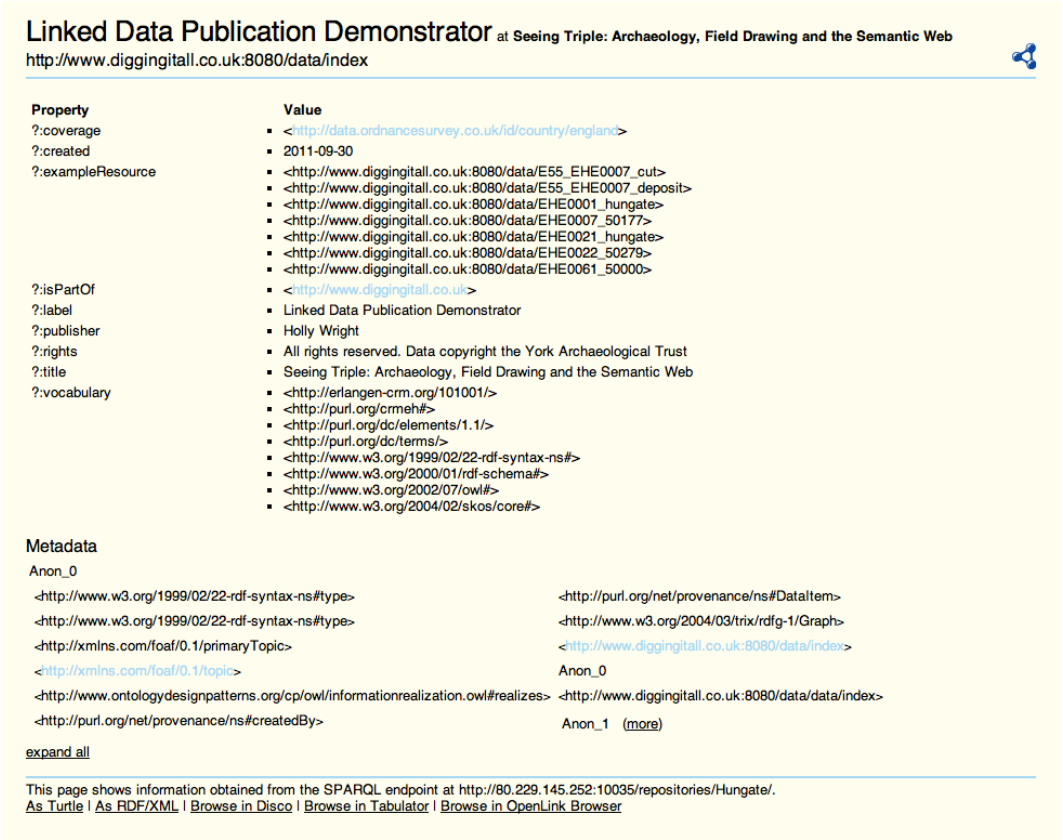


Figure 73: Screenshot of the start page of the Linked Data publication demonstrator created with Pubby, and loaded with the data from Hungate. Pubby is software which sits on top of an existing SPARQL endpoint, allowing publication of data from within an RDF store as Linked Data. The data can then be browsed in an external RDF browser, or downloaded in Turtle or RDF/XML serialisation format.

4.6 Spatial approaches

As a practical implementation, this chapter has focussed primarily on the first two research questions set out in the introduction to this thesis, but the third question requires further consideration, as it relies on still emerging ideas and technologies:

> Archaeological field drawing is a fundamental part of field recording. How can the point, line and polygon data comprising the visual archaeological record be included alongside the textual record with regard to the Semantic Web?

The chapter has demonstrated that geolocated point data can be accommodated, but what about more complex archaeological information made up of lines and polygons? As there are no direct answers currently available, how are researchers outside of archaeology attempting to do this? One of the primary venues where this question began to be formally explored was at the Bentley user conference, held in London in 2007. Bentley is a software development company which specialises in CAD and design management products for large infrastructural projects. The conference research seminar was called *Creating Spatial Information Infrastructures: Towards a Spatial Semantic Web*, and pulled together ideas which had begun to appear in Bentley seminars as far back as 2004. The 2007 seminar was the first to pull together the more general idea that the Semantic Web must have an inherently spatial component, with differing requirements from textual data. It resulted in a 2008 publication of the same name, which included both the seminar papers and further invited papers in order to give the subject wide coverage (van Oosterom and Zlatanova 2008, vii). The papers spanned work in Europe, the US and India, and as might be expected, all were focussed on large-scale national and international projects involved in work at the infrastrucural level, reflecting high levels of expertise and specialisation. They spanned contributions from transportation projects, earth science, mapping agencies, environmental change projects and land administration, along with explorations of best practices techniques for the general infrastructure sector.

Despite the large scale of these projects, from a practical standpoint, the issues they were trying to address are the same as what archaeologists will need to resolve in order to fully represent the drawn record created during field recording using the Semantic Web. They defined the specific challenge thus:

> Many spatial objects, such as areal coverages and linear events, require special treatment both at the (continuous) object level as well as the location-dependent property value level (e.g. elevation map), adding to the ontological complexity. And for spatial data sets, relationships are more frequently computed rather than stored. This presents problems for reasoners, which assume that all data relationships are explicit (Scarponcini *et al.* 2008, xvii).

The papers which followed proposed various ideas about how to implement the 'special treatment' to which they refer, but as the technology was still so immature, there were no definitive answers (Dolbear and Hart 2008, 100). A relevant example for archaeologists is the INSPIRE project. INSPIRE stands for Infrastructure for Spatial Information in Europe, and is an ongoing EU initiative to make the complex and distributed geographical information about human and environmental interaction, gathered from many EU countries and agencies, more integrated and meaningful. The goal of the project is to create a resource to allow EU policymakers better and faster access to spatial information, in order to make more informed decisions about environmental change. The particular problems they saw with their existing data structures were data inconsistency, redundancy, lack of documentation, incompatibility, proprietoriality, and resistance to data sharing (Annoni *et al.* 2008, 2). There were many challenges encountered in trying to address these issues, but the most technically difficult were identified as being problems with inconsistent data, and bridging the many EU languages in use. Within this there were naming conflicts, due to different names used in different languages (similar to archaeological naming conflicts, with terms for

the same place changing temporally), scale, precision and resolution conflicts, conflicts between the constraints governing how data is captured, and the actual values found within the data (Annoni *et al.* 2008, 5-6). These problems will feel familiar to those working with archaeological data.

The development of INSPIRE is ongoing, and the current workpackages are due to be completed in 2019 (INSPIRE 2012). In the most recent publicly available project status report, INSPIRE announced plans to make public draft regulation for the interoperability of spatial data sets in 22 languages, including coordinate reference systems, geographical grid systems, geographical names, administrative units, addresses, cadastral parcels, transport networks, hydrography and protected sites. They also plan to make their metadata validator and geoportal software available as open source through the Open Source Observatory and Repository (INSPIRE 2010). This might be directly relevant to archaeologists, especially those working in the EU. By coordinating efforts with the ongoing work of INSPIRE it might be possible to not only make use of the technology they are developing, but perhaps have a level of data interoperability as well. Participants in the INSPIRE project are also investigating making their data available as Linked Data, which could provide further accessibility for archaeologists (Schade and Lutz 2010).

Another project included in the 2007 survey, which might be of particular interest to UK archaeologists, was from the national mapping agency; the Ordnance Survey (OS). The OS had already formed a Geosemantics research group to explore the potential of Semantic Web technology, including authoring of ontologies, information integration, and the representation and manipulation of data using RDF. The OS saw the Semantic Web as a means to help bridge the gap between the rich representation understood by those who gather their data, and the limitations inherent in the way that data is recorded. The example they give is how the OS 'requires surveyors to capture "real-world objects" such as houses, warehouses, factories and so on. However, within the data, these different objects

are simply identified as “buildings” (although a textual description may be also associated with the object)’ so use of the Semantic Web was a potential way of recovering some of this ‘information loss’ (Dolbear and Hart 2008, 92). One of the key issues identified by the OS, was the lack of research within the semantics community into incorporating spatial reasoning with semantic reasoning. A further issue was the fact that much geospatial data does not contain explicit semantics associated with topological information anyway, as it is often generated as needed based on a geometric query (Dolbear and Hart 2008, 96). In 2007, the OS was just beginning to explore how to deal with these issues, and their paper outlines some of the possibilities (including the use of D2R), but no clear solutions were available at that time.

Despite these limitations, the OS has continued to expand its Semantic Web-based research and services, especially in the realm of Linked Data, though still with an understanding of the limitations with regard to spatial data:

> ...to provide a spatial referencing system as a component of the Linked Data Web may appear somewhat optimistic. But, the Linked Data Web can nonetheless work well with non-coordinate based data, and it can also at least store the geometry related to spatial objects even if it cannot index or query it directly, enabling the data to be used by GIS for analysis and display purposes... The Linked Data Web can also express topology, mereology and other discrete relationships, such as establishing that two data refer to the same real world object. So whilst the Linked Data Web is not ideal for all GI [Geographic Information], it can handle many kinds of GI (Hart 2009).

The OS have continued to do pioneering work with Linked Data in particular, both with technology and public access. The current initiative, called OS OpenData, includes a public forum for users called OS Open, and a wiki with

guidance and tutorials for using the data. There are also examples of the data in use, and the opportunity for developers to showcase work using the data. The conversion of OS datasets to Linked Data format is ongoing, so more possibilities will become available over time (Ordnance Survey 2012).

In 2011, the second book specifically highlighting the Semantic Web and spatial data was published. Entitled *Geospatial Semantics and the Semantic Web: Foundations, Algorithms, and Applications*, this text includes papers from an entirely different group of researchers from the Europe and the US. Much like emergence of SII, the rationale for bringing together this group of papers was the creation of a new research area, this time termed Geospatial Semantics (GS). Unlike SII, GS is born out of the proliferation of 'everyday applications ranging from personal digital assistants, to Web search applications and local aware mobile services, to specialized systems in critical applications such as emergency response, medical triaging, and intelligence analysis to name a few' (Ashish and Sheth 2011, v), so rather than the large infrastructural projects which drove earlier research, three years later, the focus is now on the small-scale, the portable and the personal.

In this group of papers, two stand out as being of particular interest to archaeologists. The first examines an area within the FinnONTO project, which explores ways of dealing with the changes of place names over time, and attempts to create a time series of spatial ontologies that can then be used to index and map spatio-temporal regions and their corresponding names at different times (Hyvönen *et al.* 2011, 1). This research moves beyond the changing names associated with geospatial point data, and wrestles with changes in regional boundaries and definitions over time. The specific goals of the project include using regional data to create more accurate annotations, true geospatial querying of data using spatio-temporal relationships, using existing relationships to discover additional data, using semantic reasoning to infer new ontological relationships, and create new visualisation tools for users to interact with the data.

The use of the concept of 'regions' is key to this research, and can be defined as a name, a time span, a size or a polygonal area and can be political, religious, geological or historical, etc. The ontology model allows geospatial queries of regional data like 'overlaps' or 'is overlapped by'. These overlaps can be spatial; reflecting things inhabiting the same area, or temporal; showing coexistence in time. The problems of incomplete polygon data, or data where regional boundaries are uncertain or in dispute, are also explored. Within this, the project attempts to at least create partial data models using ontology creation and inference to enrich the result. The ontology uses three core properties to describe a region at a particular period of time, including a name, a geolocated polygon and an unbroken time interval corresponding to that name and polygon. Data associated with a region with a particular set of core properties can then be linked to that region, thereby giving it an associated spatio-temporal context (Hyvönen *et al.* 2011, 4-5).

Conversely, when qualitative data is available but has no associated geospatial or temporal information with which to link it, the project explored ways of augmenting the data using semantic reasoning. To do this, they created a baseline repository of spatial and temporal data for the general region they wanted to analyse. This included information about any sub-regions present, including name, type, polygonal area, size, temporal changes and topological relationships. It also included when sub-regions formed, merged or disappeared. This allowed much more of the unassociated data to be automatically linked to the core properties within the ontology, enriching the entire dataset (Hyvönen *et al.* 2011, 7-12).
The ontology created for this project has now been published for re-use, and applications have been developed to use the data. The primary application is the Web portal *CultureSampo–Finnish Culture on the Semantic Web 2.0*, which allows search terms to be displayed on a map, within historical areas, on historical maps and showing any nearby objects. These searches are carried out using geolocated point data and polygon data, and the various maps can be overlain to show change over time (Hyvönen *et al.* 2011, 14-16). This work certainly has the potential to be of direct use to archaeologists, and should be explored further.

The second paper describes an attempt to expand the SPARQL query language to accommodate complex spatial and temporal data, called SPARQL-ST. The research proposes the formal syntax and semantics for the query language, and demonstrates an implementation built on a relational database and evaluates the performance of the demonstration. The research uses a standards-based approach and bases the ontological modelling on GML for the spatial features, and uses RDF reification (see section 2.5.2) for the temporal features (Perry *et al.* 2011, 62-64). The article is highly technical, and is a description of the actual syntax and its implementation, but its importance lies in its approach. Rather than creating a bespoke ontology to work around the spatio-temporal limitations of RDF and SPARQL, this work sets out to create a standards-based extension to give it the missing functionality that would be universally applicable across all disciplines, including archaeology. This work has been taken up by the Open Geoapatial Consortium and is now ongoing. As such the current development of SPARQL-ST concepts will be discussed further in the next section.

While by no means comprehensive, these examples give a sense of how researchers in other disciplines have begun to approach the 'special treatment' necessary to bring the missing spatial dimension to the Semantic Web. At the same time, the temporality of spatial data within these examples comes through as both a further dimension in need of complex representation, and as a potential avenue towards a solution. This is all to the good for archaeology, and shows that temporal solutions, the next important hurdle won't be far behind.

4.6.1 Exploring the geospatial potential of the practical application

This chapter's practical exploration illustrated how data derived from field drawings and related field data can be extracted, aligned to an appropriate ontology, assigned URIs, converted to RDF, housed within an RDF store, queried, visualised and published, but more could be done with this data. As extracted from GIS and converted to CSV, the data from Cottam and Hungate contains more geospatial information than could be carried forward into STELLAR. The data

was generated in GIS as polygons, which allowed calculations for the perimeter, area, and centroid of each context, along with the georeferenced points making up each polygon. Being a recently developed prototype, STELLAR only supports one geospatial data element for contexts, called 'Context_location', which is defined as an element which '…could be a number of spatial referencing systems. For STELLAR Linked Data purposes we have opted simply for a single X, Y, Z point based on WGS84 coordinates, following MIDAS quickpoint syntax' (Binding 2011), and was therefore used to describe the centroid for each context. MIDAS stands for Monument Inventory Data Standard, and is the data standard used in the UK for monument inventories (Bell *et al.* 2005). As the Cottam and Hungate data originates in the UK, and was projected alongside maps from the Ordnance Survey, the OSGB36 Terrestrial Referencing System (TRF) was used initially.

OSGB36 is the traditional UK National Grid (Ordnance Survey 2010, 23), but as Linked Data is meant to be as interoperable as possible in its raw state, the STELLAR team opted for the World Geodetic System 1984 (WGS84) as the preferred coordinate system, therefore the Cottam and Hungate data was transformed into WGS84. Because data at the site level requires fairly high precision, consumers of this data may opt to convert it back to OSGB36 (or the relevant coordinate system appropriate for the country from which the data originates), but it would depend on the research requirements. WGS84 is the basis for most modern GPS coordinate systems, and given the general prevalence of GPS technology, allows the greatest interoperability (Ordnance Survey 2010, 18-21). It is also the coordinate system used by most geospatial Web applications like Google Maps and OpenStreetMap, so aligning STELLAR data with WGS84 makes it immediately available for use by a wide variety of popular applications. Google Maps uses the Mercator projection (Google Maps 2011), so this was the projection assigned to the Cottam and Hungate data. GeoNames, the ontology and geographical database used to provide descriptive information for many Semantic Web applications also uses WGS84, so data from STELLAR could easily be made interoperable with GeoNames as well (GeoNames 2011b).

MIDAS quickpoint syntax refers to the current MIDAS XML 2.0 schema, which is a W3C standards-based format for describing information in the historic environment domain, and forms the core of the Forum on Information in Heritage Standards (FISH) Interoperability Toolkit (FISH 2011). Making this part of STELLAR conform to MIDAS adds further potential for interoperability. The schema includes the quickpoint syntax, along with the Well Known Text (WKT) markup language spatial type for a single point; WKT being the Open Geospatial Consortium's (OGC) text-based markup language for describing simple vector geometry on a map (Open Geospatial Consortium 2010, 61). MIDAS combines both the quickpoint and WKT designations for a point, and describes it thus:

> The element is designed to provide a convenient [way] to record the centroid of the spatial appellation, or a coarse approximation of its location. While use of the entity/wkt element is encouraged, it is recognised that not all systems will be able to easily parse wkt notation. This element therefore acts as a 'shortcut', designed to hold a single coordinate pair representing all spatial entities within an appellation (MIDAS 2011).

Using an example from Cottam, the XML might look like this:

```
<spatialappellation type="contextlocation">
  <quickpoint>
    <!—quick-and-dirty X and Y grid reference>
    <srs>WGS84</srs>
    <x>-0.510952</x>
    <y>54.086988</y>
  </quickpoint>
  <entity spatialtype="Point" uri="4004" namespace="Cottam">
    <wkt srs="WGS84">POINT(54.086988 -0.510952)</wkt>
  </entity>
</spatialappellation>
```

Expansion of the STELLAR templates to include two-dimensional polygons representing the extent of a single context in a field drawing, might be called something like 'Context_extent', and presumably would have a similar structure.

A polygon in MIDAS has no equivalent ‘quickpoint’ syntax, and requires regular WKT notation. An example might look like this:

```
<spatialappellation type=”contextextent”>
  <entity spatialtype=”Polygon” uri=”4004” namespace=”Cottam”>
    <wkt srs=”WGS84”>
    POLYGON((54.088836 -0.508838, 54.088625 -0.508833,
    54.088834 -0.508841, 54.088831 -0.508834))</wkt>
  </entity>
</spatialappellation>
```

Further geospatial data elements could be defined within STELLAR and aligned to MIDAS, not only area and perimeter, but information for place, address, named place, location, grid reference, geopolitical location, type of geometry, a bounding box (which could represent the extent of the entire site as a polygon), a spatial appellation to describe both a centre point for the site and its extent, and all the relevant metadata for the geospatial information (MIDAS 2011). Some textual information could also be aligned to the more universal GeoNames if deemed appropriate. If further geospatial functionality were added to STELLAR, and the resulting data included within an RDF store, the next issue becomes a matter of how to use it. Attempts to incorporate geospatial indexing into the functionality of commercial RDF stores have been made by several providers, the first being in 2007 as part of the Parliament RDF store created by Raytheon BBN Technologies, and each takes a different approach (Battle and Kolas In Press). Freely available examples include software like OWLIM-SE by OnToText, and AllegroGraph, and there is also a costlier spatial extension to the Oracle 11g database which has geospatial Semantic Web functionality.

OWLIM-SE includes geospatial indexing and several geospatial extensions based on the W3C Basic Geo Vocabulary, and allows querying to determine whether points are inside or outside a circle or polygon, which points are inside a circle or polygon, and to compute the distance between two points (OnToText 2011). AllegroGraph has created its own set of proprietary operators which can be used within SPARQL queries to return geospatial information, including the radius

around a specified point (for Cartesian coordinates), a bounding box calculated from two points, and a haversine radius around a specified point (for spherical and longitude/latitude coordinates). AllegroGraph was also meant to have support for querying polygons, which is one of the reasons it was chosen for this research, but its completion had not materialised at the time of this writing (Franz Inc. 2011a).

The Oracle 11g Enterprise Edition database has comprehensive RDF store functionality, and the addition of the spatial extension allows a broad range of geospatial querying. It supports geometries that are WKT literals, and OGC Simple Features geometry types. It defaults to WGS84, though virtually any coordinate system can be used, and transformations between coordinate systems are done transparently as part of a query. It allows querying of geospatial data including topological relations, distance, within distance, buffer, nearest neighbour, area, length, centroid, intersection, union and difference. This represents the type of functions archaeologists who are used to working with GIS programs would require. Oracle Spatial would seem a good solution for any archaeologist wishing to use geospatial data, but most would find the cost prohibitive, and the level of specialist knowledge required is similar to that of using a *framework*, putting it out of reach, both financially and technologically, for most archaeologists (Beauregard *et al.* 2011, 71-9).

Bespoke geospatial functionality housed within commercial RDF stores is not an optimal long-term solution however, and efforts have been underway to create standards-based means of working with geospatial data which separates it from any proprietary format, and real momentum has been growing within the last two years. In June of 2010, the OGC convened a *GeoSemantics Summit* in Silver Springs, Maryland. The summit was introduced by Josh Lieberman, who edited the report on the OGC's first foray into work on geospatial standards specifically for use with the Semantic Web (Lieberman *et al.* 2006), which resulted in the formation of the OGC Geosemantics Domain Working Group the same year. The stated goals of the summit were to 'examine semantic mediation, linked geodata,

and other trends in the application of geosemantics, and to initiate development of a reference model for the use and adaptation of OGC standards to further enable critical geosemantic applications' (Lieberman 2010). The issues they wished to address were:

- **Federation of disparate geospatial services and data across domain and community boundaries:**
    - Increasing need to perform semantic mediation between the concepts and vocabularies brought into these federations in order for discovery and exploitation of geodata resources to be successful.
    - Need for a consistent geosemantic framework for successful mediation to occur.
- **Publication of linked geodata:**
    - "Putting it out there and see what happens" has enormous potential for geospatial enablement and fusion.
    - Also has enormous potential for confusion and duplication as link scheme development and "triplification" are carried out in many different ways.

The participants began to define a core group working the area of 'geosemantics', and in September of 2010 the *Workshop on Linked Spatiotemporal Data* was held in Zurich, in conjunction with the 6th International Conference on Geographic Information Science, and attracted an overlapping group of participants (LSTD 2010). One of the outcomes of this workshop was the decision by the Semantic Web Journal to create a special issue on 'Linked Spatiotemporal Data and Geo-Ontologies', due to be published in the Autumn/Winter of 2011 (Janowicz 2011), which will no doubt give a comprehensive overview of the state of work thusfar. In the meantime however, one initiative presented at both meetings has been taken on by the OGC for development as a standard, and is called GeoSPARQL. The development of GeoSPARQL is a specific attempt to address the issues set out

in the *GeoSemantics Summit* by aligning and unifying the various vocabularies, query languages and experiments in enabling spatial reasoning put forth over the last 10 years into a single standard (Battle and Kolas In Press).

GeoSPARQL is a geospatial extension to the SPARQL query language, and its purpose is to define a vocabulary for representing geospatial data in RDF, as well as providing an extension to SPARQL for querying that data. The OGC GeoSPARQL working group is headed by two members of the Oracle Semantic Technologies Center. In fact, fully half of the working group works for Oracle, the rest representing American and European private companies, along with representatives from the Ordnance Survey and US Geological Survey. The convener of the working group is Oracle, and the voting member is a researcher named Matthew Perry (Perry and Herring 2011). Perry has been at the heart of the development of GeoSPARQL, and his prototypical work in this area can be traced to his 2008 PhD entitled 'A Framework to Support Spatial, Temporal and Thematic Analytics over Semantic Web Data' where he proposed a SPARQL extension for both spatial and temporal data called 'SPARQL-ST'. Perry interned with Oracle while completing his PhD (Perry 2008), and was subsequently hired shortly after its completion. Presumably his work influenced Oracle's involvement in the creation of the standard. SPARQL-ST was the first attempt to create a geospatial extension to SPARQL, and the first to allow indexing in coordinate systems other than WGS84, but the original query syntax also deviated from the SPARQL standard, and was therefore not considered a viable basis for GeoSPARQL by the OGC (Battle and Kolas In Press).

As of July 2011 a draft of GeoSPARQL was submitted for public comment by the OGC and the GeoSPARQL Working Group has stated they hope to have it ready for an OGC standardization vote by the end of year (Open Geospatial Consortium 2011a). In its current state, it is described as a standard to both represent geospatial data in RDF, and define a geospatial extension to the SPARQL query language. It will have a modular design consisting of five components, including a

'Core' component that will define the relevant RDFS/OWL classes, a 'Geometry' component for serialising the data, a 'Geometry Topology' component for topological queries, a 'Topological Vocabulary' component for defining geospatial properties (predicates), and something called a 'Query Rewrite' component, which defines transformation rules for queries between spatial objects. Each of the components forms a separate requirements class, so users can pick and choose which ones will be supported in their application (Open Geospatial Consortium 2011b, xiv-2).

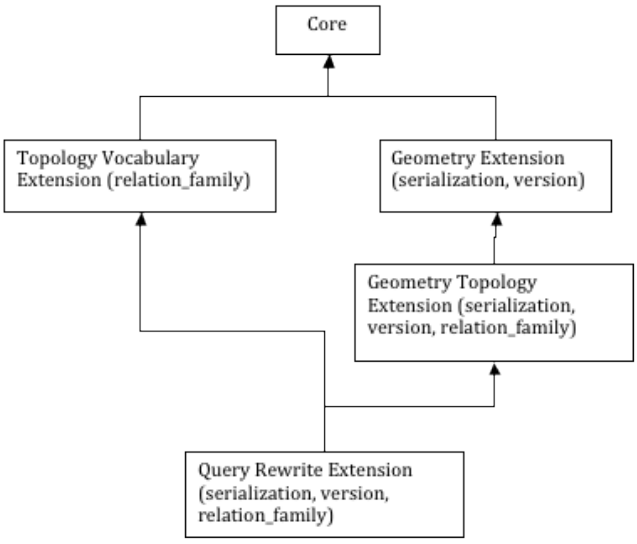


Figure 74: The relationship of the five GeoSPARQL components. Reproduced from the OGC GeoSPARQL working group standard proposal report *GeoSPARQL: a Geographic Query Language for RDF Data* (Open Geospatial Consortium 2011b, 2).

The GeoSPARQL standard as currently proposed, includes serialisations for WKT and GML, but the working group states they will likely add KML and GeoJSON serialisations in future. Because STELLAR conforms to MIDAS, which uses WKT and Simple Features, it may be possible to begin experimenting with GeoSPARQL fairly soon. The GeoSPARQL standard proposal also states there are 'large amounts of existing feature data represented either in a GML file (or similar serialization) or in a datastore supporting the general feature model. It would be beneficial to develop standard processes for converting (or virtually converting and exposing) this data to RDF' (Open Geospatial Consortium 2011b, 31), so ideas about further functionality are already underway.

4.7 Future work

The CRM-EH began as a way to better understand how data generated by archaeological fieldwork at English Heritage was being created and used. As such, it is understandable it was modelled on the way archaeology is typically undertaken in the UK, i.e. using single context recording. More than once, during discussion of the CRM-EH at CAA meetings, critical comments have been made stating the CRM-EH is not appropriate for use outside the UK, and is therefore of little use. Those making the comments often seem to be missing the point. The CRM-EH was never intended as a universal domain ontology for archaeology. It was always presented as an experiment in ontology creation based on a mapping of actual data use, and it just happened that initially the people willing to develop it and share what they created worked for English Heritage. The fact that so much work and testing has gone into the CRM-EH in the form of the Revelation, STAR and STELLAR projects means it has had the chance to mature and grow. This process has been very transparent and included the creation of tools and outcomes made freely and publically available, but it may also be this good deed that has added to the perception that the CRM-EH is being presented as a universal domain ontology for archaeology.

Is it possible or even desirable to create a universal domain ontology for archaeology? Could the CRM-EH be used as a basis for this development; where single context recording is one of several standard recording systems that could be included? The name CRM-EH was chosen because it was based on the CIDOC-CRM and originally modelled on English Heritage data, (and it had to be called something), but the creators no longer think of the CRM-EH as something proprietary to English Heritage, and would not object to a change (Keith May pers. comm. April 2011). Because of the extensive time and effort put into its development, perhaps the CRM-EH is the best starting point for exploring the development of a domain ontology for archaeology. It could be adapted to provide universal interoperability at a coarse level of granularity, and then sub-domains

could be created for use with different recording systems. The exploration done for this research makes further interoperability seem both desirable and within reach, and the next logical step is to allow interaction with data from outside of the UK, which may or may not have been gathered using single context recording. Continued work and discussion in this area would be most welcome.

STELLAR already provides a massive amount of the bridge required to move geospatial data from a native vector format like a shapefile to useable RDF, but for this research, it still took two other steps to bridge the gap between a shapefile and STELLAR. FWTools does a quick and straightforward job of translating the shapefile into GML, but as nothing existed to convert the GML to CSV in an automated way, the STELLARPreloader Java application was developed. It might be possible to incorporate the functions of the *ogr2ogr* transformation into the STELLARPreloader and reduce the number of steps needed to go from shapefile to CSV ready for use with STELLAR down to one, but it still means running command line Java in between WYSIWYG GIS programs and STELLAR.
If the user is only comfortable with STELLAR.Web, then it still represents a technological jolt, which may make the process seem too difficult.

The STELLARPreloader is available on the CD included with this thesis, and permission has been given by the developer to make the code freely available for use and further adaptation, but optimally a nice WYSIWYG version incorporating the functionality of the *ogr2ogr* transformation from FWTools and the STELLARPreloader should be developed. There is little standing in the way of combining these steps, as apart from initially defining which columns in attribute tables associated with a shapefile should be included, no manipulation or decisions are required along the way. This tool could be made available alongside STELLAR.Web (or whatever non-prototype version of STELLAR.Web is ultimately created). Optimally, a WYSIWYG interface could also be developed for choosing which columns within a CSV file correspond to the STELLAR mapping.

Just this sort of interface has been developed as part of the Ports Network Project. It is a mapping tool called a 'Data Inspector Wizard', where someone with an understanding of the data (though not necessarily an understanding of the ontology) is guided through the process of choosing the correct mapping through a GUI. It uses Natural Language Processing to help predict which mappings might be correct, but the user ultimately has control over the choices. Importantly, the Data Inspector Wizard also generates a separate mapping of the choices in the form of an XML configuration file, which can be made available alongside the data (Isaksen *et al.* 2009b, 132-3). This allows potential users to view the mapping choices, and make decisions about its potential interoperability with other data. Uploading a file into STELLAR.Web, choosing a desired template and then adding mapping steps similar to the Data Inspector Wizard would not only allow further automation to the process (without a loss of control) it would likely lead to more consistent alignments by users. The ability to generate a mapping file at the end of the process would also be very useful, as not only would it be time saving, it would remind users of the importance of making their mapping choices transparent for other users, and increase the likelihood that they would include this information with their data.

In addition to seeing the potential for expanding the CRM-EH through this research, the usefulness of the STELLAR templates cannot be emphasised enough. The STELLAR.Web format will likely be adequate for the vast majority of archaeologists, which will effectively allow anyone to prepare their data either for use within an RDF store, or publication as Linked Data, with virtually no specialist knowledge of Semantic Web technologies. For those who need greater control, STELLAR.Console allows considerable customisation and the ability to create custom templates. The use of templates allowing a modular format works well. A single mapping file would be unwieldy, and because the ontology allows the data to be 'stitched together' within the RDF store once it is loaded, the format of the templates can change and grow without losing interoperability. The decision to construct STELLAR using templates being sound, it would be

beneficial to have additional templates. As the current implementation is meant to be a demonstrator, only a core group of templates was developed. The assumption was, if found to be useful (and further funding secured), the templates could be expanded (Ceri Binding pers. comm. February 2011).

There are many directions in which the STELLAR templates could be expanded, but for this research, it would be useful to have more geospatial data represented. Currently, it is only possible to incorporate a single set of x/y coordinates for a centroid for each site, context, find, etc, but rather than create a separate template comprised only of geospatial values it would probably be sufficient to add a few more fields to the existing templates. Looking at what would be necessary to accommodate the additional information exported from the Cottam and Hungate drawings, fields could be added for area and perimeter values, and the group of coordinates that make up a polygon. If a full complement of geospatial data could be added, then creating a new template only for geospatial data might be a better solution, but this would need to be thought through. If the links between the templates became more complex, the potential for mistakes or misinterpretations when aligning the data within the CSV files increases.

If templates were expanded to take full advantage of GeoSPARQL, two new STELLAR templates could be created. GeoSPARQL is meant to be interoperable between both ‘feature’ and ‘geometry’ data. Geospatial ‘features’ are any real-world entities that have a location, whereas ‘geometries’ are the actual geolocated points, lines and polygons representing a ‘feature’ (Battle and Kolas In Press). The GeoSPARQL ontology uses the ‘geo’ namespace abbreviation and has the following structure:

Class:
geo:SpatialObject

Subclass:
geo:Feature
geo:Geometry

Connecting property:
geo:hasGeometry

Using the basic GeoSPARQL structure, perhaps two new geospatial templates could be created for STELLAR to link to any existing templates which have geospatial data, one for geo:Feature and one for geo:Geometry, which could then be linked to each other via geo:hasGeometry when actual vector data was available alongside the 'feature' data. As many of the STELLAR templates would have geospatial data specific to their data types, the application would have to be thought through. Under this scenario, two geospatial templates would have to be added for any STELLAR templates that have a 'location' designation. In its current configuration, this would include CRMEH_Contexts, CRMEH_Investigation_projects and CRMEH_Groups, for a total of six new templates, therefore a simpler solution would probably need to be found.

This research has focussed on the CRM-EH templates included with STELLAR, but there are general CIDOC-CRM and SKOS templates available as well. There is a CLassical Art Research Online Research Services (CLAROS) template for objects included in STELLAR. CLAROS is an international research federation using the CIDOC-CRM as an interoperability foundation for information about objects and images across multiple museum and university collections, and as the CRM-EH is a CIDOC-CRM extension, has coarse grained interoperability with CLAROS (CLAROS 2011). Another non-CRM-EH template that might be useful could be based on the GeoNames OWL ontology for geospatial placenames. A wide variety of organisations using Semantic Web technologies are using GeoNames, including well known applications like Wikipedia (and therefore DBPedia), Ordnance Survey, the US Department of Transport, and the statistics

agencies of many countries, so a STELLAR template for GeoNames would allow interoperability in yet another direction for the CRM-EH (GeoNames 2011a).

A couple of other items which might be useful alongside STELLAR would be a generic compatibility claim form, so that users could easily fill in how they mapped their individual data to each of the STELLAR templates. Having the form available alongside the other STELLAR tools would be a good reminder that compatibility claim forms are important to include whenever data is published for re-use, and by making the process easier, would likely ensure more users created them. If there is interest in using tools like D2R to publish relational data virtually as RDF, then it might be useful to create a D2RQ mapping file for the CRM-EH. That way, someone using any part of the CRM-EH would not have to create a mapping file, and in much the same way as the STELLAR templates ensure interoperability with the CRM-EH, a definitive D2RQ mapping file would ensure all Linked Data published using D2R would be interoperable. It would also ensure data in RDF created with STELLAR would be interoperable with relational data published using D2R.

If geospatial data aligned to GeoSPARQL could be incorporated into STELLAR, and the GeoSPARQL extension to the SPARQL query language supported within a SPARQL endpoint, then real geospatial querying of archaeological data using Semantic Web technologies will be possible. Having to learn SPARQL will continue to be a hindrance for non-specialist users however, and visual interfaces need to be created where SPARQL and GeoSPARQL are doing the work behind the scenes. The STAR demonstrator is still probably the closest implementation of a generic user interface for archaeological field data available, but visually sophisticated interfaces for Semantic Web data in archaeology are beginning to appear, and some have geospatial components.

The CLAROS project (mentioned previously) launched its first public interface in May of 2011, which allows free text searching and faceted searching by category,

place, period and/or data collection, which can be further refined through map and timeline views (CLAROS 2011). The Herodotus Encoded Space-Text-Imaging Archive (HESTIA) project has created three different geospatially based interfaces for navigating the locations mentioned in Herodotus' *Histories*, including Herodotus in GIS, which allows data to be queried with the users standard GIS desktop program, Herodotus in GoogleEarth which allows the data to be queried using GoogleEarth and KML, and Herodotus' Narrative Timeline which links the text of the Histories in both time and space, in the original Attic Greek and in English translation (HESTIA 2010).

The Pelagios: Enable Linked Ancient Geodata In Open Systems (PELAGIOS) project also makes several key datasets related to the Ancient World interoperable, and does so with geolocation at its heart, but specifically trying to incorporate Linked Data. The first version of the PELAGIOS Graph Explorer was introduced in August of 2011. Not only does it provide an intuitive, graphical interface for navigating the data using geolocation, the interface itself takes advantage of the structure of Semantic Web data as a navigation tool (PELAGIOS 2011). While CLAROS and HESTIA create interfaces that mimic traditional interfaces, but have interoperable datasets behind them, PELAGIOS creates a true graph data interface. Users are presented with circles representing each of the available datasets, and clicking on the datasets connects them and combines their data, and the thicker the line between each dataset, the more sites they have in common. Choosing to view them on a map brings up all the locations the chosen datasets have in common, mousing over one of the circles in the datasets brings up a polygon showing the area the data falls into, and clicking on a site brings up a list of records. Clicking on a record from the list takes the user directly to the record. Despite being in its first test version, the PELAGIOS Graph Explorer is an elegant, simple and very intuitive way to navigate through heterogeneous datasets using geolocation. Even though only x,y datapoints are used in all these projects, it shows how geospatial data is at the heart of these early exemplars for visualising and navigating Semantic Web data for archaeology.

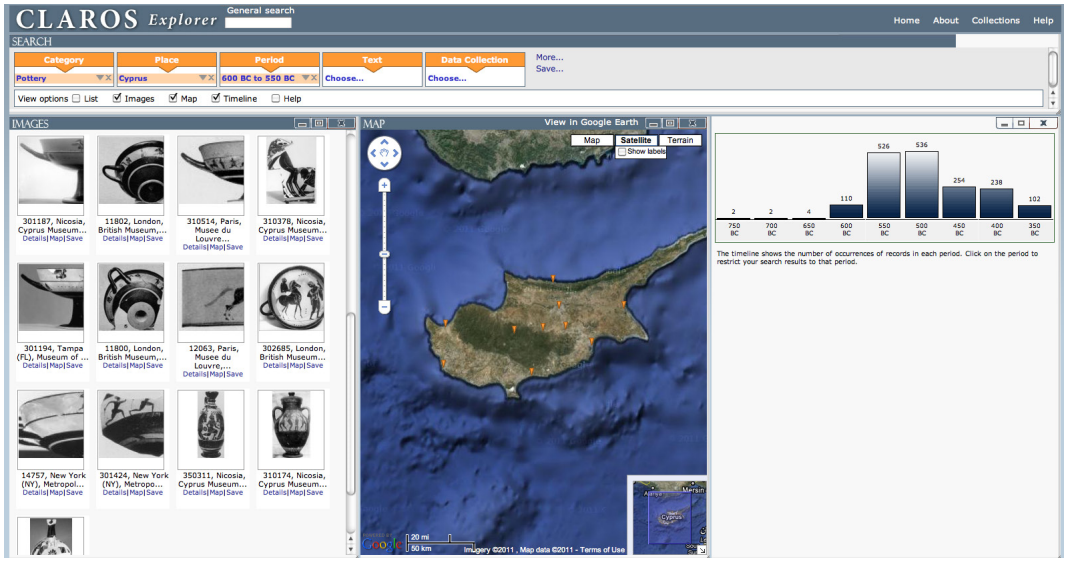


Figure 75: Screenshot of the sophisticated CLAROS interface, which allows faceted and free text exploration of archaeological resources, which can be refined through map and timeline views. Query shown is pottery found in Cyprus dating from around 600 BC. Images returned are sourced from a large interoperable dataset. http://explore.clarosnet.org/XDB/ASP/claroshome/.

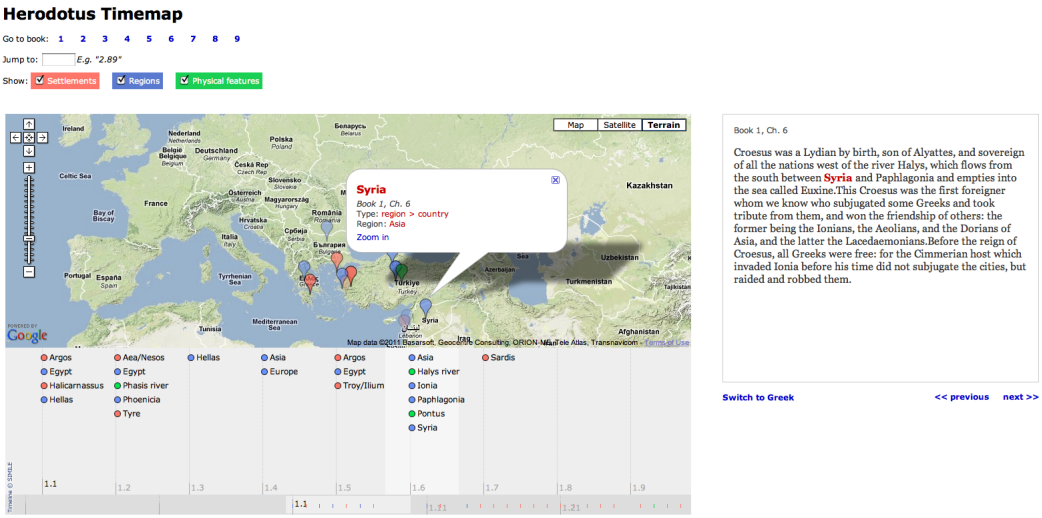


Figure 76: Screenshot of the Herodotus Encoded Space-Text-Imaging Archive (HESTIA) project's *Herodotus' Narrative Timeline,* which links the text of Herodotus' *Histories* in both time and space, in the original Attic Greek and in English translation. Text mentioning Syria in Book 1, Chapter 6 shown. http://www.open.ac.uk/Arts/hestia/herodotus/basic.html.

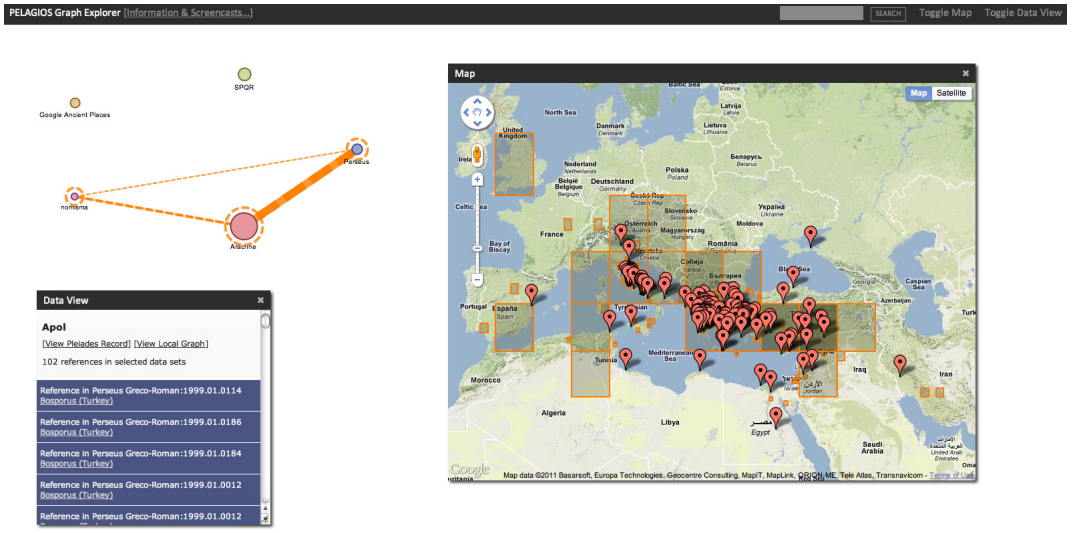


Figure 77: Screenshot of the Pelagios: Enable Linked Ancient Geodata In Open Systems (PELAGIOS) project Graph Explorer, showing three datasets selected; the thickness of the line indicating that the Arachne and Perseus data have the most sites in common. Viewing the sites on a map shows the site location, and clicking on a site (in this example, Apol) shows a list of references. Clicking on a reference brings up the specific record. http://pelagios-project.blogspot.com/.

Another promising visual interface that takes a more traditional RDF store query approach, is the Semantic Explorer for Archaeology (SEA). SEA is part of the *Tracing Networks: Craft Traditions in the Ancient Mediterranean and Beyond* project and is a bespoke Web interface allowing querying, interaction, visualisation and statistical analysis across the seven Tracing Networks datasets. Of particular interest is the incremental query view interface, which allows users to build SPARQL queries through a WYSIWYG interface and then apply filters to the data. The process is similar to the graphical query builder in Gruff, but arguably easier to navigate for a non-specialist, despite being more text-based. Statistical analysis can then be applied and visualised using pie or bar charts, or if the data is geolocated, viewed on a map. Map-based query users can also choose whether they want to view the current location of an object (within a museum collection, for example) or the location of the site where the object was found. The query itself can also be visualised similarly to Gruff (Solanki *et al.* 2011). Though SEA is currently under development for use only with the data from *Tracing Networks,* there is great potential here for generic adaptation for use with a broad range of archaeological data.

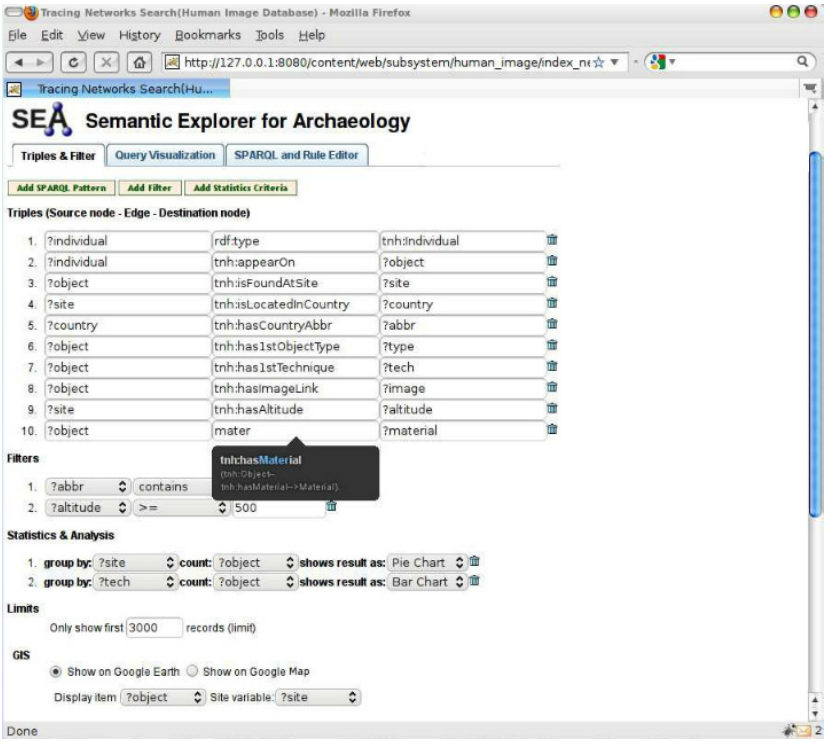


Figure 78: Screenshot of the Semantic Explorer for Archaeology (SEA) query interface, which is part of the *Tracing Networks: Craft Traditions in the Ancient Mediterranean and Beyond* project. The screenshot shows the query building interface, including the application of filters. Statistical analysis can then be applied and the results visualised as pie or bar charts, or in the case of geolocated results, visualised on a map (Solanki *et al.* 2011).

Based on these sophisticated recent applications, exemplars of good visual interfaces incorporating geospatial data are now being developed. With the recent developments in geospatial querying like GeoSPARQL, it may be possible to move beyond the use of single x,y coordinates into using true 'geosemantics'. The creation of interface tools like these mean non-specialists will be able to use Semantic Web data, and the advent of tools like STELLAR means non-specialist archaeologists will be able to create Semantic Web data as well. In addition to true geospatial querying, the creation of generic visualisation tools which have the functionality of STAR, CLAROS, HESTIA, PELAGIOS and SEA, and can be used with any archaeological data are also needed. Clearly we are on the cusp of several developments coming together to make it possible to do this in the not-so-distant future, and there is no shortage of work to be done.

4.8 Conclusion

The research presented here has attempted to answer the questions posed at the start of the chapter. It explored whether it was yet possible for non-specialist archaeologists to take advantage of the Semantic Web, looked at some of the obstacles to be overcome, and whether there are generic tools available which can be adapted for use with archaeological data. It also explored the special characteristics and requirements of data derived from archaeological field drawing, including the use of geospatial vector data alongside textual data within the Semantic Web. In order to demonstrate interoperability, spatial data from two different archaeological sites were used, including plan drawings and their associated data, which were created by two distinct organisations with differing field methods, data collection techniques, and data manipulation practices. The datasets were chosen because they fell into the same early medieval, eighth-tenth century AD, Anglian and Anglo-Scandinavian time period, and while they were related archaeologically, they differed technologically.

It was hoped that combining these two structurally heterogeneous, yet archaeologically related datasets using Semantic Web principles would facilitate new questions, and perhaps new comparisons between urban and rural settlement in York and the surrounding area, that might not have been possible by other means. At a basic level, this was possible, as simply combining the data in an interoperable way and then being able to query across both datasets simultaneously allowed whatever patterns and information were held within the data to be easily seen, but further technological development would be required to pose more nuanced questions. Perhaps when the Anglo-Scandinavian dataset from Hungate is complete, it will be achievable. Is was possible however to take some of the other research objectives further, and in some cases, expand them to a greater degree than was anticipated at the outset.

For non-specialists, working with archaeological data using Semantic Web technologies is not easy. The learning curve to understand what the Semantic Web is, and how it might benefit archaeological practice is steep, even at a theoretical level. Thinking about how to organise and use data from within a 'haystack' rather than housed within the structure of a traditional relational database, takes a mental shift. Learning about ontologies and what is appropriate for use with their data is another complexity that takes real time and understanding for proper use. Then there is the level of specialist computing knowledge that has been necessary to actually use Semantic Web technologies. Just to get data into RDF, correctly aligned to an appropriate ontology, required either solid knowledge of UNIX and a programming language like Java, or sufficient understanding of how the Semantic Web works at a technical level to combine a variety of disparate WYSIWYG programs in a way that produces the desired result. In either case, the process is messy, difficult, and well beyond what any non-specialist would wish to undertake. As defined by Leif Isaksen, for the Semantic Web to see uptake by archaeologists, this process must be:

1. Quick. While there can be no definition as to what constitutes 'quick enough', 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 thereby more burdensome.
2. Cost-effective. It must use freely available (ideally Open Source) software and require minimal, if any, technical support.
3. 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.
4. Transparent. The archaeologist must understand enough of the production process to feel confident in its output (Isaksen 2009, 16).

While these points refer primarily to the production of Linked Data, they are just as true for creating RDF for any Semantic Web application. For archaeologists to use

the Semantic Web, they must be able to process their data quickly and easily. They will need to continue the well-worn tradition of using tools which are free or very low-cost (often developed for other sectors) and don't require significant adaptation or support. They will also need to know their data will be correctly aligned to their chosen ontology and correctly formatted, and be able to verify it. STELLAR has demonstrated, at least for those using single context recording, that this is indeed within reach. Given the ease with which it is now possible to set up a generic RDF store with a WYSIWYG interface, archaeologists can expect to process their data, and work with it within hours, with a minimal amount of specialist knowledge. STELLAR does not just show it is worth developing tools for archaeologists to translate their data into RDF, but it emphasises the importance of using tools like templates to ensure consistancy, and therefore true interoperability.

The situation with data derived from field drawing is more complex, and provisions for including spatial data alongside the textual continue to evolve. Several commercial providers of RDF store software have attempted to create proprietary geospatial features to fill this void, but a standards-based, non-proprietary solution that stands apart from any particular software and can be used equally with all, is the better scenario. Fortunately, work on GeoSPARQL, the OGC standard to define a vocabulary for representing geospatial data in RDF, as well as providing an extension to SPARQL for querying that data, is well underway, and should be submitted for an OGC standardisation vote by the end of 2011 (Open Geospatial Consortium 2011a). The usefulness of the CRM-EH as an ontology for archaeological fieldwork has been demonstrated here, though some of the successful translation into Semantic Web technologies is down to its foundation in single context recording. Single context recording allows spatial relationships to be kept intact within an RDF triple, no matter how abstracted the data becomes within a Semantic Web graph. How this could be done using other types of archaeological recording systems is unclear, but should be explored. Expansion of the STELLAR templates to include more geospatial data would be welcome as well, especially if it were possible to build them around the new

GeoSPARQL standard. It is hoped that the practical application presented in this chapter, along with the discussion of the potential geospatial functionality that may be possible in the near future, has demonstrated that it is possible to create tools that will allow non-specialist archaeologists to use the Semantic Web. It is hoped that the workflow and demonstrator tools can be expanded to include the full complement of archaeological data, including other types of recording practice, so that the sharing and use of archaeological field data will be possible, wherever and whenever it is useful.

That said, if the Semantic Web technologies explored in this chapter do come into use more generally within archaeology, what might the impact be? Looking back on the history of archaeological practice, the introduction of new technologies has always been subject to theoretical analysis, as it should be, and the Semantic Web will be no different. In the seminal 1992 text on the rise of computing in archaeology *Archaeology and the Information Age: A global perspective*, much attention is paid to access and exclusivity; to whether the use of information technology will be a great leveller, or a way to increase the hold researchers and research institutions have on our shared heritage. To put *Archaeology and the Information Age* into historical perspective, it is an edited volume resulting from a series of meetings held at the Second World Archaeology Congress (WAC) in Barquisimeto, Venezuela in September of 1990. The resulting text includes papers circulated to the participating researchers prior to the meeting, and so reflect work carried out in the late 1980s (Ucko 1992, vii). As the discussions were part of a WAC, where the spirit of the congress is to be international and inclusive, it is understandable that ideas about the accessibility of archaeological data loomed large:

> If archaeological data dissemination, and archaeological dialogue more generally, is intended to take place over such networks, is it simply naive to hope for what might be termed the democratization of archaeological knowledge on a global

> basis? Alternatively, are we on the brink of a period in which transaction processing, usually associated with the banking and merchandizing business worlds, will be necessary to exchange data, and does this herald the rise of specialist brokers of archaeological data? (Reilly and Rahtz 1992, 13)

This was written on the eve of the World Wide Web bursting onto the scene, so the 'networks' to which they refer would have been pre-Web forms of Internet-based connections and email, and the 'transaction processing' associated with the business world now reflects the way we negotiate most things in our everyday Web-based lives, but both their hope and their fear are remarkably prescient with regard to the Semantic Web. The 'democratization of archaeological knowledge' sees its purest form thusfar with the creation of Linked Data; raw data made available for use and re-use by anyone with access to the ever more ubiquitous Web. At the same time, the creation and use of Semantic Web data has largely been the purview of 'specialist brokers' either because of the difficulty inherent in creating data to share with others, thereby excluding the average archaeologist from incorporating their own data within the Semantic Web, or the lack of useful tools and user interfaces to allow non-specialist archaeologists to use the data of others.

This is why tools like STELLAR are just as key to the 'democratization' of the Semantic Web for archaeology as Linked Data. Just making the data available is very important, but if archaeologists are going to participate fully, it is of equal importance that they are able to use and incorporate their own data. Tools like STELLAR allow the technically difficult, but intellectually mundane aspects of creating Semantic Web data, to be carried out easily. The only non-automated point in the creation process; the point at which decisions must be made as to how the data should be aligned to the ontology, is the point where the archaeologist must step in. This allows the individual most familiar with the data to make the first level of Semantic Web interpretation, and simultaneously

frees them from reliance on ‘specialist brokers’ for processing their data. While STELLAR is currently only available for use with data created using the single context recording tradition, and should not be considered a panacea for use with all archaeological field data, it does clearly show that further ‘democratization’ of Semantic Web data for archaeologists is achievable. The next important step being the development of similar tools for other recording traditions, which would hopefully include a reasonable level of interoperability with ontologies like the CRM-EH to further this trend.

Much of the technical work demonstrated in this chapter goes beyond what an archaeologist wishing to use Semantic Web data needs. Showing that it was possible to set up an RDF store and Web interface with minimal specialist knowledge and using free and generic tools was a useful exercise, but by no means necessary to get started. After running the data through STELLAR, it could simply be loaded into a generic visualisation tool like Gruff, in just the way you open any file in a desktop application, the archaeologist is up and running, and data downloaded from other SPARQL endpoints or Linked Data repositories could be incorporated as well. Using an application like Gruff is no different than learning any other computer application, like database or GIS software. It can be therefore argued that the availability of tools like STELLAR and Gruff help to address Reilly and Rahtz’s foundational concern; that it is important to not only provide access to the data derived by the destructive process of excavation, which forms our shared heritage, but to the technology and expertise necessary to use it.

In addition to the implications to archaeological research inherent in access to data, technology and expertise, how does organising data by aligning it to an ontology and storing it in graph format change the way we perceive and understand that data? It can be argued that aligning data to an ontology like the CRM-EH is a much more flexible way of organising data than within a traditional relational data structure. Relational data is typically stored within a static and closed format, where the more closely the data conforms to the structure, the

more accurately it can be queried. It is also typically meant to suit a particular purpose, and not required to be similar enough to other databases to allow the data to be combined. By aligning data to an ontology like the CRM-EH, it is really being stored within layers of description, with varying levels of granularity, and not just within the CRM-EH, but within the CIDOC-CRM as well. For example, because of the nested and hierarchical nature of ontologies, the triple which includes EHE0003_AreaOfInvestigation and its value, describing the name of a particular archaeological site, is automatically mapped to the superclasses of EHE0003, which are E53_Place and E1_CRM_Entity respectively. Add to this the augmentory functionality of things like SKOS, and the flexibility inherent in mapping to an ontology, or a group of related ontologies becomes apparent. The ability to add new mappings along the way also makes it easier to find ways to accommodate the addition of data which might not have been originally envisioned as being part of the data store.

Once the data is aligned to an ontology and stored in RDF format as graph data, perceptually, it is in a fundamentally different format from relational data. While the human mind may be structured in a nonlinear, non-hierarchical way which is much more similar to graph data, most archaeologists have long experience using relational databases, and we have trained ourselves to think about organising data in the way relational databases are structured, in order to use them properly. When confronted with an unfamiliar database, we spend time looking at the data structure and acceptable values to get a sense of it. Working with graph data can feel much more mercurial. Browsing or ‘clicking through’ triples and being able to move in many directions without a sense of where you are within the data can be disconcerting. Even if a user familiarises themselves with the structure of the ontologies in use, often only a subset of a given ontology is actually present. Querying for a list of the actual classes and properties in use within the datastore will show the nature of data, but may not feel like the most intuitive way to understand how data is structured for those first learning about the Semantic Web.

Another important difference between graph and relational data is the way graph data can be pulled apart and stitched back together in different ways. How does this very abstracted way of working with data influence our perceptions of it? After our long history of working with relational data, do we archaeologists need to train our minds to think in a more natural and nonlinear way before we can use it to its full potential? Thusfar, the answer to this seems to be yes, as strong demonstrations of the Semantic Web answering new and different archaeological research questions with Semantic Web technology are still thin on the ground. Bespoke projects with simple, intuitive interfaces like Pelagios do the best job of allowing users to cognitively engage with the data, but these are still primarily ways to query across aggregated datasets, and then drill down to the specific information of interest. Now that it is becoming easier for archaeologists to use the Semantic Web, it is time to begin creating exemplars of what new research questions might be asked and answered using this technology. By demonstrating what might be possible, it will become easier to understand and work with this very abstract data format, and in doing so, begin to build the necessary wider critique of its usefulness, biases and potential impact on archaeological practice.

Having established the importance of visual data, and particularly field drawing, to the archaeological record in Chapter 3, what are the implications of using this kind of data with Semantic Web technology? Does the abstraction of graph data add different forms of cognitive challenge and interpretation than other types of digital visualisation? It is useful to return to Stuart Piggott's pre-digital understanding of the translation of what is seen by the excavator to what is recorded:

> Archaeological draughtsmanship involves the construction of technical cryptograms, and as in all ciphers these must be made according to rules carefully observed by both transmitter and recipient. As symbol, all illustration is a transcript of reality... The draughtsman's illustrations are no more passive agents of

> communication than the author's words they complement and expand. A drawing must say something or it is failing in its primary purpose, exactly as a sentence or a paragraph of text must say something economically or elegantly, in clarity or in confusion (Piggott 1965, 165).

By aligning data to not only a set vocabulary, however flexible and expandible, but to a group of relationships found within that data, is to build a further sort of 'technical cryptogram' to which Piggot refers. The 'rules carefully observed by both transmitter and recipient' are made explicit in the choice of ontology, and by documenting and communicating how the mapping choices were made. As such, the inclusion of the drawn archaeological record within Semantic Web data is no different than any of its analogue predecessors, or indeed the several iterations of interpretation through which digital field drawing data is already subjected. As long as the rules are made explicit, the user can make up their own mind about relative usefulness and appropriateness of the data to their research.

If done in a suitably transparent way, this further level of translation and interpretation should constitute an acceptable trade off for the potential ways in which the data might be used. That said, while the process of translating field data into RDF for use with Semantic Web technologies should be valid within archaeological practice, its use remains largely mute. It was shown earlier in this chapter that it is possible to extract polygon data for future incorporation within the Semantic Web, but the technology to make use of it does not yet exist. As important tools like GeoSPARQL continue to develop, it is useful to speculate as to how this might take shape, and what the theoretical implications of its use might be.

To fully express the archaeological record within the Semantic Web, archaeologists will need to convey the visual nature of the drawn archaeological record, at least as a form of analytical tool. As such, the various iterations of

GIS software seem the most logical interface form to emulate. The ability to geolocate points, lines and polygons and visually represent and interact with them alongside what is typically defined as attribute data (though for the Semantic Web, what would constitute attribute data would be undifferentiated from the polygon data held within the RDF store) and augmentory layers to help understand the data more fully. The theoretical implications surrounding the use of GIS within archaeology are well documented (Wheatley and Gillings 2002, 8; Lock 2003, 182; Zubrow 2006, 22-23), and presumably visualising Semantic Web data in a similar way would be subject to the same sort of theoretical arguments. As GIS has found wide use within archaeology for many years, its approach has apparently been deemed sufficiently appropriate to warrant its continued favour, so the use of a GIS-like analytical interface for Semantic Web data should have similar potential.

Envisioning the advantages of being able to use the full complement of data from archaeological field drawing, with a GIS-like interface which can respond to dynamically generated GeoSPARQL queries against Semantic Web data is quite exciting. In particular, the non-linear nature of data structured as RDF triples might allow for very fluid visualisations of temporal phasing which would not be possible with traditional means. The basis of single context recording being that interpretation should be separated from the excavation process as much as possible, the abstract nature of Semantic Web data might be considered an advantage to keeping the information in an oblique format, until the time comes to actively analyse it. Once the data is ready for analysis however, the ability to create complex and nuanced queries which are visual, fluid and are able to combine data in new and structurally different ways could be an elegant solution to the interpretive questions archaeologists have been trying to answer throughout the history of the discipline. Being able to see and understand the stratigraphy and fundamental spatial relationships with new patterns and relationships which can be easily changed and reshuffled based on new ideas and criteria might offer a further way to extend the usefulness of existing data, and to create a deeper

understanding of what that data can teach us about our human history. As such, it is important to continue to move forward with the development of the necessary technologies, in the hopes that archaeologist will be able to take advantage of them in the very near future.

Chapter Five

Conclusion

Throughout its development, there has been no shortage of naysayers ready to proclaim the vision of the Semantic Web dead; its long incubation period becoming an equation for a useless technology. This, coupled with a decent amount of confusion over what the Semantic Web is, what it is called, if it should be called something else, or if it should be done differently, all continue to stir the pot. In September of 2010, eminent technologist Tom Coates spoke at the sixth *dConstruct* gathering in Brighton, where part of his presentation included a slide that simply said 'Death to the Semantic Web' (Coates 2010). While Coates was largely trying to be humorous and provocative, his point was that the Semantic Web is designed to be implemented from the top down, and as an advocate of social media as a means to build the Web of Data, its continued development was an impediment in need of slaying. For Coates, the incremental networks built by users of the Web are the way forward, and interestingly he chose an archaeological example of the road system built by Darius the Great across the Persian Empire to illustrate his point.

The creation of Darius' road system generated a knock-on effect of new and complex interconnections, built through the grassroots contributions of others. Archaeologists will appreciate this causality, but will also take issue with his analysis, as many things had to happen in order for Darius to be in a position to have the road system built in the first place, and decisions had to be made about where to build the roads and why. The assumption made by Coates is that the Web

is the road system, and everything else should be allowed to grow organically from there, but what if the Semantic Web is an extension of the road system necessary for this organic growth? How can the infrastructure that will lead to real interoperability be created if Web standards are ignored in favour of (largely) commercial solutions? The hard lessons of the Browser Wars of the 1990s seem to have been quickly forgotten.

Figure 79: Tom Coates proclaiming 'Death to the Semantic Web' at *dConstruct* in September of 2010. Flikr photo by happy.apple: http://www.flickr.com/photos/29022619@N03/4968410475/.

As argued in the second chapter of this thesis, there is plenty of room for development in both directions. To revisit the cave analogy, the stalactite of the Semantic Web may not meet up with the stalagmite of Web 2.0 to form a perfect cave column, but it shouldn't be a matter of getting there by any expedient means either. The Web of Data must be made genuinely useful as more than an immediate resource for individuals; it must be useful within and across knowledge domains like archaeology in a lasting and sensible way, and that requires building some secondary roads in the right places first. The perception that the top-down approach of the Semantic Web as envisioned by Berners-Lee is simply too difficult, or will require so much infrastructural work as to negate its value has been well explored with regard to Cultural Heritage by Isaksen *et al.* (2009a;

2011; 2010). Kirk Martinez and Leif Isaksen recount the specific history of how the Cultural Heritage sector has struggled to engage with the Semantic Web:

> The DigiCULT Project brought together a panel of 13 European experts in 2003 to discuss Semantic Web development in cultural heritage but the plethora of nascent (and competing) technologies at that time…resulted in the Semantic Web being described as a 'Shangri-La' surrounded by a 'veil of mystery'. Nonetheless, a number of participants concluded that 'they would put their money on the Semantic Web' whilst other contributors maintained that 'the heritage sector is likely to be left behind'. Five years later, the Semantic Web Think Tank project…concluded in 2008 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 appeared to create 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.' (Martinez and Isaksen 2010, 31-2).

They analyse the reporting from the Think Tank meetings, and find that a variety of concepts and technologies were under discussion as though they were all part of the Semantic Web. In reality, many were not, and of those, the majority were associated with Web 2.0. They believe this is an indicator of confusion about the Semantic Web, even after extensive discussion, and this confusion was the major source of the negative feelings about it. (Martinez and Isaksen 2010, 32). If so, this means much more needs to be done to clarify and demonstrate what the Semantic Web is and is not, rather than take it as a rejection of the vision as unworkable for the Cultural Heritage sector.

They go on to emphasise that these two areas are not in competition within the Cultural Heritage domain, and both are useful in different and complementary ways (Martinez and Isaksen 2010, 33). Surely this is the case, and any call to slay the Semantic Web so as not to impede the progress of Web 2.0 is at best unnecessary, and at worse, wrong-headed. The existence of the Semantic Web (and at this point it seems fair to say that it exists) does not hinder the growth of the social Web by channelling development energies into a fruitless direction. If anything, interest in the Semantic Web has made developers focus on how the Web could and should move forward, resulting in a more sophisticated understanding of its growth. At the same time, Web 2.0 has kept the Web growing, changing and full of energy, thereby heightening the participatory activity in a way that makes the sharing of information, which is central to the Semantic Web, feel more natural and acceptable.

This change of perception is important. When the seminal Scientific American article by Tim Berners-Lee and his fellow authors was published in 2001, the futuristic scenarios it predicted, with every piece of information about a person's interests, habits and commitments available for use on the Web, probably sounded downright invasive to many people, even though it was information about themselves (Berners-Lee *et al.* 2001). The growth of the social Web over the last decade has shown that people seem to value connection over privacy however, and we have Web 2.0 to thank (or blame) for that. This change has been so pervasive that it has led to an expectation that information *should* be shared. Even governments, which may have only felt people had a right to access their data previously, now began to feel the pressure to actually make it *accessible*. Without Web 2.0, the cultural shift making people *want* to share their data, which is of fundamental importance to the success of the Semantic Web, would be much more difficult.

Archaeology has not been immune from this either, despite being notorious for research going unpublished. Funders now typically require dissemination and

publication commitments, and in some cases the public archiving of primary data as well (Takeda *et al.* 2010). The move within the last few years by the UK Arts and Humanities Research Council (AHRC) requiring the data produced by their funded projects be deposited with the Archaeology Data Service in some form of publicly available archive is an example of this. Archaeological excavation is a destructive act, where the primary data archive becomes the archaeological resource. This dictates that data should be made available for use by future archaeologists, and if the expansion of Web 2.0 encourages this mindset, it is all to the good.

If sharing data has become a comfortable and commonplace phenomenon on the Web thanks to Web 2.0, and continued development of the Semantic Web is providing the infrastructure to bring about a more meaningful Web of Data, how best do we model that meaning for archaeology? What is an acceptable depth of meaning for archaeological data? Martinez and Isaksen recount the practical and theoretical problems of choosing to map to a generic ontology like the CIDOC-CRM directly:

> …it is better for cultural resource providers to first map data to their own local ontologies, and only then to align them with more generic Domain Ontologies (such as the CIDOC CRM) as a second step…It is quite difficult enough to map between two explicit world-views, let alone, convert between data formats and implicit ontologies at the same time! Secondly, there is a growing realisation of the importance of multivocality in areas of contested heritage…We do not want to throw the baby out with the bath water by creating 'one ontology to rule them all'. It is our belief that the next step to be taken in the Cultural Heritage sector is for organisations to render their own ontologies explicit using tools…Once they can express the 'deep' nature of the data in their possession, we can begin to

> align them and, in so doing create a much more powerful body of information than has previously been available (Martinez and Isaksen 2010, 44).

Here they acknowledge the importance of the understanding archaeologists have about the characteristics of their own data, and how that understanding must be preserved and respected. They also acknowledge the practical difficulty in mapping an actual dataset to a generic ontology, and the level of expertise required being far beyond what most archaeologists would care to attempt. Both of these points are important, and the challenges they address were certainly encountered during this research. Attempts to map the very simple Cottam and Hungate datasets even to the CRM-EH 'by hand' for someone unfamiliar with mapping to an ontology, was a frustrating experience, primarily as it was difficult to understand where the CIDOC-CRM ended and the CRM-EH began. As a result, one of the real strengths of the STELLAR templates was being able to see one's own data aligned to the ontology correctly. It is as much a teaching tool for learning what properly mapped data should look like, and how it all works together once it is in the RDF store, as it is a mapping and RDF conversion tool. To be able to see your own data mapped in an easily understandable way is very powerful.

During the creation of the CRM-EH, anything within the general Cultural Heritage domain was mapped to the CIDOC-CRM directly, and anything specific to archaeology then became part of the CRM-EH. By creating the CRM-EH ontology as an extension of the CIDOC-CRM for the archaeology domain, the STAR project chose to go down the route of maximum interoperability, therefore can it be considered a 'local ontology' as described by Martinez and Isaksen? Does the CRM-EH provide fine enough granularity to model most archaeological datasets to a level where the connections between the data are sufficiently meaningful to answer real research questions? If not, can the CRM-EH at least serve as a bridge for use between even more specialised 'local' sub-domain

ontologies and the CIDOC-CRM? Looking at the ontology developed by Isaksen *et al*. for the Port Networks Project, (which must have informed their views about generic versus local mapping), it was made specifically for Roman marble and amphorae finds, and mapped to a local ontology. In addition to universal classification concepts coming from SKOS and GeoNames, an ontology made up of two different namespaces (called archvocab and heml) was created with the instance data concepts modelled specifically for the Port Networks Project (Isaksen *et al.* 2009b, 4), including:

Classes	**Properties**
Excavation	inExcavation
Context	inContext
Find	locationRef
	ofForm
	ofMaterial
	ofType
	TerminusAnteQuem
	TerminusPostQuem
	hasQuantity

Looking at these concepts, it is possible to see the beginnings of a potential mapping to the CRM-EH using the format of the STELLAR templates:

Class	**STELLAR Template**
Excavation	CRMEH_INVESTIGATION_PROJECTS
Property	
inExcavation	investigation_id

Class	**STELLAR Template**
Context	CRMEH_CONTEXTS
Property	
inContext	within_excavation_id
	context_id

Class	**STELLAR Template**
Find	CRMEH_FINDS
Properties	
	Within_context_id
	find_id
locationRef	no current equivalent, but find_location could be created
ofForm	no current equivalent, but find_form could be created
ofMaterial	find_material
ofType	find_type
TerminusAnteQuem	no current equivalent, but could be production_period
TerminusPostQuem	no current equivalent, but could be production_period
hasQuantity	no current equivalent, but find_quantity could be created

Both the CRMEH_INVESTIGATION_PROJECTS and CRMEH_CONTEXTS templates within STELLAR include a location field, so it would likely be a simple matter to add a find_location field to the CRMEH_FINDS template. Martinez and Isaksen cite the need for the properties ofForm, ofMaterial and ofType as being a function of their particular data, where the generic 'type' designation is not sufficient. They discuss how ceramics specialists wish to account for the shape of a find and prefer ofForm, as a separate designation from the more generic ofType property. For dating, the Ports Network Project uses the upper and lower temporal limits of TerminusPostQuem and TerminusAnteQuem, whereas the STELLAR finds template uses the single designation of production_period. In addition, the STELLAR finds template has no equivalent for hasQuantity, which can hold a variety of measurement values, but the CRMEH_SAMPLE_MEASUREMENTS could probably be adapted for use with the finds template, as it includes measurement_type, measurement_unit, and measurement_value, which would make information about the quantity more machine readable.

These differences all seem quite minor, and with further expansion of the STELLAR templates, the ontology created for the Port Networks Project could therefore be mapped easily to the CRM-EH and therefore to the CIDOC-CRM. As the CRM-EH has sufficient scope to allow the STELLAR expansions, it can be argued that the CRM-EH could be used as a local ontology for the Ports Networks Project, and that STELLAR could allow it to be done quickly and easily

by non-specialist archaeologists. If this is the case, then it could also be argued that tools like STELLAR can adequately allow expression of the 'deep' nature of the data described by Martinez and Isaksen. This is just one example, but only through testing with more local domain ontologies and experimentation with a wide variety of datasets, will it be possible to fully determine the strengths and limitations of tools like STELLAR.

The Port Network Project is also based on the individual 'context' used in the single context recording tradition, but as discussed in the fourth chapter of this thesis, projects using other field recording traditions would not be compatible with the CRM-EH in its current configuration. Again, whether the CRM-EH could or should be adapted for use with other recording systems is a matter for debate, but at some point archaeologists using different types of field recording traditions will likely wish to have some level of interoperability with each other. It will also be important to ensure legacy datasets can be accommodated; single context recording only having been developed in the latter part of the 20th century.

In order to fully express the 'deep' nature of archaeological field recording, and in particular, the data derived from field drawing, further expansion of the STELLAR templates will be necessary in this direction as well. In its current configuration, it is only possible to express the spatial coordinates for a single x,y centroid for an area where some sort of fieldwork has been carried out (investigation_location), a context (context_location) or a group of contexts (group_location). As shown in the Port Networks Project, a find_location field should be included, but there is also scope within the CRM-EH to include polygon data. If the STELLAR templates could be expanded to incorporate groups of x,y coordinates which form polygons, then the extent of fieldwork locations could be defined, along with the extent of each context. This information was included in the data exported from the Cottam and Hungate drawings, along with area and perimeter data for each context, and is included in the CSV files ready for import into STELLAR. If STELLAR were expanded to include this additional geospatial data, it could be easily incorporated as well.

As it currently stands, STELLAR does a good job of handling stratigraphic information; with the ability to create relationships in one direction and then to automatically infer the reverse. By adding polygon data, STELLAR would go a long way to approximating the three dimensional nature of archaeological field recording as typically carried out. In other words, it would be possible to include two-dimensional plan data represented as x,y coordinates, along with the third dimension represented by stratigraphic relationships. Of course, this is not the same as true x,y,z three-dimensional point data, but with 3D laser scanners becoming more prevalent and affordable, the need to express RDF data in three dimensions may not be far away.

As discussed in the third chapter, field drawing in archaeology is about transformation, especially field drawing carried out as part of excavation. As each context is recorded and transformed by an individual, it is then removed and the visual record of that context becomes the primary archaeological data resource. Once the field drawing is collated with the other drawings to form the two-dimensional field record of the excavation, it goes through transformation again through further interpretation and distillation of understanding. This transformation is taking an increasingly varied path, with the majority of field drawing still being undertaken by hand drawing on *permatrace*, but then possibly undergoing digitisation through 'retrospective conversion' or in some instances being 'born digital'. As the history of field drawing shows, it is an integral part of the archaeological record and as such, must be included in any efforts to make archaeological information part of the Semantic Web, alongside the textual and photographic record.

The goal of the initial transformation of the visual and spatial information derived from an archaeological resource in Edward Tufte's 'three-space world' into the 'flatland' that is archaeological field drawing, is to preserve sufficient meaning to allow understanding. The goal of the subsequent transformation from whatever analogue or digital form of 'flatland' is used, into a graph data structure built

from RDF triples, is the same. The result is a complete abstraction of information that is meant to be understood visually, therefore is it still possible to express the meaning within the data? The comprehensive nature of single context recording (aligned to ontology) creates an archaeological record that can be pulled apart and put together without losing its stratigraphic relationships. Because individual contexts are self-contained units of information, which derive meaning from their stratigraphic connections with other contexts, their structure is much like the nodes and edges making up the graph data model used in RDF. It can therefore be argued that data derived from archaeological field drawings, at least within the tradition of single context recording, is translatable to the Semantic Web with sufficient meaning intact, and it is justifiable to use this technology with this type of data.

At the same time, the production of field drawings as the visual part of the archaeological record remains fundamental to understanding the resource, and somehow Semantic Web data that is meant to be visually understood will need to be accommodated. The graph data structure of RDF triples can be successfully visualised with generic tools like Gruff or bespoke GUIs like PELAGIOS, but the nature of the interface is a direct reflection of its structure. Map interfaces like those used in HESTIA and CLAROS are another good way to navigate through decentralised graph data within the Cultural Heritage domain, but have been limited to single x,y coordinate points at the site level. To represent visual archaeological data at the site level will require polygons to define the site, and for sites where excavation has been carried out, the contexts within it. This will require the advent of a new visual interface, preferably one that is generic and could be used for any archaeological data set created with single context recording. This is an area for further work to be explored upon completion of this thesis.

Of course, it will not only be important to visualise data at the site level, but to query it as well. Now that GeoSPARQL is well on its way to becoming an OGC

standard that will bring non-proprietary geospatial functionality to any SPARQL endpoint, it will be possible to make the types of useful spatial queries familiar to users of GIS. The results of these queries are meant to be returned as textual data, but ultimately it would be optimal to incorporate them into a visual interface. This is important not only because non-specialist users should not be expected to learn to write SPARQL, much less GeoSPARQL, but also because a query interface that displays results visually rather than textually allows truer expression of the visual and spatial nature of the data, if not in Tufte's 'three-space world' then at least in something approximating 'flatland'.

It will be fruitful to revisit the data from the Cottam and Hungate drawings as the technology moves on. When the Hungate excavation is complete, and all of the Anglo-Scandinavian contexts identified, it may then be possible to formulate and pose some real archaeological research questions that would have been difficult or impossible using data from relational databases with differing structures.
There were two main questions posed at the start of this research. First, whether the data from archaeological field drawings can be incorporated into Semantic Web data along with the textual. While this has been answered in the affirmative, much more needs to be done to fully realise its potential, for both visualisation and querying. Second, whether it is possible for non-specialist archaeologists to make use of the Semantic Web using free and generic tools. This has also been answered in the affirmative, and the means to do so is available right now. Some computing knowledge is needed, including a basic understanding of UNIX during the initial setup of the RDF store, along with a basic understanding of Semantic Web principles, but the majority of interaction can be through WYSIWYG interfaces. Thanks to STELLAR, at the very least, someone with a modicum of computing skill can set up the system, but once in place virtually anyone can add, maintain and work with the data with a minimal amount of training. This is a huge leap forward for making the Semantic Web accessible for everyday use by archaeologists.

The main conclusion of this research is that much more needs to be done to articulate and demonstrate what the Semantic Web is, how it works, and most importantly, how it is useful to archaeologists. The fact that an archaeologist with a very modest amount of computing knowledge could be walked through the workflow discussed in the fourth chapter of this thesis, see their own data modelled in RDF, aligned to an appropriate ontology, and displayed in a generic visualisation tool like Gruff in a matter or hours, should go a long way to dispelling the idea that there is a 'gap between the vision and the reality of the Semantic Web' as expressed by the participants in the Semantic Web Think Tank. So much has been done in the three years since the Think Tank met. Continued development of WYSIWYG interfaces by specialists that are not reliant on SPARQL queries to navigate the data, will be critical for demonstrating what the Semantic Web can do, including the development of generic interfaces which can be used across many types of archaeological datasets. Most importantly, specialists must work to articulate and demonstrate what new and different archaeological research questions can now be asked and potentially answered by using Semantic Web technology, and only then will it be possible to firmly place it into the archaeologist's toolkit. For those who have said the Semantic Web is dead, or should be slain, the kernel of the Semantic Web appears to be well planted within archaeology, and with more work and care, will continue to grow.

The contribution to knowledge and understanding made by the preceding thesis included an exploration of the Semantic Web with relation to archaeology, and whether it is yet possible for non-specialist archaeologists to create, use and share their data using Semantic Web technologies and principles. It also considered whether spatial data derived from field drawings can be incorporated alongside textual data, to ensure a more complete archaeological record is represented. To determine if these two related questions could be answered, a practical application was undertaken, followed by a discussion of the results and recommendations for future work. One of the primary tenets of the Semantic Web being interoperability of data, two archaeological sites were chosen because they were related

archaeologically, but differed technologically. Both datasets included field drawings from which data could be extracted, along with augmentory databases to enhance the demonstration. The data was carried through a complete workflow, from extraction, alignment to an ontology, translation into RDF, querying and visualisation within an RDF store, and through to publication as Linked Data.

This practical application was completed primarily using newly available generic tools, which required a minimal amount of specialist knowledge during most phases of the process. It demonstrated it is currently possible for non-specialist archaeologists to work with their data using Semantic Web technologies, including some data derived from field drawings, and showed how the Semantic Web may allow archaeologists to use their data in new ways, and that it is a fruitful area for further work.

Appendix A

List of files on CD

WrightHE_thesis_2011.pdf

STELLAR

STELLAR_preloader.jar
STELLAR_mapping.doc
STELLAR_command.rtf

Primary_data_files

Primary_data_Cottam

95.dwg
cont_with.txt
context_relationship_table.gif
contexts.txt
earlier_than.txt
later_than.txt

Primary_data_Hungate

hungate_area_h2_contexts.csv
hungate_area_h2_deep_trench_contexts.csv
hungate_area_h2_deep_trench_strat.csv
hungate_area_h2_deep_trench.dxf
hungate_area_h2_strat.csv
hungate_area_h2.dxf

CAD_files

CAD_Cottam

cottam_process.bak
cottam_process.dwg
cottam.bak
cottam.dwg

CAD_Hungate

hungate_merged_georef.bak
hungate_merged_georef.dwg
hungate_merged.bak
hungate_merged.dwg
hungate_process.bak
hungate_process.dwg
hungate.bak
hungate.dwg

GIS_files

GIS_Cottam

cottam_area_wgs84.img
cottam_area_wgs84.img.xml
cottam_area_wgs84.rrd
cottam_farm_wgs84.img
cottam_farm_wgs84.img.xml
cottam_farm_wgs84.rrd
cottam_GIS_file_metadata.docx
cottam_GIS_project_metadata.docx
cottam_wgs84_web.dbf
cottam_wgs84_web.prj
cottam_wgs84_web.sbn
cottam_wgs84_web.sbx
cottam_wgs84_web.shp
cottam_wgs84_web.shp.xml
cottam_wgs84_web.shx
featur2_georef_wgs84.img
featur2_georef_wgs84.img.vat.dbf
featur2_georef_wgs84.img.xml
featur2_georef_wgs84.rrd
magno2_georef_wgs84.img
magno2_georef_wgs84.img.vat.dbf
magno2_georef_wgs84.img.rrd

SE

SE_road_wgs84.dbf
SE_road_wgs84.prj
SE_road_wgs84.sbn
SE_road_wgs84.sbx
SE_road_wgs84.shp
SE_road_wgs84.shp.xml
SE_road_wgs84.shx

GIS_Hungate

google_satellite_wgs84.img
google_satellite_wgs84.img.xml
google_satellite_wgs84.rrd
hungate_GIS_file_metadata.docx
hungate_GIS_project_metadata.docx
hungate_wgs84_web.dbf
hungate_wgs84_web.prj
hungate_wgs84_web.sbn
hungate_wgs84_web.sbx
hungate_wgs84_web.shp
hungate_wgs84_web.shp.xml
hungate_wgs84_web.shx

SE

SE_road_wgs84.dbf
SE_road_wgs84.prj
E_road_wgs84.sbn
SE_road_wgs84.sbx
SE_road_wgs84.shp
SE_road_wgs84.shp.xml
SE_road_wgs84.shx

CSV_files

- **CSV_Cottam**
  - *contemp_with.csv*
  - *context_note.csv*
  - *contexts.csv*
  - *cottam_draw.csv*
  - *earlier_than.csv*
  - *investigation_projects_cottam.csv*
- **CSV_Hungate**
  - *hungate_contexts.csv*
  - *hungate_draw.csv*
  - *hungate_strat.csv*
  - *investigation_projects_hungate.csv*

GML_files

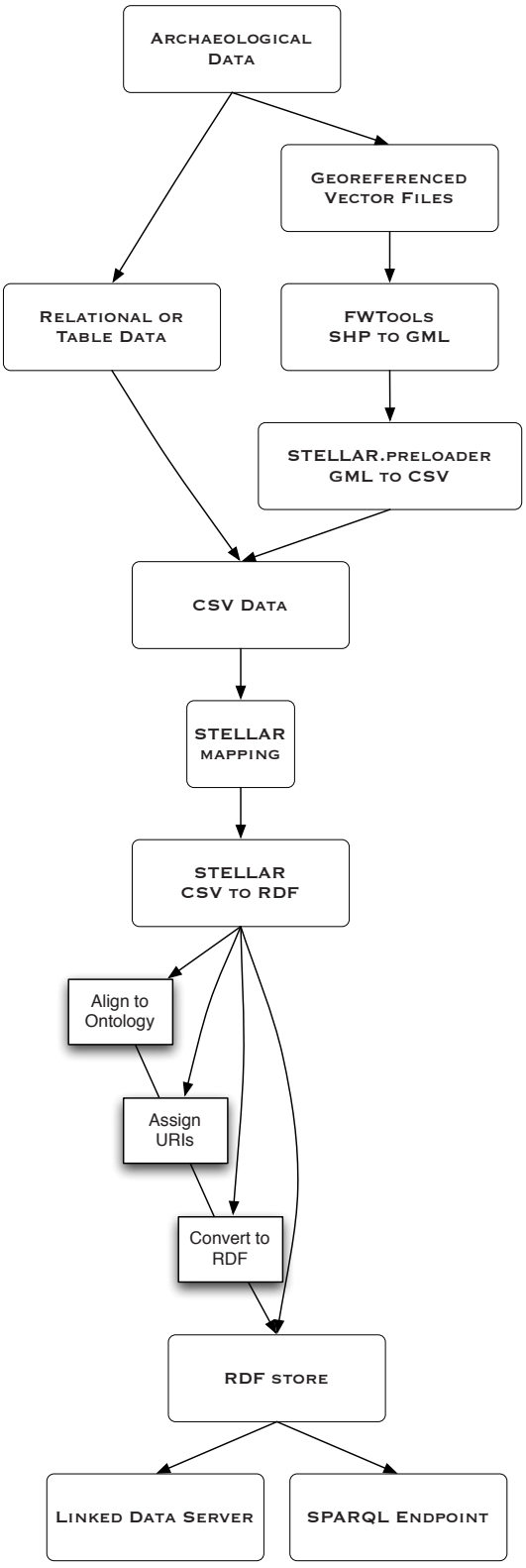
- *cottam.gml*
- *cottam.xsd*
- *hungate.gml*
- *hungate.xsd*

RDF_files

- **RDF_Cottam**
  - *contemp_with.rdf*
  - *context_note.rdf*
  - *contexts.rdf*
  - *cottam_draw.rdf*
  - *earlier_than.rdf*
  - *investigation_projects_cottam.rdf*
- **RDF_Hungate**
  - *hungate_contexts.rdf*
  - *hungate_draw.rdf*
  - *hungate_strat.rdf*
  - *investigation_projects_hungate.rdf*
  - **Pubby_data**
    - *hungate_contexts_LD.rdf*
    - *hungate_draw_LD.rdf*
    - *hungate_strat_LD.rdf*
    - *investigation_projects_hungate_LD.rdf*

Appendix B

Thesis Workflow



Appendix C

Selected Glossary of Acronyms

ADS: Archaeology Data Service

AHRC: Arts and Humanities Research Council

AI: Artificial Intelligence

ASCII: American Standard Code for Information Interchange

CAA: Computer Applications (and Quantitative Methods) in Archaeology

CAD: Computer Aided Design

CERN: *Organisation européenne pour la recherche nucléaire*

CfA: Centre for Archaeology

CIDOC-CRM: *le Comité International pour la DOCumentation de l'ICOM* (International Council of Museums)-Conceptual Reference Model

CLAROS: CLassical Art Research Online Research Services

CRM: Conceptual Reference Model

CRM-EH: Conceptual Reference Model-English Heritage

CSS: Cascading Style Sheets

CSV: Comma Separated Values

CWE: Collaborative Working Environment

DAML: DARPA (Defense Advanced Research Projects Agency) Agent Markup Language

DC: Dublin Core

DOI: Digital Object Identifier

DSWG: Documentation Standards Working Group

DTD: Document Type Definition

DWG: 'Drawing' file format

DXF: Drawing eXchange Format

EDM: Electronic Distance Meter

ERM: Entity Relationship Model

ECRM: Erlangen Conceptual Reference Model

FISH: Forum on Information in Heritage Standards

FOAF: Friend Of A Friend

FOL: First Order Logic

FRBR: Functional Requirements for Bibliographic Record

GDAL: Geospatial Data Abstraction Library

GIF: Graphics Interchange Format

GI: Geographic Information

GIS: Geographic Information System

GML: Geography Markup Language *or* Generalized Markup Language

GPS: Global Positioning System

GS: Geospatial Semantics

GUI: Graphical User Interface

HEIRNET: Historic Environment Information Resources Network

HESTIA: Herodotus Encoded Space-Text-Imaging Archive

HTML: HyperText Markup Language

IADB: Integrated Archaeological DataBase

INSPIRE: Infrastructure for Spatial Information in Europe

ISO: International Organization for Standardization

ISWC: International Semantic Web Conference

IW: Inference Web

JPG (JPEG): Joint Photographic Experts Group file format

JSON: JavaScript Object Notation

KOS: Knowledge Organization System

KML: Keyhole Markup Language

LRC: Landscape Research Centre

MIDAS: Monument Inventory Data Standard

MCFE: Mobile Computing in a Fieldwork Environment

N3: Notation3 Resource Description Framework serialisation format

NLP: Natural Language Processing

NS: NameSpaces

OGC: Open Geospatial Consortium

OIL: European Ontology Inference Layer

OS: Ordnance Survey

OWL: Web Ontology Language

PELAGIOS: Pelagios: Enable Linked Ancient Geodata In Open Systems

PML: Proof Markup Language

PUNS: Publication of Archaeological Projects: a user needs survey

RDF: Resource Description Framework

RDF/XML: Resource Description Framework/eXtensible Markup Language serialisation format

RDFS: Resource Description Framework Schema

SEA: Semantic Explorer for Archaeology

SGML: Standard Generalized Markup Language

SKOS: Simple Knowledge Organization System

SPARQL: SPARQL Protocol and RDF Query Language

SII: Spatial Information Infrastrcutures

SQL: Structured Query Language

STAR: Semantic Technologies for Archaeological Resources

STELLAR: Semantic Technologies Enhancing Links and Linked data for Archaeological Resources

SVG: Scalable Vector Graphics

SUAT: Scottish Urban Archaeology Trust

TIF (TIFF): Tagged Image File Format

TST: Total Station Theodolite

TXT: Text file format

URI: Uniform Resource Identifier

URL: Uniform Resource Locator

VASLE: Viking and Anglo-Saxon Landscape and Economy

VAST: Virtual Reality, Archaeology and Cultural Heritage

VERA: Virtual Environments for Research in Archaeology

VRE: Virtual Research Environment

W3C: World Wide Web Consortium

WKT: Well-Known Text

WYSIWYG: What You See Is What You Get

XHTML: eXtensible HyperText Markup Language

XML: Extensible Markup Language

YAT: York Archaeological Trust

Bibliography

Aasman, J (2011) 'Will Triple Stores Replace Relational Databases?', http://www.information-management.com/newsletters/database_metadata_unstructured_data_triple_store-10020158-1.html. Page consulted 16 May 2011.

Adkins, L, and Adkins, R A (1989) *Archaeological Illustration.* Cambridge: Cambridge University Press.

Alesso, H P, and Smith, C F (2004) *Developing Semantic Web Services.* Wellesley: A K Peters, Ltd.

Allemang, D, and Hendler, J (2008) *Semantic Web for the Working Ontologist.* Boston: Morgan Kaufmann Publishers.

Allen, S (2009) Rocking the Boat. *Yorkshire Archaeology Today***,** 9-11.

Annoni, A, Friis-Christensen, A, Lucchi, R, and Lutz, M (2008) Requirements and Challenges for Building a European Spatial Information Infrastructure: INSPIRE. IN van Oosterom, P, and Zlatanova, S (eds) *Creating Spatial Information Infrastructures: Towards a Spatial Semantic Web.* Boca Raton: Taylor & Francis Group, LLC.

Antoniou, G, Bikakis, A, Dimaresis, N, Genetzakis, M, Georgalis, G, Governatori, G, Karouzaki, E, Kazepis, N, Kosmdakis, D, Kritsotakis, M, Lilis, G, Papadogiannakis, A, Pediaditis, P, Terzakis, C, Theodosaki, R, and Zeginis, D (2008) Proof explanation for a nonmonotonic Semantic Web rules language. *Data and Knowledge Engineering,* 64**,** 662-687.

Antoniou, G, and van Harmelen, F (2004) *A Semantic Web Primer.* Cambridge: The MIT Press.

Apple Inc. (2010) 'Discovering Ancient Pompeii with iPad', http://classics.uc.edu/pompeii/images/stories/ipad/Apple%20-%20Discovering%20ancient%20Pompeii%20with%20iPad.pdf. Page consulted 20 June 2011.

Archaeology Data Service (2011a) 'Archaeology Data Service', http://archaeologydataservice.ac.uk/. Page consulted 12 April 2011.

Archaeology Data Service (2011b) 'What is the ADS, and what do we do?', http://archaeologydataservice.ac.uk/about/background. Page consulted 20 June 2011.

ArcTron (2007) 'ArcTron Website', http://www.arctron.com/. Page consulted 10 October 2010.

Artz, D, and Gil, Y (2007) A survey of trust in computer science and the Semantic Web. *Journal of Web Semantics,* 5**,** 58-71.

Ashish, N, and Sheth, A (eds) (2011) *Geospatial Semantics and the Semantic Web: Foundations, Algorithms, and Applications.* New York: Springer.

Austin, T, Brown, A, Brown, D, Coles, G, Dawson, D, Dodd, A, Fernie, K, Gardiner, J, Kilbride, W, Lock, G, Longworth, C, Merriman, N, Miller, P, Murray, D, Reeve, J, and Wise, A (2000) *Digital Archives from Excavation and Fieldwork: Guide to Good Practice.* York: Archaeology Data Service.

Backhouse, P (2006) Drowning in Data? IN Evans, T L, and Daly, P (eds) *Digital Archaeology: Bridging Method and Theory.* London: Routledge.

Bateman, J (2006) Pictures, Ideas and Things: The Production and Currency of Archaeological Images. IN Edgeworth, M (ed) *Ethnographies of Archaeological Practice.* Lanham: AltaMira Press.

Battenfeld, I, Beckmann, I, Schultze, J, and Türk, H (2009) Unifying Archaeological Databases using Triples. *Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology (COINFO).* Beijing, IEEE.

Battle, R, and Kolas, D (In Press) GeoSPARQL: Enabling a Geospatial Semantic Web. *Semantic Web Journal*.

Beauregard, B, Das, S, Perry, M, Rieb, K, and Sundara, S (2011) 'Oracle Database Semantic Technologies: Understanding How to Install, Load, Query and Inference', http://download.oracle.com/otndocs/products/semantic_tech/pdf/semtech_howto.pdf. Page consulted 11 July 2011.

Bell, T, Larcombe, A, and Veter, Y (2005) 'MIDAS XML and the Forum for Information Standards Toolkit', http://www.heritage-standards.org.uk/files/midas_caa2005.ppt. Page conusulted 12 January 2012.

Berners-Lee, T (2000) *Weaving the Web.* London: Texere.

Berners-Lee, T (2004) Forward. IN Fensel, D, Hendler, J, Lieberman, H, and Wahlster, W (eds) *Spinning the Semantic Web.* Cambridge: MIT Press.

Berners-Lee, T (2009a) 'The Semantic Web as a language of logic', http://www.w3.org/DesignIssues/Logic.html. Page consulted 8 March 2011.

Berners-Lee, T (2009b) 'Tim Berners-Lee on the next Web', http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html. Page consulted 15 April 2011.

Berners-Lee, T, Hendler, J, and Lassila, O (2001) The Semantic Web. *Scientific American.*

Binding, C (2011) 'STELLAR mapping and extraction guidelines', http://reswin1.isd.glam.ac.uk/stellar/STELLAR.Introduction.pdf. Page consulted 22 March 2011.

Bizer, C (2010) 'The D2RQ Plattform - Treating Non-RDF Databases as Virtual RDF Graphs', http://www4.wiwiss.fu-berlin.de/bizer/d2rq/. Page consulted 12 May 2011.

Bizer, C, and Cyganiak, R (2010) 'D2R Server: Publishing Relational Databases on the Semantic Web', http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/#about. Page consulted 2 July 2011.

Bizer, C, Heath, T, and Berners-Lee, T (2008) Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems,* 5**,** 1-22.

Blomerus, P, and Eiteljorg II, H (2009) 'Managing the Content of AutoCAD Models with Layers', http://csanet.org/newsletter/fall09/nlf0901.html. Page consulted 26 October 2010.

Boucher de Perthes, J (1864) *Antiquités Celtiques et Antédiluviennes: Mémoir sur L'industrie Primitive et les Arts a Leur Origine.* Paris: Jung-Treuttel.

Bowden, M (1991) *Pitt Rivers: The life and archaeological work of Lieutenant-General Augustus Henry Lane Fox Pitt Rivers.* Cambridge: Cambridge University Press.

Brachman, R J, and Levesque, H J (2004) *Knowledge Representation and Reasoning.* San Francisco: Morgan Kaufmann Publishers.

Brewer, J, and Kilbride, W (2006) 'HEIRNET User Survey 2005 Report', http://www.britarch.ac.uk/HEIRNET/survey/section1.htm. Page consulted 27 March 2011.

Byrne, K (2010) *Populating the Semantic Web: Combining Text and Relational Databases as RDF.* Saarbrücken: Lambert Academic Publishing.

CAA (2010) 'About CAA...', http://www.leidenuniv.nl/caa/about_caa.htm. Page consulted 15 October 2010.

Carver, M O H (2009) *Archaeological Investigation.* Oxford: Routledge.

Castro, E (2007) *HTML, XHTML & CSS.* Berkeley: Peachpit Press.

Chenhall, R G (1968) The impact of computers on archaeological theory: An appraisal and projection. *Computers and the Humanties,* 3**,** 15-24.

CIDOC CRM (2010) 'The CIDOC Conceptual Reference Model', http://www.cidoc-crm.org/. Page consulted 20 June 2011.

CIDOC CRM (2011) 'What is the CIDOC CRM', http://www.cidoc-crm.org/index.html. Page consulted 4 March 2011.

Clark, J T, and Hagemeister, E M (eds) (2007) *Digital Discovery: Exploring New Frontiers in Human Heritage.* Budapest: ARCHAEOLINGUA.

Clark, P R (1993) Sites without Principles; post-excavation analysis of 'pre-matrix' sItes. IN Harris, E, Brown III, M R, and Brown, G J (eds) *Practices of Archaeological Stratigraphy.* London: Academic Press Limited.

Clarke, A, Fulford, M, and Rains, M (2002) Nothing to Hide - Online Database Publication and the Silchester Town Life Project. in Doerr, M, and Sarris, A (eds) *The 29th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA).* Heraklion, Hellenic Ministry of Culture.

Clarke, A, and O'Riordan, E (2009) Managing Change: Introducing innovation into well-established systems. in Frischer, B, Webb Crawford, J, and Koller, D (eds) *Computer Applications in Archaeology.* Williamsburg, USA, Archaeopress.

CLAROS (2011) 'About CLAROS', http://explore.clarosnet.org/XDB/ASP/clarosHome/about.html. Page consulted 25 July 2011.

Coates, T (2010) 'Everything the Network Touches', http://dconstruct.s3.amazonaws.com/2010/podcast/dconstruct2010-coates.mp3. Audio accessed 22 July 2011.

Collis, J (2001) *Digging Up the Past: An Introduction to Archaeological Investigation.* Thrupp, Stroud, Gloucestershire: Sutton.

Connelly, P (2007) Great Expectations for the Hungate Excavation. *Yorkshire Archaeology Today***,** 1-4.

Connelly, P (2008) Hungate Excavations: Season 2 Draws To An End. *Yorkshire Archaeology Today***,** 1-3.

Connelly, P (2010) Three Is The Magic Number! An Update On The Hungate Excavations. *Yorkshire Archeology Today***,** 1-4.

Connelly, P (2011) Hungate 2011: The Final Year! *Yorkshire Archaeology Today***,** 1-6.

Cowgill, G L (1967) Computer Applications in Archaeology. *AFIPS Joint Computer Conferences.* Anaheim, California, USA, Association for Computer Machinery.

Cripps, P, Greenhalgh, A, Fellows, D, May, K, and Robinson, D (2004) *Ontological Modelling of the work of the Centre for Archaeology*. English Heritage.

Cripps, P, and May, K (2010) To OO or not to OO? Revelations from Ontological Modelling of an Archaeological Information System. in Niccolucci, F, and herman, S (eds) *Computer Applications and Quantitative Methods in Archaeology (CAA) 2004.* Prato, Archaeolingua.

Crofts, N, Doerr, M, Gill, T, Stead, S, and Stiff, M (eds) (2010) *Definition of the CIDOC Conceptual Reference Model 5.0.2.* ICOM/CIDOC Documentation Standards Group, CIDOC CRM Special Interest Group.

Cyganiak, R, and Bizer, C (2010) 'Pubby: A Linked Data Frontend for SPARQL Endpoints', http://www4.wiwiss.fu-berlin.de/pubby/. Page consulted 2 May 2011.

D'Andrea, A (2008) CIDOC-CRM in Data Management and Data Sharing. *The 36th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA).* Budapest, In Press.

Daconta, M C, Obrst, L J, and Smith, K T (2003) *The Semantic Web: A guide to the Future of XML, Web Services, and Knowledge Management.* Indianapolis: Wiley Publishing, Inc.

Davies, J, Fensel, D, and van Harmelen, F (eds) (2003) *Towards The Semantic Web: Ontology-Driven Knowledge Management.* Chichester: John Wiley & Sons Ltd.

Dolbear, C, and Hart, G (2008) Opportunities and Challenges in Exploiting Semantics as an Aid to Information Integration: A National Mapping Agency Perspective. IN van Oosterom, P, and Zlatanova, S (eds) *Creating Spatial Information Infrastructures: Towards a Spatial Semantic Web.* Boca Raton: Taylor & Francis Group, LLC.

DuCharme, B (2011) *Learning SPARQL.* Sebastopol: O'Reilly.

ECRM (2011) ‘Erlangen CRM/OWL’, http://erlangen-crm.org/about. Page consulted 2 July 2011.

Eiteljorg II, H, Fernie, K, Huggett, J, and Robinson, D (2002) ‘CAD: A Guide to Good Practice’, http://ads.ahds.ac.uk/project/goodguides/cad. Page consulted 28 June 2011.

Eiteljorg II, H, and Limp, W F (2008) ‘Archaeological Computing’. Second ed. Bryn Mawr, PA Center for the Study of Architecture.

Ellis, S, and Wallrodt, J (2010) ‘iPads at Pompeii’, http://classics.uc.edu/pompeii/index.php/news/1-latest/142-ipads2010.html. Page consulted 30 June 2011.

English Heritage (2007a) *3D Laser Scanning for Heritage: Advice and guidance to users on laser scanning in archaeology and architecture.* Swindon: English Heritage Publishing.

English Heritage (2007b) *Understanding the Archaeology of Landscapes: A guide to good recording practice.* Swindon: English Heritage Publishing.

Feigenbaum, L, Herman, I, Hongsermeier, T, Neumann, E, and Stephens, S (2007) The Semantic Web in Action. *Scientific American,* 297**,** 90-97.

Feigenbaum, L, and Prud’hommeaux, E (2011) ‘SPARQL By Example’, http://www.cambridgesemantics.com/2008/09/sparql-by-example/. Page consulted 20 July 2011.

Fensel, D, Hendler, J, Lieberman, H, and Wahlster, W (eds) (2003) *Spinning the Semantic Web.* Cambridge: MIT Press.

Fenton-Thomas, C (2005) *The Forgotten Landscapes of the Yorkshire Wolds.* Stroud, Gloucestershire: Tempus Publishing Limited.

Fiadeiro, J, Tuosto, E, Law, E, Solanki, M, and Hong, Y (2009) ‘Collaborative Working Environment (CWE) and Ontology’, http://www.tracingnetworks.org/content/web/collaborative_system.jsp. Page consulted 22 June 2011.

Fischer, L, Frischer, B, and Wells, S (eds) (2009) *Making History Interactive: Abstracts of The 37th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA).* Williamsburg: CAA2009 Williamsburg.

FISH (2011) 'FISH Interoperability Toolkit', http://www.heritage-standards.org.uk/. Page consulted 15 July 2011.

Fisher, C R, Terras, M, and Warwick, C (2009) Integrating New Technologies into Established Systems: a case study from Roman Silchester. in Frischer, B, Webb Crawford, J, and Koller, D (eds) *Computer Applications and Quantitative Methods in Archaeology.* Williamsburg, Archaeopress.

Franz Inc. (2011a) 'Geospatial support in SPARQL queries', http://www.franz.com/agraph/support/documentation/current/sparql-geo.html. Page consulted 2 July 2011.

Franz Inc. (2011b) 'AllegroGraph 4.3 Introduction', http://www.franz.com/agraph/support/documentation/current/agraph-introduction.html#header2-7. Page consulted 11 July 2011.

Franz Inc. (2011c) 'AllegroGraph 4.3 Web View', http://www.franz.com/agraph/support/documentation/v4/agwebview.html. Page consulted 2 July 2011.

Franz Inc. (2011d) 'SPARQL API Reference', http://www.franz.com/agraph/support/documentation/current/sparql-reference.html. Page consulted 20 July 2011.

Gaffney, V, and Exon, S (1999) 'From Order to Chaos: Publication, Synthesis and the Dissemination of Data in a Digital Age', http://intarch.ac.uk/journal/issue6/gaffney/index.html. Page consulted 30 May 2011.

Gaines, S W (1974) Computer Use at an Archaeological Field Location. *American Antiquity,* 39**,** 454-462.

Gamble, C (1997) General Editor's Preface. IN Molyneaux, B L (ed) *The Cultural Life of Images: Visual Representation in Archaeology.* London: Routledge.

Garshol, L M (2004) Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all. *Journal of Information Science,* 30**,** 378-391.

GeoNames (2011a) 'GeoNames', http://www.geonames.org/. Page consulted 30 June 2011.

GeoNames (2011b) 'About GeoNames', http://www.geonames.org/about.html. Page consulted 2 April 2011.

Geroimenko, V (2004) *Dictionary of XML Technologies and the Semantic Web.* London: Springer-Verlag London Limited.

Gil, Y, and Artz, D (2007) Towards content trust of web resources. *Journal of Web Semantics,* 5**,** 227-239.

Goddard, S (2000) The importance of illustration in archaeology and exemplary work of Robert Gurd. *Sussex Archaeological Collections,* 138**,** 7-13.

Golbeck, J (2008) Weaving a Web of Trust. *Science,* 321**,** 1640-1.

Goldfarb, C (1996) 'The Roots of SGML -- A Personal Recollection', http://www.sgmlsource.com/history/roots.htm. Page consulted 20 February 2011.

Google Maps (2011) 'Google Maps Javascript API V3 Map Types', http://code.google.com/apis/maps/documentation/javascript/maptypes.html#Projections. Page consulted 25 June 2011.

Grabau, A M (1960) *Priciples in Geology.* New York: Dover Publications.

Hall, A R (2004) A Historiographical Introduction to Anglo-Scandinavian York. *Aspects of Anglo-Scandinavian York.* York: Council for British Archaeology.

Hall, R (2007) Anglo-Saxon & Viking Age York. IN Nuttgens, P (ed) *The History of York.* Pickering: Blackthorn Press.

Hall, R A, Rollason, D W, Blackburn, M, Parsons, D N, Fellows-Jensen, G, Hall, A R, Kenward, H K, O'Connor, T P, Tweddle, D, Mainman, A J, and Rogers, N S H (2004) *Aspects of Anglo-Scandinavian York.* York: Council for British Archaeology.

Harold, E R, and Means, W S (2002) *XML in a Nutshell: A Desktop Quick Reference.* Sebastopol: O'Reilly & Associates, Inc.

Harris, E C (1989) *Principles of Archaeological Stratigraphy.* London: Academic Press Limited.

Harris, E C, Brown III, M R, and Brown, G J (eds) (1993) *Practices of Archaeological Stratigraphy.* London: Academic Press Limited.

Hart, G (2009) 'Linking to the past, geographically speaking: The Linked Data Web and Historical GIS', http://www.ordnancesurvey.co.uk/oswebsite/partnerships/research/publications/docs/2009/Linking_to_the_Past_GeoS.pdf. Page consulted 25 February 2011.

Hartig, O (2008) Trustworthiness of Data on the Web. *STI Berlin & CSW PhD Workshop.* Berlin, Germany, In Prep.

Hawker, J M (2001) *A Manual of Archaeological Field Drawing.* Hertford: RESCUE - The British Archaeological Trust.

Heath, T (2009) 'Linked Data? Web of Data? Semantic Web? WTF?', http://tomheath.com/blog/2009/03/linked-data-web-of-data-semantic-web-wtf/. Page consulted 16 December 2010.

Hebeler, J, Fisher, M, Blace, R, and Perez-Lopez, A (2008) *Semantic Web Programming.* Indianapolis: Wiley Publishing, Inc.

Hendler, J (2009) 'My Take on the Semantic Web Layercake', http://www.cs.rpi.edu/~hendler/presentations/LayercakeDagstuhl-share.pdf. Page consulted 8 January 2011.

HESTIA (2010) 'Project HESTIA: the Herodotus Encoded Space-Text-Imaging Archive', http://www.open.ac.uk/Arts/hestia/index.html. Page consulted 22 February 2011.

Hitzler, P, Krötzsch, M, and Rudolph, S (2010) *Foundations of Semantic Web Technologies.* Boca Raton: Taylor and Francis Group, LLC.

Hong, Y, and Solanki, M (2010) A Framework for Transforming Archaeological Databases to Ontological Datasets. IN Melero, F J, Cano, P, and Revelles, J (eds) *Fusion of Cultures: Abstracts of The 38th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA).* Granada: CAA2010 Granada.

Hope-Taylor, B (1966) Archaeological Draughtsmanship: Principles and Practice Part II: Ends and Means. *Antiquity,* 40**,** 107-113.

Hope-Taylor, B (1967) Archaeological Draughtsmanship: Principles and Practice Part III: Lines of Communication. *Antiquity,* 41**,** 181-189.

Hopkinson, G, and Winters, J (2003) 'Problems with Permatrace: a note on digital image publication', http://intarch.ac.uk/journal/issue14/hopkinson_index.html. Page consulted 10 June 2011.

Huggett, J (2004) 'The Past in Bits: towards an archaeology of Information Technology? ', http://intarch.ac.uk/journal/issue15/huggett_index.html. Page consulted 25 July 2010.

Hunter-Mann, K (2009) The Vikings come to Hungate. *Yorkshire Archeology Today***,** 4-8.

Hyvönen, E, Tuominen, J, Kauppinen, T, and Väätäinen, J (2011) Representing and Utilizing Changing Historical Places as an Ontology Time Series. IN Ashish, N, and Sheth, A (eds) *Geospatial Semantics and the Semantic Web: Foundations, Algorithms, and Applications.* New York: Springer.

IBM DevelopWorks (2011) 'New to XML', http://www.ibm.com/developerworks/xml/newto/. Page consulted 20 February 2011.

Imhof, E (2007) *Cartographic Relief Presentation.* Redlands: ESRI Press.

Inference Web (2011) 'Inference Web', http://inference-web.org/. Page consulted 8 March 2011.

INSPIRE (2010) 'INSPIRE Status Report', http://inspire.jrc.ec.europa.eu/documents/INSPIRE_/INSPIRE_status_report_Oct_2010.pdf. Page consulted 1 February 2012.

INSPIRE (2012) 'About INSPIRE', http://inspire.jrc.ec.europa.eu/index.cfm/pageid/48. Page consulted 1 February 2012.

Isaksen, L (2009) Linking Archaeological Data: A Framework for Academic Microprovision on the Semantic Web. *Mini-thesis submitted in continuation of a PhD to the School of Electronics & Computer Science.* Southampton, University of Southampton.

Isaksen, L, Martinez, K, and Earl, G (2009a) Archaeology, Formality & the CIDOC CRM. *Interconnected data worlds: Workshop on the implementation of CIDOC-CRM.* Berlin.

Isaksen, L, Martinez, K, and Earl, G (2011) Semantic Technologies in Cultural Heritage Past, Present and Future. *Cultural Heritage & the Semantic Web.* British Museum, London.

Isaksen, L, Martinez, K, Gibbins, N, Earl, G, and Keay, S (2009b) Linking Archaeological Data. in Frischer, B, Webb Crawford, J, and Koller, D (eds) *Making History Interactive: Abstracts of The 37th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA).* Williamsburg, Archaeopress.

Isaksen, L, Martinez, K, Gibbins, N, Earl, G, and Keay, S (2010) Interoperate with whom? Formality, Archaeology and the Semantic Web. *Web Science 2010.* Raleigh, North Carolina, Web Science Trust.

Janowicz, K (2011) 'Semantic Web Journal: Special Issue on Linked Spatiotemporal Data and Geo-Ontologies', http://www.semantic-web-journal.net/blog/special-issue-linked-spatiotemporal-data-and-geo-ontologies. Page consulted 20 June 2011.

Jena (2011) 'Jena – A Semantic Web Framework for Java', http://jena.sourceforge.net/. Page consulted 28 June 2011.

Jerem, E, Redő, F, and Szeverényi, V (eds) (2008) *On the Road to Reconstructing the Past: Abstracts of The 36th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA).* Budapest: Archaeolingua Foundation.

Jones, S, MacSween, A, Jeffrey, S, Morris, R, and Heyworth, M (2001) 'From The Ground Up: The Publication of Archaeological Projects: a user needs survey', http://www.britarch.ac.uk/pubs/puns. Page consulted 10 June 2011.

Kahney, L (1999) 'Programmer Reaches His Xanadu', http://www.wired.com/print/science/discoveries/news/1999/08/21430. Page consulted 17 February 2011.

Karmacharya, A, Cruz, C, Boochs, F, and Marzani, F (2009) ArcheoKM: Toward a Better Archaeological Spatial Datasets Management. *Computer Applications and Quatitative Methods in Archaeology (CAA).* Williamsburg.

Karmacharya, A, Cruz, C, Boochs, F, and Marzani, F (2010a) ArchaeoKM: Managing Archaeological data through Archaeological Knowledge. IN Melero, F J, Cano, P, and Revelles, J (eds) *Fusion of Cultures: Abstracts of The 38th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA).* Granada: CAA2010 Granada.

Karmacharya, A, Cruz, C, Boochs, F, and Marzani, F (2010b) Use of Geospatial Analyses for Semantic Reasoning. *Knowledge-Based and Intelligent Information and Engineering Systems.* Berlin: Springer.

Karmacharya, A, Cruz, C, Marzani, F, and Boochs, F (2008) Industrial Archaeology: Case study of Knowledge Management for Spatial Data of Findings. *2nd International Workshop on Personalized Access to Cultural Heritage.* Hannover.

Kendall, T (2007) Archaeology on Block H. *Yorkshire Archeology Today***,** 5-8.

Kendall, T (2009) Hungate: from HI to H2. *Yorkshire Archaeology Today***,** 1-3.

Kendall, T (2010) Jet and Glass and rocks 'n' Bones: Hungate Block H in 2010. *Yorkshire Archaeology Today*, 4-11.

Klee, P (1953) *Pedagogical Sketchbook.* London: Faber and Faber.

Lang, M (2009) ArcheoInf—Allocation of Archaeological Primary Data. *Computer Applications and Quantitative Methods in Archaeology (CAA).* Williamsburg.

Lang, M, and Türk, H (2010) Recent Developments in the ArcheoInf Project - Towards an Ontology of Archaeology. IN Melero, F J, Cano, P, and Revelles, J (eds) *Fusion of Cultures: Abstracts of The 38th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA).* Granada: CAA2010 Granada.

Laroche, J F (2010) 'The Pocket Computer Museum', http://pocket.free.fr. Page consulted 18 October 2010.

Li, Y, and Dailey, D (2011) 'A proposal for extending SVG's capabilities for online mapping and GIS', http://www.svgopen.org/2011/registration.php?section=abstracts_and_proceedings. Page consulted 25 July 2011.

Lieberman, J (2010) Greeting and Introduction Geosemantic Summit. *2010 OGC Geosemantics Summit.* Silver Spring.

Lieberman, J, Pehle, T, Morris, C, Kolas, D, Dean, M, Lutz, M, Probst, F, and Klien, E (2006) Geospatial Semantic Web Interoperability Experiment Report. in Lieberman, J (ed), Open Geospatial Consortium Inc.

Lock, G (2003) *Using Computers in Archaeology: Towards Virtual Pasts.* Oxford: Routledge.

Lock, G, and Brown, K (eds) (2000) *On the Theory and Practice of Archaeological Computing.* Oxford: Oxford University Committee for Archaeology.

LSTD (2010) 'LSTD 2010: Workshop On Linked Spatiotemporal Data 2010', http://stko.psu.edu/lstd2010/. Page consulted 25 June 2011.

Martinez, K, and Isaksen, L (2010) The semantic web approach to increasing access to cultural heritage. IN Bailey, B, and Gardiner, H (eds) *Revisualizing Visual Culture.* London: Ashgate.

May, K (2006) 'Integrating Cultural and Scientific Heritage: Archaeological Ontological Modelling for the Field and the Lab', http://cidoc.ics.forth.gr/workshops/heraklion_october_2006/may.pdf. Page consulted 26 June 2011.

May, K (2009) The Semantic Technologies for Archaeological Resources (STAR) project's use of SKOS. *ATHENA WP4 SKOS Workshop.* Rome.

May, K, Binding, C, and Tudhope, D (2008) A STAR is born: some emerging Semantic Technologies for Archaeological Resources. *On the Road to Reconstructing the Past: Computer Applications & Quantitative Methods in Archaeology (CAA)* Budapest, In Press.

May, K, Binding, C, and Tudhope, D (2010) Following a STAR? Shedding More Light on Semantic Technologies for Archaeological Resources. *Computer Applications and Quatitative Methods in Archaeology (CAA).* Williamsburg, Archaeopress.

May, K, and Cross, S (2004) Revelation: Practice, Technology, Dissemination and the Design of a Field Recording System. IN Frischer Ausserer, K, Börner, W, Goriany, M, and Karlhuber-Vöckl, L (eds) *Enter the Past: Proceedings of the 30th CAA conference held in Vienna, Austria.* Oxford: Archaeopress.

McCarthy, P (2005) 'Search RDF data with SPARQL', http://www.ibm.com/developerworks/xml/library/j-sparql/. Page consulted 25 June 2011.

McGuinness, D L, and Pinheiro da Silva, P (2004) 'Explaining Answers from the Semantic Web', http://www.ksl.stanford.edu/KSL_Abstracts/KSL-04-03.html. Page consulted 8 March 2011.

Melero, F J, Cano, P, and Revelles, J (eds) (2010) *Fusion of Cultures: Abstracts of The 38th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA).* Granada: CAA2010 Granada.

Meng, L (2008) Cartography and Visualization. IN Xiong, H, and Shekhar, S (eds) *The Encyclopedia of GIS.* New York: Springer.

MIDAS (2011) 'Documentation for MIDAS Spatial Schema', http://www.heritage-standards.org.uk/midas/docs/spatial/index.html?url=/midas/docs/spatial/spatial.html. Page consulted 12 July 2011.

Miller, P, and Richards, J D (1995) The good, the bad, and the downright misleading: archaeological adoption of computer visualization. IN Huggett, J, and Ryan, N (eds) *Computer Applications and Quantitative Methods in Archaeology 1994.* Oxford: TEMPVS REPARATVM.

Museum of London Archaeology Service (1994) *Archaeological Site Manual.* London: Museum of London.

Naughton, J (2000) *A Brief History of the Future: The origins of the Internet.* London: Pheonix.

Nelson, T (2009) *Geeks Bearing Gifts.* Mindful Press.

Nuttgens, P (2001) *York: The Continuing City.* London: Faber and Faber.

Ogbuji, U (2006) 'Thinking XML: The XML decade', http://www.ibm.com/developerworks/xml/library/x-think38/index.html. Page consulted 20 February 2011.

OnToText (2011) 'Geo-spatial indexing in OWLIM', http://www.ontotext.com/owlim/geo-spatial. Page consulted 12 May 2011.

Open Geospatial Consortium (2010) OpenGIS® Implementation Standard for Geographic information - Simple feature access. IN Herring, J (ed) http://www.opengeospatial.org/standards/sfa. Page consulted 1 August 2011.

Open Geospatial Consortium (2011a) 'GeoSPARQL SWG', http://www.opengeospatial.org/projects/groups/geosparqlswg. Page consulted 1 August 2011.

Open Geospatial Consortium (2011b) OGC GeoSPARQL - A Geographic Query Language for RDF Data. IN Perry, M, and Herring, J (eds) http://portal.opengeospatial.org/files/?artifact_id=44722. Page consulted 20 July 2011.

OpenRDF (2011) 'OpenRDF.org...home of Sesame', http://www.openrdf.org/. Page consulted 28 June 2011.

Ordnance Survey (2010) 'A guide to coordinate systems in Great Britain', http://www.ordnancesurvey.co.uk/oswebsite/gps/docs/A_Guide_to_Coordinate_Systems_in_Great_Britain.pdf. Page consulted 22 July 2011.

Ordnance Survey (2012) 'OS OpenData', http://www.ordnancesurvey.co.uk/oswebsite/products/os-opendata.html. Page consulted 1 February 2012.

Ottaway, P (2007) Roman York. IN Nuttgens, P (ed) *The History of York: From Earliest Times to the Year 2000.* Pickering: Blackthorn Press.

Passin, T B (2004) *Explorer's Guide to the Semantic Web.* Greenwich: Manning.

Payne, A (2011) 'Laser Scanning for Archaeology: A Guide to Good Practice', http://guides.archaeologydataservice.ac.uk/g2gp/LaserScan_Toc. Page consulted 20 April 2011.

Peacock, D (2005) 'Stone in Archaeology: Towards a digital resource', http://ads.ahds.ac.uk/catalogue/resources.html?stones_ahrb_2005. Page consulted 21 February 2011, doi:10.5284/1000246.

PELAGIOS (2011) 'Pelagios: Enable Linked Ancient Geodata In Open Systems', http://pelagios-project.blogspot.com/. Page consulted 15 August 2011.

Perry, M (2008) A Framework to Support Spatial, Temporal and Thematic Analytics over Semantic Web Data. PhD Thesis: Wright State University.

Perry, M and Herring, J (2011) 'GeoSPARQL SWG Charter', http://ontolog.cim3.net/forum/socop-forum/2010-05/docb2aMvfG39U.doc. Page consulted 20 May 2011.

Perry, M, Jain, P, and Sheth, A (2011) SPARQL-ST: Extending SPARQL to Support Spatiotemporal Queries. IN Ashish, N, and Sheth, A (eds) *Geospatial Semantics and the Semantic Web: Foundations, Algorithms, and Applications.* New York: Springer.

Piggott, S (1947-8) The excavations at Carinpapple Hill, West Lothian, 1947-1948. *Proceedings of the Society of Antiquaries of Scotland,* 82**,** 68-123.

Piggott, S (1965) Archaeological Draughtsmanship: Principles and Practice Part I: Principles and Retrospect. *Antiquity,* 39**,** 165-176.

Pilides, D, Hermon, S, Amico, N, Chamberlain, M, D'Andrea, A, Iannone, G, and Ronzino, P (2010) The Hill of Agios Georgios, Nicosia: 3D analysis of an on-going archaeological excavation. IN Melero, F J, Cano, P, and Revelles, J (eds) *Fusion of Cultures: Abstracts of The 38th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA).* Granada: CAA2010 Granada.

Pinheiro da Silva, P, McGuinness, D L, and Fikes, R (2006) A proof markup language for semantic web services. *Information Systems,* 31**,** 381-395.

Posluschny, A, Lambers, K, and Herzog, I (eds) (2008) *Layers of Perception.* Frankfurt: Römisch-Germanische Kommission des Deutschen Archäologischen Instituts.

Powers, S (2003) *Practical RDF.* Sebastapol: O'Reilly.

Powlesland, D (1986) On-Site Computing: In the Field with the Silicon Chip. IN Richards, J D (ed) *Computer Usage in British Archaeology.* Birmingham: The Institute of Field Archaeology.

Powlesland, D (1991) From the Trench to the Bookshelf: Computer Use at the Heslerton Parish Project. IN Ross, S, Moffett, J, and Henderson, J (eds) *Computing for Archaeologists.* Oxford: Oxford University Committee for Archaeology.

Powlesland, D, May, K, Rackham, J, and Tipper, J (2009) 'DigIT: Archaeological Summary Report and Experiments in Digital Recording in the Field', http://intarch.ac.uk/journal/issue27/powlesland_index.html. Page consulted 16 October 2010.

Rains, M (2007) Silchester – A Virtual Research Environment for Archaeology. in Posluschny, A, Lambers, K, and Herzog, I (eds) *The 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA).* Frankfurt, Römisch-Germanische Kommission des Deutschen Archäologischen Instituts.

Rains, M (2010) 'Integrated Archaeological Database', http://www.iadb.org.uk/. Page consulted 20 October 2010.

Rains, M (2011) 'Integrated Archaeological Database', http://www.iadb.org.uk/. Page consulted 20 May 2011.

Reilly, P (1991) Graphic Systems. IN Ross, S, Moffett, J, and Henderson, J (eds) *Computing for Archaeologists.* Oxford: Oxford University Committee for Archaeology.

Reilly, P, and Rahtz, S (1992) Introduction: archaeology and the information age. IN Reilly, P, and Rahtz, S (eds) *Archaeology and the Information Age: A global perspective.* London: Routledge.

Richards, J D (1993) 'York Environs Project, Cottam: An Anglian Site on the Yorkshire Wolds, Project Outline and Research Design', http://archaeologydataservice.ac.uk/archives/view/cottam_ba/html.cfm?CFID=30&CFTOKEN=A26216B6-CE90-4320-8FCB5D6504532107&. Page consulted 26 June, 2011.

Richards, J D (1998) Recent Trends in Computer Applications in Archaeology. *Journal of Archaeological Research,* 6**,** 331-382.

Richards, J D (1999) Cottam: An Anglian and Anglo-Scandinavian settlement on the Yorkshire Wolds. *The Archaeological Journal,* 156**,** 1-111.

Richards, J D (2000) *Viking Age England.* Stroud, Gloucestershire: Tempus Publishing Limited.

Richards, J D (2001a) 'Anglian and Anglo-Scandinavian Cottam: linking digital publication and archive', http://intarch.ac.uk/journal/issue10/richards/. Page consulted 25 January 2010.

Richards, J D (2001b) 'Burrow House Farm, Cottam: an Anglian and Anglo-Scandinavian Settlement in East Yorkshire', http://ads.ahds.ac.uk/catalogue/projArch/cottam_ba/overview.cfm. Page consulted 10 December 2010.

Richards, J D (2001c) 'Burrow House Farm, Cottam: an Anglian and Anglo-Scandinavian Settlement in East Yorkshire' [data-set]. York: Archaeology Data Service [distributor] (doi:10.5284/1000339).

Richards, J D, Ashby, S, Austin, T, Haldenby, D, Hummler, M, Jelley, E, Richardson, J, and Roskams, S (in prep) Cottam, Cowlam and environs: An Anglo-Saxon estate on the Yorkshire Wolds.

Richards, J D, Naylor, J, and Holas-Clark, C (2009) 'Anglo-Saxon Landscape and Economy: using portable antiquities to study Anglo-Saxon and Viking Age England', http://intarch.ac.uk/journal/issue25/richards. Page consulted 10 December 2010.

Richards, J D, and Ryan, N (1985) *Data Processing in Archaeology.* Cambridge: Cambridge University Press.

Rosenberg, T (2008) New Beginnings and Monstrous Births: Notes Towards an Appreciation of Ideational Drawing. IN Garner, S (ed) *Writing on Drawing.* Bristol: Intellect Books.

Roskams, S (2001) *Excavation.* Cambridge: Cambridge University Press.

Ryan, N, and Ghosh, S (2005) Ubiquitous Data Capture for Cultural Heritage Research. in Ryan, N, Cinotti, T S, and Raffa, G (eds) *Smart Environments and their Applications to Cultural Heritage: UbiComp'05.* Tokyo, Japan, Archaeolingua.

Ryan, N, Pascoe, J, and Morse, D R (1998) Enhanced Reality Fieldwork: the Context Aware Archaeological Assistant. *Life Sciences Educational Computing,* 9**,** 18-20.

Ryan, N S, Pascoe, J, and Morse, D R (1999) FieldNote: extending a GIS into the field. in Barcelo, J A, Briz, I, and Vil, A (eds) *New Techniques for Old Times: Computer Applications in Archaeology, 1998.* Barcelona, Spain, BAR International Series, S757.

Scarponcini, P, Camateros, S, Custers, O, Zlatanova, S, and van Oosterom, P (2008) Introduction. IN van Oosterom, P, and Zlatanova, S (eds) *Creating Spatial Information Infrastructures: Towards a Spatial Semantic Web.* Boca Raton: Taylor & Francis Group, LLC.

Schade, S, and Lutz, M (2010) Opportunities and Challenges for using Linked Data in INSPIRE. *Workshop on Linked SpatioTemporal Data 2010*. Zurich.

Segaran, T, Evans, C, and Taylor, J (2009) *Programming the Semantic Web.* Sebastopol: O'Reilly Media, Inc.

Shannon, V (2006) 'A 'more revolutionary' Web', http://www.iht.com/articles/2006/05/23/business/web.php?page=1. Page consulted 12 March 2009.

Shaw, R (2007) Earthwork Excavation: scanning archaeological excavations. IN English Heritage (ed) *3D Laser Scanning for Heritage: Advice and guidance to users on laser scanning in archaeology and architecture.* Swindon: English Heritage Publishing.

Silberman, S (1998) 'Upgrading the Human OS', http://www.wired.com/culture/lifestyle/news/1998/12/16752. Page consulted 17 February 2011.

Sizov, S (2007) What Makes You Think That? The Semantic Web's Proof Layer. *IEEE Intelligent Systems,* 22**,** 94-99.

Smyth, A P (1975-9) *Scandinavian York and Dublin: the history and archaeology of two related Viking kingdoms.* Dublin: Templekieran Press.

Solanki, M, Hong, Y, and Rebay-Salisbury, K (2011) SEA: A Framework for Interactive Querying, Visualisation and Statistical Analysis of Linked Archaeological Datasets. *Computer Applications and Quantitative Methods in Archaeology (CAA).* Beijing.

Spence, C (1993) Recording the archaeology of London: the development and implementation of the DUA recording system. IN Harris, E, Brown III, M R, and Brown, G J (eds) *Practices of Archaeological Stratigraphy.* London: Academic Press Limited.

STAR (2011) ‘Semantic Technologies for Archaeological Resources: STAR Project’, http://hypermedia.research.glam.ac.uk/kos/star/. Page consulted 29 June 2011.

Stoertz, C (1997) *Ancient Landscapes of the Yorkshire Wolds.* Swindon: Royal Commission on the Historical Monuments of England.

Story, D (2000) ‘Extensible Graphics With SVG’, http://www.oreillynet.com/pub/a/network/2000/04/28/feature/svg.html. Page consulted 20 June 2011.

Swartz, A (2002) ‘The Semantic Web in Breadth’, http://logicerror.com/semanticWeb-long. Page consulted 8 March 2011.

Takeda, K, Brown, M, Coles, S, Carr, L, Earl, G, Frey, J, Hancock, P, White, W, Nichols, F, Whitton, M, Gibbs, H, Fowler, C, Wake, P, and Patterson, S (2010) Data Management for All - The institutional Data Management Blueprint Project. *6th International Digital Curation Conference.* Chicago.

Tauberer, J (2006) ‘What is RDF’, http://www.xml.com/pub/a/2001/01/24/rdf.html. Page consulted 21 February 2011.

The Mighty Boosh (2005) BBC 1 television broadcast, 30 August.

Trigger, B G (1989) *A History of Archaeological Thought.* Cambridge: Cambridge University Press.

Tudhope, D, Binding, C, Jeffrey, S, May, K, and Vlachidis, A (2011a) ‘A STELLAR Role for Knowledge Organization Systems in Digital Archaeology’, http://www.asis.org/Bulletin/Apr-11/index.html. Page consulted 2 July 2011.

Tudhope, D, Binding, C, Jeffrey, S, May, K, and Vlachidis, A (2011b) STELLAR – Tools for Interoperable Archaeology. *Computer Applications and Quantitative Methods in Archaeology UK (CAAUK).* Birmingham.

Tufte, E R (1990) *Envisioning Information.* Cheshire, Connecticut: Graphics Press.

Ucko, P J (1992) Forward. IN Reilly, P, and Rahtz, S (eds) *Archaeology and the Information Age.* London: Routledge.

van Oosterom, P, and Zlatanova, S (eds) (2008) Creating Spatial Information Infrastructures: Towards the Spatial Semantic Web. Boca Raton: Taylor & Francis Group, LLC.

Vossen, G, and Hagemann, S (2007) *Unleashing Web 2.0: From Concepts to Creativity.* Boston: Morgan Kaufmann Publishers.

Wallrodt, J (2011) 'Drawings Workflow', http://paperlessarchaeology.com/category/ipad/. Page consulted 20 June 2011.

Walton, C (2007) *Agency and the Semantic Web.* Oxford University Press.

Warmerdam, F (2011) 'FWTools: Open Source GIS Binary Kit for Windows and Linux', http://fwtools.maptools.org/. Page consulted 20 June 2011.

Watt, A (2002) *Designing SVG Web Graphics.* Indianapolis: New Riders Publishing.

Weitzner, D, Hendler, J, Berners-Lee, T, and Connolly, D (2004) 'Creating a Policy-Aware Web: Discretionary, Rule-based Access for the World Wide Web', http://www.mindswap.org/users/hendler/2004/PAW.html. Page consulted 8 March 2011.

Wheatley, D, and Gillings, M (2002) *Spatial Technology and Archaeology: The Archeological Applications of GIS.* London: Taylor and Francis.

Whitehead, J (1996) 'Orality and Hypertext: An Interview with Ted Nelson', http://www.ics.uci.edu/~ejw/csr/nelson_pg.html. Page consulted 17 February 2011.

Wickstead, H (2008) Drawing Archaeology. IN Duff, L, and Sawdon, P (eds) *Drawing - The Purpose.* Bristol: Intellect Books.

Wolf, G (1995) ‘The Curse of Xanadu’, http://www.wired.com/wired/archive//3.06/xanadu.html?person=ted_nelson&topic_set=wiredpeople/. Page consulted 17 February 2011.

World Wide Web Consortium (1999) ‘Resource Description Framework (RDF) Model and Syntax Specification’, http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/. Page consulted 2 July 2011.

World Wide Web Consortium (2001) ‘URIs, URLs, and URNs: Clarifications and Recommendations 1.0’, http://www.w3.org/TR/uri-clarification/. Page consulted 2 July 2011.

World Wide Web Consortium (2004a) ‘W3C RDF Primer’, http://www.w3.org/TR/rdf-primer/. Page consulted 14 February 2011.

World Wide Web Consortium (2004b) ‘Web Ontology Language (OWL)’, http://www.w3.org/2004/OWL/. Page consulted 4 March 2011.

World Wide Web Consortium (2004c) ‘RDF Vocabulary Description Language 1.0: RDF Schema’, http://www.w3.org/TR/2004/REC-rdf-schema-20040210/. Page consulted 20 June 2011.

World Wide Web Consortium (2008a) ‘SPARQL Query Language for RDF’, http://www.w3.org/TR/rdf-sparql-query/. Page consulted 20 June 2011.

World Wide Web Consortium (2008b) ‘Cool URIs for the Semantic Web’, http://www.w3.org/TR/cooluris/. Page consulted 2 July 2011.

World Wide Web Consortium (2008c) ‘SPARQL Protocol for RDF’, http://www.w3.org/TR/rdf-sparql-protocol/. Page consulted 20 June 2011.

World Wide Web Consortium (2009a) ‘SKOS Simple Knowledge Organization System Primer’, http://www.w3.org/TR/skos-primer/. Page consulted 20 June 2011.

World Wide Web Consortium (2009b) ‘OWL 2 Web Ontology Language Primer’, http://www.w3.org/TR/owl2-primer/. Page consulted 20 June 2011.

World Wide Web Consortium (2009c) ‘SKOS Simple Knowledge Organization System Namespace Document - HTML Variant’, http://www.w3.org/2009/08/skos-reference/skos.html. Page consulted 25 June 2011.

World Wide Web Consortium (2010) ‘Linking Open Data Project’, http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData. Page consulted 10 January 2011.

World Wide Web Consortium (2011a) ‘W3C Standards’, http://www.w3.org/standards/. Page consulted 20 June 2011.

World Wide Web Consortium (2011b) ‘W3C Mission’, http://www.w3.org/Consortium/mission. Page consulted 20 June 2011.

World Wide Web Consortium (2011c) ‘RDF Current Status’, http://www.w3.org/standards/techs/rdf#w3c_all. Page consulted 20 June 2011.

World Wide Web Consortium (2011d) ‘Publications of the W3C Semantic Web Activity’, http://www.w3.org/2001/sw/Specs. Page consulted 20 June 2011.

World Wide Web Consortium (2012) ‘About the World Wide Web Consortium (W3C)’, http://www.w3.org/Consortium/mission. Page consulted 11 January 2012.

Wright, H (2006) ‘Archaeological Vector Graphics and SVG: A case study from Cricklade ’, http://intarch.ac.uk/journal/issue20/wright_index.html Page consulted 10 October 2011.

Zacharias, V (2007) ‘Ban the Semantic Web Layer Cake!’, http://www.valentinzacharias.de/blog/2007/04/ban-semantic-web-layer-cake.html. Page consulted 9 January 2011.

Zeldman, J (2003) *Designing With Web Standards.* Indianapolis: New Riders Publishing.

Zhou, M (2011) ‘CAA 2011 Beijing Conference Schedule’, http://www.caa2011.org/. Page consulted 20 June 2011.

Zubrow, E B W (2006) Digital Archaeology: A historical context. IN Evans, T L, and Daly, P (eds) *Digital Archaeology: Bridging Method and Theory.* London: Routledge.