

ResearchOnline@JCU

This file is part of the following reference:

Hardy, Dianna Lynn (2008) *Searching heterogeneous and distributed databases: a case study from the maritime archaeology community*. Masters (Research) thesis, James Cook University.

Access to this file is available from:

<http://eprints.jcu.edu.au/1849>

Every reasonable effort has been made to gain permission and acknowledge the owner of copyright material. If you are a copyright owner who has been omitted or incorrectly acknowledged, please contact ResearchOnline@jcu.edu.au and quote <http://eprints.jcu.edu.au/1849>

**Searching Heterogeneous and Distributed
Databases: A Case Study from the Maritime
Archaeology Community**

Thesis submitted by
Dianna Lynn HARDY
Bachelor of Arts - Computer Science, Graduate Diploma - Archaeology

In partial fulfillment of the Degree of Master of Social Science by Research at
James Cook University.

Archaeology, School of Anthropology, Archaeology and Sociology
James Cook University

March 2008

Statement of Access

I, the undersigned, the author of this thesis, understand that James Cook University will make it available for use within the University library and, by microfilm or other photographic means, allow access to users in other approved libraries. All users consulting this thesis will have to sign the following statement:

“In consulting this thesis I agree not to copy or closely paraphrase it in whole or in part without written consent of the author; and to make proper written acknowledgement for any assistance which I have obtained from it”.

Beyond this, I do not wish to place any restrictions on access to this thesis.

.....

(signature)

.....

(date)

Statement on Sources

DECLARATION

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from published or unpublished work of others has been acknowledged in the text and a list of references is given.

.....

(signature)

.....

(date)

Acknowledgements

I wish to thank the providers of the maritime archaeological databases that were used in this project: David Nutley (Heritage NSW), Peter Harvey (Heritage Victoria) and Vivienne Moran (Townsville Maritime Museum). Without their willingness to allow me access to their data, this project would not have been possible.

I had the benefit of having two great supervisors for this thesis: David Roe and Ian Atkinson (JCU). Both have given graciously of their time and energy during the course of this thesis. I wish to also thank Nigel Chang (JCU) who stepped in at the last phase of this project to give his input.

My thanks also go to Nigel Sim and David Laing who were very helpful in providing access to their applications (PGL and ArchaeoView) and explaining the backend processes. Trina Myers provided welcome assistance in understanding semantic search, and the odd book or article at the opportune moment.

Finally, my friends and family have been a tremendous support to me in this effort, and I thank them for their patience.

Abstract

Much of the data from archaeological investigations currently reside in databases with dissimilar file formats and structures. In addition, data from individual excavations and other research are frequently placed in separate databases that are maintained and accessed solely by the group responsible for the project. Due to the differing file formats, lack of access via a cohesive network and issues regarding ownership and use of data, maritime archaeologists have found it difficult to query such databases in order to perform cross-site analyses. This thesis seeks to provide a framework for federating maritime archaeological databases in order to make such queries and cross-site analyses possible. During this research two important question emerged, 1. Are there tools available to federate these databases? ,and 2. how can the search results be appropriately targeted when searching across such a variety of data sources?

This research began by developing a case study centred on databases provided by three maritime heritage organizations in Australia. An informal analysis of feedback from these contributors and others in the maritime archaeological community informed the preliminary design of a prototype system. One of the key issues identified by the community was a lack of funding for new tools. Therefore, the decision was made to use only "open-source" software which is available at no cost. The initial prototype system developed here employed the application 'Storage Resource Broker' (SRB). This software acts as a broker by providing access to distributed sources of data via a search engine that queries the combined resources. The holders of the individual data can set access permissions so that users only see the data to which they have been granted access.

As the research progressed another key issue was identified; although there are currently open source tools available which are capable of integrating distributed data sets, the tools are difficult to use, and require a significant level of time, technical ability and planning in order to fully implement. A related issue is the difficulty of combining data sets which may have with little data in common. To overcome these issues it was necessary to develop a separate application that works *in concert* with SRB and requires little technical ability to deposit databases. The prototype system allows a data depositor to provide a schema or description of the data itself, and to use the functionality built into the system to create a mapping between columns of data which contain similar information. Integral to the prototype is an embedded metadata catalogue (MCAT) that lists semantic metadata for each resource which allows the system to return better search results.

The final results of the research show that while it is possible to integrate maritime archaeological datasets, in order to implement a data sharing strategy, data standards for archaeological resources must be established. In addition, tools geared toward the average user must be established for creating ontologies and handling other semantic issues.

Table of Contents

Statement of Access	ii
Statement on Sources	iii
Acknowledgements	iv
Abstract	v
Table of Contents	vii
List of Figures	x
List of Tables	xii
Chapter 1 – Searching Federated Databases	1
1.1 Introduction	1
1.2 Background	2
1.3 Maritime Archaeology in Australia	2
1.4 Aims of This Research	6
Chapter 2 – Data Sharing: Federation and Search Technologies	8
2.1 Maritime Archaeology Data Sharing	8
2.1.1 Data Sharing Models	8
2.1.2 An Introduction to Databases	9
2.1.3 Current Systems Available for Sharing Maritime Archaeological Data	14
2.1.4 Digital Libraries	17
2.2 e-Research	22
2.2.1 Middleware Applications to Federate Datasets	24
2.2.2 Problems Sharing Large Amounts of Data	26
2.3 Semantic Web	29
2.3.1 Semantic Web Potential	29
2.3.2 Semantic Web Architecture	32
2.3.3 Specifics Regarding a Data Sharing System	42
Chapter 3 – Tools of the Semantic Trade: Metadata and Ontologies	47
3.1 Metadata	47
3.1.1 Metadata tags	47
3.1.2 Why is Metadata Important	48
3.1.3 The Purpose of Metadata	48
3.1.4 Metadata Standards	50
3.1.5 Metadata Harvesting	51
3.1.6 Metadata and Multimedia	53
3.1.7 Problems with Metadata Research	53
3.2 Ontologies	54
3.2.1 History of Ontologies	54
3.2.2 From Dewey to Google	55
3.2.3 Failed Ontologies: the Metric System	57
3.2.4 Characteristics of Ontologies	58
3.2.5 Levels of Ontologies	59
3.2.6 Why Develop an Ontology?	62
3.2.7 The Architecture of an ontology	63
3.2.8 How to Define an Ontology	66

3.3	Ontology Tools	67
3.3.1	Ontology Specification Languages	67
3.3.2	Methods for Creating Ontologies	68
3.4	Overall Usefulness of Metadata and Ontologies	69
Chapter 4 – Discussion of Methodology		70
4.1	Informal Survey of Data Sharing in Maritime Archaeology	71
4.2	Determine Whether There are Existing Ontologies	71
4.3	Analysis of Data Samples	71
4.3.1	Lack of Cohesion Between Data Sets	72
4.3.2	Mapping Field Names	75
4.3.3	Correctness of Data	76
4.3.4	Limitations on Case Study	77
4.4	Storage Resource Broker	78
4.5	Study 1: Personal Grid Library	79
4.6	Review Process	82
4.6.1	Archaeological Data Service	82
4.6.2	Crosswalks	84
4.6.3	Discovery Versus Cross-dataset Analysis	85
4.6	Study 2: ArchaeoView	86
Chapter 5 – Results and Discussion		93
5.1	Informal Survey of Data Sharing in Maritime Archaeology	93
5.1.1	Current Situation Regarding Data Sharing?	94
5.1.2	What Does This Community Want to Achieve?	95
5.1.3	What Problems Current Exist Regarding Data Sharing?	95
5.1.4	What Problems Must a Data Sharing Program Handle?	95
5.1.5	What Human Factors Will Have an Impact?	96
5.2	Analysis of Data Sharing Systems: PGL and ArchaeoView	97
5.2.1	Study 1: Personal Grid Library (PGL)	97
5.2.2	Study 2: ArchaeoView	102
5.3	Use of Data Federation and Semantic Search with Maritime Archaeology Datasets	108
5.3.1	Data Federation Results	108
5.3.2	Semantic Search Results	110
5.3.3	System Usability Results	113
5.3.4	Pairing Data Federation with Semantic Search	114
5.4	Research Implications	117
5.4.1	Maritime Archaeology Concerns	117
5.4.2	Information Technology Concerns	118
Chapter 6 – Conclusions and Further Research		121
6.1	Summary of Conclusions	121
6.1.1	Implementation Issues	124
6.2	Recommendations for Further Work	125
6.3	Implications for Archaeological Methodology	127
6.4	Conclusion	128

References	129
Appendix A – Glossary of Terms	135
Appendix B – Sample Ontology – XML Schema for NSW data	148
Appendix C – Use cases	152
Appendix D- Survey questions	153

List of Figures

1.1	Data Life Cycle	5
2.1	Models of data sharing	10
2.2	Features of a Z39.50 session	21
2.3	Flow crystallography data using JAINIS	24
2.4	PARADESIC system for archiving digital data	25
2.5	SRB used as a middle-ware application	26
2.6	Australasian Digital Theses advanced search screen	28
2.7	Semantic Web architecture	32
2.8	Unicode, a system for listing text on computer systems	34
2.9	A URI contains the elements URL and URN	35
2.10	Sample XML showing defined tags	36
2.11	Sample element rendered in XML	36
2.12	Sample biological ontology	39
2.13	Dublin Core metadata usage on university web site	43
3.1	Proposed ontology structure for library information system	62
3.2	Five layer model of ontology structure	65
4.1	Dataset VIC, format: Microsoft Excel	73
4.2	Dataset NSW, format: XML and as viewed in Microsoft Excel	74
4.3	Dataset TMM, format: Microsoft Access	75
4.4	Sample SQL query against NSW and VIC datasets	76
4.5	Storage Resource Broker architecture	79
4.6	Architecture for Personal Grid Library	80
4.7	Personal Grid Library User Interface	81
4.8	Archaeological Data Service search results	83
4.9	ADS individual resource listing	84
4.10	Basic architecture of ArchaeoView	87
4.11	Search engine: ArchaeoView	87
4.12	Results of the query: ArchaeoView	89
4.13	Data upload screen: ArchaeoView	90
4.14	Mapping a new dataset: ArchaeoView	91
5.1	Architecture of PGL using SRB	101
5.2	Data structure within ArchaeoView	104

5.3	System architecture of ArchaeoView	104
5.4	Intersection between datasets	110

List of Tables

1.1	Use Cases describing functionality and usability requirements	6
2.1	Example of an archaeological dataset	11
2.2	Normalization rules for databases	12
2.3	Sample loans table	13
2.4	Database terminology	13
2.5	File formats used by ADS	20
2.6	Types of questions answerable by a GIS	27
2.7	Potential answers to query question using Semantic Web	31
2.8	Sample RDF properties define subject ‘Dianna’	37
2.9	Parsing a RDF triple	37
2.10	Dublin Core metadata identifiers	43
2.11	Typical components of an archaeological project plan	44
2.12	Contents and purpose of a site/artefact database	45
2.13	Functionality requirements for maritime data sharing system	46
3.1	Metadata categories	49
3.2	Metadata characteristics	49
3.3	Resource object lifecycle	50
3.4	Metadata harvesting	52
3.5	Possible extensions to words used in each criteria	55
3.6	Dewey’s category 200	56
3.7	Categorisation effectiveness	57
3.8	Linnaean taxonomy	61
3.9	Online libraries of ontologies	69
4.1	Details of sample datasets	72
4.2	Inconsistent formatting in datasets	77
4.3	Limitations on case study	78
4.4	Common misalignments between schemas	85
5.1	Use cases for this research	100
5.2	Review of use cases for this study	106
5.3	Usability concerns for Studies 1 and 2	114
6.1	Suggested methods for use of combined search technologies	123

Chapter One – Searching federated databases

Much of the data from archaeological investigations currently reside in databases with dissimilar formats and structures. In addition, data from individual excavations and other research are frequently placed in separate databases that are maintained and accessed solely by the group responsible for the project. Due to the differing file formats, lack of access via a cohesive network, and issues regarding ownership and use of data, maritime archaeologists have found it difficult to query such databases in order to perform cross-site analyses.

1.1 Introduction

The 2001 UNESCO Convention on the Protection of the Underwater Cultural Heritage (UNESCO 2001) specifies that provisions for the archiving of data relating to excavations must be made in the project plan of every maritime archaeological project. Although ratification of this convention has not been universal, Australian maritime archaeologists do follow this standard and maintain paper and digital records of their field activities. Providing access to this data however has been difficult to achieve. Even if the data is in digital form, there is no existing method in place to supply the location of data, handle the many inconsistencies caused by multiple format types, structures and versions, and determine access rights to the data.

Each archaeological project results in the creation of multiple data sets containing information regarding that research. This thesis seeks to determine whether this data can be made available in a data-sharing environment. The aim to this research is to provide a framework for allowing searches across separate maritime archaeological databases in order to make queries and cross-dataset analysis possible. In consideration of this research focus, two further subsidiary questions emerged:

Are there tools available to federate or combine these data sets?

How can the search results be appropriately targeted when searching across a variety of data sources?

To explore these questions, a case study was developed centering on databases provided by three maritime heritage organizations in Australia. This project is based on a stated need of the archaeological community to be able to integrate datasets of varying complexity and purpose held by state and national agencies, regional museums and universities.

1.2 Background

Due to differing file formats, lack of access via a cohesive network, and issues regarding ownership and use of data, maritime archaeologists have found it difficult to query more than one database at a time in order to perform cross-site analyses. A previous maritime database integration project that reveals some of the complexities of this endeavor is the Australian National Shipwreck Database (Green et al 1993). To solve the issue of multiple file formats and data structures associated with data obtained from the multiple shipwreck sites around the Australian coast, the system designers required that all data be entered into a single homogenous database. A major difficulty with this scheme was that it required a substantial cooperative effort between data providers and mandated a high level of time and labor to input the data into the host system. The data was restricted by the original design of the database to a description of the ship, rather than a complete listing of artefacts found at the wreck site. The system was not capable of scaling because it is limited to a precise type of data. A system is capable of scaling when it is able to grow naturally in size and capability as the datasets increase in number and complexity. Systems which are not capable of scaling are limited in their long term benefits. In order to add data of a different sort to the Australian National Shipwreck Database, a redesign of the underlying database structure would be required. Due to these issues, and others such as legislative implications, lack of standards in terminologies and data security concerns, relatively few museums and researchers have been willing to input their data into these types of systems.

1.3 Maritime archaeology in Australia

The maritime archaeological community is quite small, with less than 40 full time professional members actively working in greater Australia. Often there are few practitioners working in a very large geographic area, making the sharing of information difficult due to the 'tyranny of distance'. To illustrate the issues associated with data sharing in this environment, a brief description of the history of the development of the field in Australia may be of benefit. Maritime archaeology is generally described as:

The study of human interaction with the seas, lakes, and rivers through the archaeological study of manifestations of maritime culture, including vessels, shore-side facilities, cargoes, and even human remains (Delgado 1997).

The field of maritime archaeology in Australia is young, having developed in the last 30 years. Many of the early 'pioneers' of the field are still working. In the late 1960s several

Dutch East India Company wrecks were discovered off the western coast of Australia. Once the discovery was made public the wrecks became the target of looting by salvage divers and other opportunists. In some cases the wrecks sustained extensive damage because of the use of explosives by treasure hunters (Henderson 1986, Hosty and Stuart 1994). Only wrecks in deep water or along totally uninhabited coasts have completely escaped salvage (Muckelroy 1978). The Western Australian Maritime Museum began extensive rescue salvage operations resulting in the excavation of thousands of artefacts from the Dutch wrecks. In the 1970s attention turned to the preservation of shipwreck sites through state and Commonwealth legislation. Maritime units in each State were formed to oversee the protection of shipwreck sites (Gibbs 2004) and universities such as Flinders University and James Cook University developed programs to train maritime archaeologists. As a consequence of this legislation, the maritime archaeological community now consists of three types of groups which have different research foci: museums, state based agencies and universities. In general, museums have a public education focus, and concentrate their efforts on the acquisition, curation and interpretation of artefacts. State based agencies are engaged principally in heritage management projects overseen by the Commonwealth Shipwreck Officer. The academic communities at universities are engaged in shipwreck research as well as other types of maritime investigation which consider all types of maritime culture such as technical, social, economic, political and religious spheres (Muckelroy 1978). One caveat is that although these groups have differing research goals, the distinctions are not as broad as this description may make it appear. There is often significant overlap between groups and projects that make the sharing of data a concern of great importance.

Recreational divers may be said to form an additional group of interested participants, although a consideration of their data access needs is beyond the scope of this project. Maritime archaeological finds are often of great interest to individuals or small groups who may operate independently but often have the luxury of time to engage in their efforts. Archaeology as a whole holds a tremendous fascination for many people; the Archaeological Data Service in York, England which provides web access to archaeological publications and data regarding sites in Britain estimates that over 45% of their users are members of the public (Archaeological Data Service).

Recently maritime researchers have attempted to perform analysis across multiple sites in order to draw a more complete picture of the economic, political and cultural processes at work in the maritime sphere and moreover to integrate maritime studies into a broader archaeological framework.

To move forward, Australian maritime archaeology (and Australian historical archaeology) must actively develop the structures for integration of data and comparative analysis of artefacts, sites and regions. Those undertaking the research must see themselves as trying to understand the nature of maritime systems – often newly established and evolving in the Australian context – within which the studies of individual sites are just components of a greater whole (Gibbs 2004).

Given that the maritime archaeology community in Australia is quite small and widely separated geographically, this type of research agenda is difficult to implement. Although the output of research data from maritime archaeologists is prolific, published articles and other media are generally descriptive rather than analytical and are often focused solely on individual sites rather than comparative overviews and analyses of sites across a region. In addition, access to the data generated in these projects is always a concern. Due to the small size of research groups in Australia, budgets are usually constrained, and there often is no dedicated IT staff available to assist maritime archaeology projects.

Two further issues make the sharing of data a priority in Australia. First, many of the original researchers in maritime archaeology are beginning to retire. In most cases a researcher's accumulated research notebooks, databases, and other records pertaining to excavations are turned over to the university or organization where they were employed. However, there is a growing concern that this data is sometimes unusable unless explanatory notes are kept describing the context of the data. Just as an artefact without provenance cannot be placed in the archaeological record, columns of numbers with no associated information detailing their significance is not useful. Second, archaeological excavation by its very nature is destructive. A shipwreck site once disturbed can never be re-assembled. Therefore the onus is placed on the researcher to take precise notes and ensure that the complete set of the data is available to the community at large.

The issue of providing access to data which is located in diverse locations and in multiple file types is not only a problem for archaeologists, but for research scientists in general. A strategy called e-Research is beginning to take shape, where advanced computation and networking tools are used facilitate collaboration among distributed researchers and communities. Science is increasingly being performed through distributed global collaborations where computational activities as well as data and scientific instruments can be delivered and shared over local area networks and the Internet (Hey and Trefethen 2002). e-Research addresses the need for researchers to gain access to technical

infrastructure regardless of their geographic location. In a discussion paper, the Australian Government recognized the need for collaborative data sharing systems.

Developments in ICT (Information and Computer Technologies) are also changing research methodologies and enabling formerly inaccessible problems to be addressed... The Australian Government has recognized that the future of research will be collaborative, across research organizations, across countries and across the globe (Sargent 2005).

The term data life cycle (DLC) describes the phases through which data passes once it has been collected. The DLC is described as: data collection, repository, processing and dissemination (see Figure 1.1 below). e-Research initiatives set up systems that enable the maintenance of and access to data throughout the DLC.

[Image removed due to copyright restrictions]

Figure 1.1 – Data Life Cycle – (<http://www.itaginfo.com>)

Data gathering techniques vary between fields with some disciplines utilizing digital tools to a greater extent than others. In many fields, a combination of paper and computer aided techniques is utilised. For example, in archaeology digital photographs of in situ artefacts in addition to sketches are often employed. Once the data have been collected, it must be deposited, and then disseminated to others within the organization. This requires the user to bring the data back from the field and place it in some sort of repository. The final step, dissemination is an area where the existing situation in maritime archaeology falls short of what is desired by researchers regarding data sharing. The prime vehicle for publishing the results of maritime archaeological research in Australia currently is AIMA, the Australasian Institute of Maritime Archaeologists. This group meets on a yearly basis and publishes a journal that allows researchers to share the results of their most recent projects, however access to the raw data from those projects is beyond the scope of their publishing mechanisms.

1.4 Aims of this research

The two main areas of research investigated in this thesis are the federation of maritime archaeological databases and the targeting of search results when searching across multiple databases. The aim of the case study is to determine the suitability of the proposed framework to aid maritime archaeological research. Implementation and usability issues are of equal value in this consideration. In order to examine the feasibility of data federation paired with semantic search, two separate systems were evaluated in view of their ability to meet the project requirements. These functionality requirements are detailed in Table 1.1 below.

A use case is a description of a typical scenario that may be encountered by a person using the proposed system. Use cases are created to specify in detail the functionality that is required for new computer systems. Each use case has an actor (the person interacting with the system), stakeholders (other people that this action impacts) and other information regarding the action being taken. In this case study there are only two persons identified: a data provider and a researcher. The data provider is the one depositing the data and the researcher is someone who is trying to gain access to the data.

ID	Use case description	Actor(s)	Stakeholder(s)
1	User is able to deposit a dataset	Data provider	Researcher Data provider
2	User is able to set access rights for data	Data provider	Researcher Data provider
3	User is able to make data discoverable	Data provider	Researcher Data provider
4	User is able to find data	Researcher	Researcher Data provider
5	User is able to view data online	Researcher	Researcher Data provider
6	User is able to download data	Researcher	Researcher Data provider
7	System is easy to use	Researcher Data Provider	Researcher Data provider
8	System allows searches across multiple datasets	Researcher	Researcher
9	System returns search results combined in one table in a web browser	Researcher	Researcher

Table 1.1 – Use cases describing functionality and usability requirements

This first chapter has provided an introduction to the research questions and the general bounds of the research. An outline of the current tools and technology available to handle this problem are explored Chapter 2. An in-depth consideration of the main building blocks of semantic search, i.e. metadata and ontologies is given in Chapter 3. The methodology used in this project is detailed in the methods section (Chapter 4), and was based on an iterative process composed of a cycle of consultation with community members, design of a potential system, and a review of its suitability for the maritime archaeological research environment. The viability of the prototype system is explored in the results (Chapter 5), and a summary of conclusions as well as future steps are specified in the final chapter (Chapter 6). Supplementary material such as a glossary of terms, a sample ontology, the use cases and data obtained from an informal survey of maritime archaeologists are available in the appendices.

The e-Research technologies described in this thesis have the potential to provide many benefits to researchers in the humanities as well as science. Rather than implementing a central storage system with the associated issues of keeping multiple copies of data in synch, this project explores the establishment of a federated environment where groups can manage their own data at their location. The goal of the project is to allow researchers to deposit, discover and gain access to research data held at separate locations. The remainder of this thesis will explore the suitability of these techniques for application to maritime archaeological data.

Chapter 2 – Data sharing: federation and search technologies

This chapter outlines previous research on data federation schemes and search technologies. e-Research is a key component of this review, and a detailed analysis is provided of the tools and technologies associated with this area of research. These analyses are directed towards the understanding of data sharing in general, and to illustrate the possible use of data sharing in the maritime archaeology community. Although e-Research has recently been extended to include the humanities, in the past the study of data sharing technologies, and the associated search techniques has been focused primarily on the sciences. This thesis seeks to expand this research to the discipline of archaeology in particular and to allow access to data that was previously unavailable.

2.1 Maritime Archaeology Data sharing

At the present, datasets for archaeological investigations are stored in separate locations. There is no mechanism available to allow researchers to share data easily. Section 2.1 details the current situation concerning maritime archaeological data sharing. In order to fully explain the implications of the lack of a means for the dissemination of maritime datasets, a background in data sharing models and databases in general is provided also.

2.1.1 Data sharing models

Regardless of the discipline, there are three distinct models of data sharing systems: centralization, distribution and federation. In a centralized system, all data is stored and managed by a central data management system. All datasets are managed by a central authority, and must adhere to the standards and rules of the central system. In this type of system, a central administrator monitors and facilitates access to the data. To deposit data, the dataset provider must make changes to the stored information to meet the guidelines mandated by the administrator. Mainframe computers are an example of this type of system. In the distributed model, data is stored at separate locations that are distributed throughout an organization. Data processing can be shared by multiple computers. In this model, the owners of the data are responsible for managing and limiting access to their own data. Distributed systems offer a high level of autonomy for data owners, but also require that the owners of the data manage access to this data. Federation offers a third model of data sharing, where resources are distributed throughout the enterprise, but appear to be located in a common central repository. Access to distributed documents is handled via a central system.

The search engine Google can be seen as an example of federated search. Google maintains a list of locations for each resource. When a user requests these resources via a search, Google presents the user with a link that contains the location for the data. This allows the resources to be stored at their original locations, and eliminates the issue of duplication of data. For the purpose of this research the definitions in Figure 2.1 will be used.

2.1.2 An introduction to databases

State based societies have always had a need for maintaining precise records regarding their citizens. Databases have provided modern societies with a means for tracking data in an easy to manage format. The invention of databases has occurred in link step with the development of computer systems. In the 1890s, a former employee of the U.S. Census Bureau, Herman Hollerith, developed the first automated information processing equipment. The punch card system that he designed was used to collate the census for 1890 and 1910. In 1911, he and another ex-Census Bureau employee founded International Business Machines (IBM). In rapid succession this technology was used to collect the new income tax (1913), manage records of inductees for World War I, and to document the work history of over 26 million people under requirements of the 1935 Social Security Act. The Census Bureau purchased the first commercially available computer called the Universal Automated Computer (UNIVAC) in 1951, specially designed for collating the 1950 census (National Research Council 1999).

Until the 1960s storage of data was highly dependent on the type of machine that was used. Two main data models were established: Conference on Data System Languages (CODASYL), a network based storage mechanism using a new programming language called Common Business-Oriented Language (COBOL), and a hierarchical storage system called the Information Management System (IMS) which was partially derived from a National Aeronautics and Space Administration (NASA) Apollo project. Following the introduction of these technologies the term 'database' emerged to describe the idea that the information on a computer could be stored, managed and accessed separately from the operating system (National Research Council 1999). Previously the computer's operating system was required to handle all interactions with the data. This would be similar to having to issue individual commands to Windows XP or Linux rather than having a software application handle these interactions.

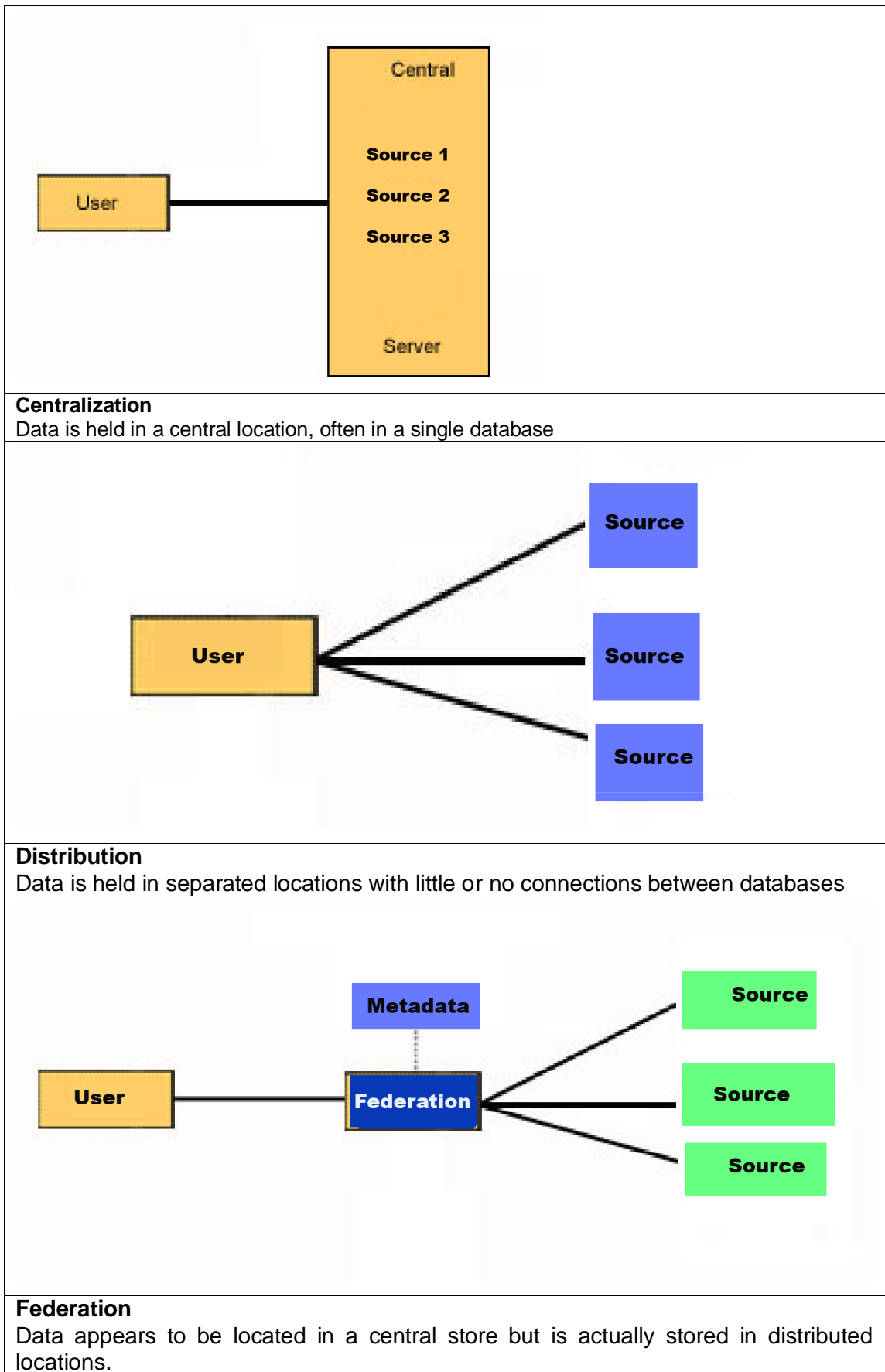


Figure 2.1 – Models of data sharing

In 1970, an Oxford-trained mathematician employed by IBM published a seminal paper where he defined a new standard for dealing with databases called the relational model (Codd 1970). Codd’s model has two key points:

Data independence from hardware and software implementation
Automatic navigation, or a high-level nonprocedural language for accessing data
(National Research Council 1999).

The goal of his data model was to enable designers to set up databases in a standard way so that software applications could use a simple language to access the data. In this way, programmers could update multiple rows in a dataset with one command.

A database is commonly defined as a collection of information that is organized so that it can be easily accessed, managed and updated (Conolly and Lake 2006). Most modern databases use the relational model to handle relationships between items in the data. For instance, an order-tracking database might maintain connections between a customer’s name and the items that the customer has ordered.

Archaeological datasets may make connections between a particular site id and grid section numbers that make up that site. Taking this further, each grid number may be segmented into specific X, Y and Z dimensions for the location of an artefact that has been excavated. Each dimension is based on the distance from a center datum that provides the point of reference. This data would be placed in a table, within columns that specify exactly what sort of data is expected in each field. Table 2.1 can be used as an example of this kind of data.

Artefact ID	Site Number	Grid Section Number	X	Y	Z	Description
212	03	24	-452	265	34	Metal nails
213	03	22	-398	154	56	Ceramic jug
214	03	23	-423	175	27	Table leg

Table 2.1 – Example of an archaeological dataset

In general, database tables are designed to follow specific rules in order to allow data to be retrieved and updated in a logical fashion. These ‘normalization’ rules attempt to ensure that duplicate records are not created, and that a change to one record does not negatively impact other related records (Kent 1983). These guidelines are summed up in Table 2.2.

1	The same information should not be stored in more than one table
2	A table should not contain duplicate information
3	Each table should contain information about a single subject
4	Each piece of data should have a unique identifier

Table 2.2 – Normalization rules for databases

An examination of Table 2.1 and the rules in Table 2.2 reveals that the dataset is well designed according to the normalization rules. All of the data in this table is related to the same subject: artefacts. Also, each artefact has a unique identifier (ID). This ID differentiates one individual artefact from another, even if they are similar in description and appearance. In only one case does the data set break the normalization rules: the site number is duplicated. To be completely normalized the table should be broken up into a few smaller tables, perhaps an artefacts table with only the artefact ID and the X, Y, Z coordinates for example. This occurs often in many databases. For instance in the case of a postal code the database might list name, address, suburb, state and postal code. Since State and Postal code may be duplicated these could be broken down into individual tables. However this would slow down search results, so in most cases the entire address is retained as a group in the same table. This is referred to as de-normalization, and is allowed in circumstances where it is necessary to aid performance. In the case of the data set in Table 2.1, the data is grouped in a manner that meets most of the rules for normalization, and is suitable to be included in a cross-database search (Codd 1970).

Databases generally include the following items: tables, records, fields, and primary keys. A database can contain one or more tables. Each table contains information pertaining to a particular subject. In order to illustrate this point, take the case of an archaeological dataset. This data consists of a database with three tables named Team_Members, Artefacts, and Sites. Team_Members would contain a list of all the archaeologists who had worked on a particular project. Similarly, the Artefacts table would contain all of artefacts found during the excavation. The Sites table would list excavation sites defined for the project. This sort of table is depicted in Table 2.3.

Since this database is properly normalized the table holds numbers that relate to items in other tables. For instance if you looked up MemberID 234 in the Team_Members table this might show the name of James Smith. Using Table 2.3 as a sample, we can define the terms listed in Table 2.4.

SiteNumber	ArtefactNumber	Date	MemberID
213	34-546	23/12/2005	5467
214	58-090	23/12/2005	234
215	23-776	24/12/2005	10987

Table 2.3 – Sample Sites table

Unique Key	A specific number that specifies a particular item. Also called a primary key. This would be SiteNumber
Foreign Key	Refers to an item in another table. In this case ArtefactNumber and MemberID are foreign keys
Field	Any column in a table is referred to as a field
Record	A row of a table is referred to as a record

Table 2.4 – Database terminology

In the early 1980s the large database manufacturing companies embarked on an aggressive marketing scheme and drove many of the competitors out of business. As a result, prices for the remaining database systems skyrocketed. In an effort to acquire small business customers, Microsoft introduced personal productivity tools such as Excel and Access, exposing a niche for personal-use software. The resulting availability of relatively inexpensive, easy-to-use database software such as Filemaker, Microsoft Access and MySQL took the development of databases out of the hands of data specialists and placed it within the reach of the average user. Large amounts of maritime archaeological data is placed in databases which have been created quickly, without much planning or forethought regarding how this data will be accessed, managed and archived in the future (Roe 2005 pers comm). Many times these tables take the same layout as the forms used to collect the data in the field. Cryptic column headings are often the norm, making it difficult for others to analyze the data at a later date. A further consideration of the problems caused by database design will be considered in the results (Chapter 5).

2.1.3 Current systems available for sharing maritime archaeological data

Data sharing in maritime archaeology is currently limited to data published in paper form: (journal articles and site surveys) and data published via the Internet (museum sites and the Australian National Shipwreck database). Due to space constraints and the cost of publishing in Australia, very little raw data is published in paper form. The major journal for maritime archaeology in Australia, the Australasian Institute for Maritime Archaeology (AIMA) Bulletin is published once a year. Articles in this journal are primarily descriptive in nature, and are brief, often less than 10 pages in length. The referee process for new articles may add as much as 6 months – 1 year to the time needed for publication. Site survey reports are an alternate mechanism used to distribute data and are published up to 5 years after the project completion (Chang pers comm 2006). This increasingly long lag time between the generation of the data and its publication in a journal does not meet the needs of archaeologists who require timely and up-to-date information. Internet publishing of non-refereed journal articles and other data from maritime projects attempts to solve some of the problems with publication schedules and cost, but can be controversial. Since the articles have not been through the referee process, this information can be seen as merely opinion, rather than substantiated fact. This is a problem in all disciplines, not just archaeology.

The peer-review process has been a fundamental part of modern science and provides a mechanism for the critical review and evaluation of scientific work. This process distinguishes publication in a scientific journal from other forms of communication (such as news releases and public talks) and implies that certain criteria have been met (Adams and Venter 1996).

The authenticity of data presented on the Web has long been an issue with researchers. Unless others of professional standing have vetted an article, it is difficult for the authenticity and rigor of the results to be established. In an effort to overcome this limitation, archaeological researchers in the UK have set up an independent, not-for-profit electronic journal that contains new articles of a high academic standing. ‘Internet Archaeology’ (IA) is published twice a year, and is the first fully refereed e-Journal for Archaeology. The first issue of IA appeared in 1996, and since that time the journal has been accessed by over 27,000 readers in 120 countries (Heyworth et al. 1995). Their primary goal is to exploit and research the emerging Internet medium for publishing of archaeological research. The publication has moved to a subscription model for funding purposes. IA publishes articles regarding excavation reports, analyses of large data sets (with links to the original data available for download), as well as data visualizations. The

IA actively seeks papers which detail completed research projects, and preliminary reports regarding on-going work are not considered. All articles and associated data are automatically archived at the Archaeological Data Service (ADS) in the UK.

Although the IA ostensibly has an international focus, due to the advisory committee being composed primarily of representatives from the British Academy, the Council for British Archaeology and a range of universities in the UK, the focus is naturally on British archaeological efforts. The IA will accept articles from foreign sources, even to the point of publishing in other languages than English. However, an examination of the articles published in the last decade reveals a heavy focus on British research. Many of the same benefits that the IA touts can be provided via a distributed data access model such as that specified in this project without the limitations imposed by a subscription model.

Some museums and heritage organizations have started to make portions of their data available via the Internet. The four systems listed below are indicative of this type of data access and were selected to show the range of data available.

Heritage NSW

The Heritage Unit for New South Wales has a substantial listing of artefacts, articles and other data which it makes available for download via its website. Included are interactive maps, a searchable database, still images, video clips, archaeological survey reports and activity guides for educational use. In addition, users can access summary shipwreck graphs listing wrecks by Country Build, Rig, Construction Type, Region, Industry, Decade and Month. Also included are videos and panorama images of shipwrecks recorded in QuickTime format. Through this web site the user can also gain access to the Australian Shipwreck Database and the New South Wales State Heritage Inventory. The focus on the Heritage NSW site is on providing information to the general public regarding maritime archaeology. Therefore no provenance information concerning the artefacts is provided. As such, in order to gain useful information for archaeological research concerning these artefacts, the researcher will need to contact the museum directly to request access to further information (Maritime Heritage Online: New South Wales).

Museum of Tropical Queensland

The Museum of Tropical Queensland in Townsville holds artefacts retrieved from several excavations of the Pandora wreck site. The ship Pandora wrecked on the Great Barrier Reef in 1791 while transporting captured mutineers from the Bounty back to England for trial. Although the artefacts from these excavations are housed and displayed at the

Townsville museum, the artefact database for this project is hosted by a regional museum, the Queensland Museum in Brisbane. The link to the database is located deep within the parent museum website, but is available for general user access. A brief search on the website reveals the size of the artefact catalog. A search for 'wood' as the keyword returns over 200 entries. Although each artefact is listed and a photo of the item is presented, relatively little other data is available via the search page. Each photo is accompanied by a listing of the register number, grouping, description and materials. As with the Heritage NSW site, no contextual information is presented for these artefacts (Museum of Tropical Queensland).

JCU Picture Archive

James Cook University hosts a digital photo archive, and makes selected photos available over the Internet. Due to copyright restrictions, users are requested to obtain permission from the donor of the photo before using it in a commercial publication. These images are in three collections: the Daintree, the North Queensland Collection, and the Nelly Bay Collection. The Daintree Collection consists of 50 photos taken during the last twenty years which detail protest rallies surrounding the declaration of the Daintree area of North Queensland as a World Heritage site. The North Queensland Collection contains a selection of 306 photos out of a group of 40,000 historical photographs of Townsville and surrounding districts engaged in industry and shipping activities. The Nelly Bay photograph collection is 120 photographs of the Nelly Bay Harbour development on Magnetic Island. It includes pictures of construction of the island break wall, and demonstrations against the development as well as images of the aboriginal flag raising. The archive can be searched via a search engine on the JCU website (JCU Picture Archive 2006), or through Picture Australia (<http://www.pictureaustralia.org/>) a service hosted by the National Library of Australia. Picture Australia contains a search tool which returns results from libraries, galleries, museums and other archives. The user can search on the JCU site by Title, Description or Source, and also by the Asset Modification Date for each item in the database.

Australian National Shipwreck database

The Australian National Shipwreck database is available online (Australian National Shipwreck Database), and the data in this system is updated by the individual state heritage units. The data is restricted to a listing of shipwrecks known to the maritime archaeological community. The listing contains details regarding basic data regarding the ship, such as length, draft, date of sinking as well as location information of the shipwreck. Since this information is provided to all users of the website, some effort has been made to only give

general descriptions of the shipwreck locations to avoid illegal salvage attempts. The actual location details of the wrecks are maintained off-line by the individual heritage agencies. Since these locations are held in separate databases, access to this information is not available online to legitimate researchers.

Although basic information is available on these five Web sites, any cross site analysis is impossible without access to more information on each artefact. The four online data providers listed here focus on conveying their information for public consumption rather than for the use of researchers. Due to this fact, data which would be of use to archaeological research projects is not provided online, and access to this information must be requested separately.

The next section of this chapter describes data sharing systems that have been put in place in other disciplines. Digital libraries present one solution to the issue of sharing multiple documents online. Following is a discussion of the purpose of digital libraries, the technology used by digital libraries in allowing searches, and a description of a digital library which contains archaeological data.

2.1.4 Digital Libraries

Universities, museums and libraries are beginning to make portions of their research data available online. These repositories are often called digital libraries and their content can vary widely from a simple archive of a few items to an extensive collection with millions of items available for viewing and download. In most cases a relational database of some sort contains the digitized images, text or other media that is provided to users via a web site search utility. Rather than static hypertext markup language (HTML) pages, each piece of data is retrieved from the database and a new page is generated at the time of the user request. The NSW Heritage Unit mentioned previously, provides access to the items in their artefact database through this type of mechanism.

Although digital or electronic libraries were first defined in the early 1990s, the first implementation of one didn't occur until 1994. In the United States, digital libraries were designated a 'national challenge application area' under the High Performance Computing and Communications Initiative (Borgman 1999) and digital libraries were made a key component of the National Information Infrastructure by the Office of Science and Technical Policy in 1994. In the UK a similar process gave rise to the Electronic Libraries Programme (eLib) with the goal to 'transform the use and storage of knowledge in higher

education institutions' (Oppenheim and Smithson 1999). The Internet Archaeology Journal was established as an offshoot of this program.

Digital libraries are multi-disciplinary in nature due to the over-lapping of skill sets necessary to their creation and maintenance. While researchers from a computer science background tend to focus on the technology and network infrastructure needed to facilitate information delivery, library and information science practitioners emphasize the importance of the content, organization of the data, publishing mechanisms, copyright issues as well as concerns regarding user behaviour and human computer interactions. These distinctions are visible also in the competing views of what composes a digital library. Librarians associated with traditional bricks and mortar institutions take a broad view of the term, and recognize a library as an organization that selects, collects, organizes, conserves, preserves and provides access to information for some group. In this definition digital libraries merely use a new delivery system to enable users to view data. The computer science community however takes a more narrow focus when discussing the term library. Emphasis is generally focused on databases, storage needs and information retrieval mechanisms (Borgman 1999).

Borgman (1999) specifies that a digital library is comprised of databases with content that may be text, images, or other media collected on behalf of a user community. This brings up an intriguing question: can all databases be considered digital libraries? Many data-centric web sites such as Lexis-Nexis, WestLaw or even Wikipedia describe themselves as digital libraries. Semantics aside, the difference between a database and a digital library can be summed up as follows: a database contains a number of items which all pertain to one particular topic, while a digital library contains links to many different items that may have nothing in common. For this reason, the terms may not be used interchangeably. Although it has become common to refer to a database as a digital archive, it does not follow that all databases or archives are digital libraries. Griffiths (1998) argues that databases on the Web do not comprise a true digital library because content is incomplete, data standards are lacking, minimal cataloguing is available and methods for information retrieval are ineffective.

Archaeological Data Service

Other groups outside of Australia have encountered difficulties in facilitating data sharing as well. The Archaeological Data Service (ADS) in the UK works in conjunction with its parent organization, the Arts and Humanities Data Service (AHDS) to:

support research, learning and teaching with high quality and dependable digital resources...by preserving digital data in the long term, and by promoting and disseminating a broad range of data in archaeology.

(<http://ads.ahds.ac.uk/project/about.html>, see also Richard 1997)

Although primarily focused on archiving archaeological data, the ADS also promotes standards and guidelines for best practices in the creation, dissemination, and preservation of archaeological research data. The group utilizes a catalogue that includes links to resources outside of the ADS network in order to provide access to data. Wise (1997) makes a key point that archaeology is in a special position in that much of the creation of data results from the destruction of primary evidence. This makes it doubly important that the data derived from these excavations be made available to the wider research community. Another ongoing problem is that previously digitized data is in danger of becoming obsolete and unreadable due to antiquated storage mechanisms unless it is moved or 'migrated' into a modern storage system and format. Large amounts of the archaeological data stored on these various types of media, document unpublished excavations, thus making access to this rare data all the more urgent. Consider the case of site reports stored on 8 inch floppy disks or cassette tapes. The ADS migrates the data from its original format into American Standard Code for Information Interchange (ASCII) text files and non-proprietary formats whenever possible. Table 2.5 describes the file formats accepted for archiving as well as the display mechanism usually chosen for the specific file type.

As can be seen by the table below, all databases, spreadsheets, statistics and texts available on the ADS system are presented in ASCII delimited text. Breaks or large spaces between items of data are delimited or specified by commas or other special characters. Images and video files are stored in Joint Photographic Experts Group (JPEG) or QuickTime (a royalty-free video format) while Geographical Information System (GIS) and map data are rendered in formats that are the current industry standard.

The ADS provides a valuable service to archaeological researchers in the United Kingdom but limits its acquisitions to data from the UK. Data sets from outside this area are not sought. In its archiving mechanisms, the ADS operates very much like a standard library. Material can be 'deposited' assuming that it meets criteria specified by the administrators of the service, in this case the digital library. This contrasts strongly with the needs of the maritime community that is interested in making the data available from its original location rather than archiving it at a central repository. For data holders who wish to

maintain their data in its original location, the ADS requires depositors to format their data to meet the requirements of a data transfer interface protocol called Z39.50 (Hammer and Favaro 1996). This protocol is described in the next section.

[Table removed due to copyright restrictions]

Table 2.5 – File formats used by ADS (<http://ads.ahds.ac.uk/project/faq.html>)

Z39.50

The ADS currently uses Z39.50, a data transfer interface protocol frequently used in digital libraries. Interface protocols are used to define the mechanisms and data transfer schemes that allow computer systems to communicate. To describe this in terms of human interaction, this would be similar to a simple trade language that is used between people who don't speak the same language. A protocol is a set of simple phrases or commands that each system understands. The Z39.50 protocol supports networked access to suitably formatted distributed databases over the Internet. By 'suitably formatted', this implies that each database has been modified to meet some criteria mandated by a central authority. To search a database held on another machine, a user enters a search phrase into a search engine. The request is translated into a special query language and sent to the remote server that holds the data that the user is trying to obtain. Once the server receives the request it translates the query into language understandable by the databases (Oracle, MySQL, etc.) and retrieves the results which it sends back to the user. On the user's machine the code must be translated back to HTML, and the results displayed to the users. Figure 2.2 gives a simplified view of the architecture necessary to use Z39.50 for data transfer requests (Wells et al 1998).

[Table removed due to copyright restrictions]

Figure 2.2 – Features of a Z39.50 session – National Library of Australia
(<http://www.nla.gov.au/nla/staffpaper/awells2.html>)

Z39.50 is frequently seen as an old technology because development on it began in the 1970s and it was implemented in successive versions in 1988, 1992, and 1995. Although the protocol has been updated, most recently in 1999 called the Bath Profile (Gethin 2001), it is limited in its ability to integrate Web databases of varying types. It works best when the data are similar in format and content. The protocol is complex to administer, configure and maintain, furthermore the data transmittal process can break easily unless precise implementation rules are followed (Troll and Moen 2001). The Z39.50 scheme requires that each database be of a similarly format, and that a data administrator set up and monitor the system as these data requests take place.

2.2 e-Research

A common thread that links all research endeavors including those in the sciences and humanities, is the requirement to collect, collate and analyze data in order to support the knowledge generation process. The newly emerging area of e-Research strives to make information technology tools and methods that support this process available to research communities. Two types of projects that fall into the e-Research category are those that are carried out in highly distributed, i.e. widely separated locations, and those that use extremely large data sets that require access to supportive network infrastructure. Examples of such projects include fields as diverse as particle physics, genome mapping and social simulations.

In 2004, the Australian government awarded the National Collaborative Research Infrastructure Strategy (NCRIS) over \$500 million to bring a more strategic direction to Australia's investment in research infrastructure. One of the priority areas for research was the development of a platform for collaboration for data access, discovery, storage and management (Sargent 2005).

As the needs for many of the specific enabling technologies (such as high-speed data communications) are shared by all disciplines, investment in them is best managed on a system-wide (rather than discipline-by-discipline) basis. This has particular ramifications for the humanities and social sciences...Ideally, investment in platforms for collaboration should provide researchers with the ability to: gain access to information relevant to their field from a variety of sources seamlessly; exchange information collaboratively with colleagues; annotate their datasets or publications; and to manage and disseminate the results of their research through supported repositories (NCRIS 2004).

The development of the necessary infrastructure to support the exchange of research data falls under the category of e-Research. e-Research describes research that uses an array of computer applications to provide support for interaction and collaboration across distributed groups of researchers. It often uses immense datasets that can require advanced computing or tools to interpret and present the data, but can equally be used in the humanities and social science research with great effect. A common characteristic is that the projects are carried out collaboratively via distributed researchers and communities (Schroeder and Fry 2007). This project falls under the category of e-Research in that the system under development will provide an accessibility framework for connecting to specific data locations. A search and data presentation methodology is a key component of this research. A number of technologies must work together to allow eResearch systems to accomplish the task of enabling collaboration and access to distributed data sources. These technologies are described in section 2.3.

To describe the breadth of the disciplines that can utilize e-Research, following are details concerning two e-Research projects: one from the scientific area and one from the humanities. The first project named JAINIS (James Cook University And Indiana State University Instrument Service) is a collaborative development between two universities. It allows researchers to remotely operate crystallography equipment in order to run experiments where the researcher is not physically present at the location where the laboratory equipment resides. This is of particular interest to scientists who perform their work at regional institutions that may not otherwise have access to advanced crystallography tools. The JAINIS software allows the researcher to better manage the data being generated by the Rigaku X-Ray Crystallography machine and be able to view the current status of the laboratory including the crystal image and diffraction pattern remotely (Atkinson, pers comm). Figure 2.3 shows the architecture for this system.

A second e-Research project is called PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures). The purpose of this research is to provide a method for digital conservation and access to materials from the Pacific region which might otherwise be lost. The project develops tools allowing collaboration with other groups in order to promote good field practice in documentation and digital archiving of endangered languages. The collaboration tools established in this research project allow the recording, digitization, annotation and access to video and audio files concerning anthropological exploration into languages and cultures of peoples from the Pacific (Thieberger 2004). Figure 2.4 shows the technical architecture for this project in regards to archiving the data (<http://www.paradisec.org.au>).

[Figure removed due to copyright restrictions]

Figure 2.3 – Flow of crystallography data using JAINIS

2.2.1 Middleware applications to federate datasets

In order to gain access to datasets that are distributed throughout an organization, a middleware tool must be used. In the computer industry, middleware is a general term for any programming application that serves to ‘glue together’ or handle the communication between two separate and often already existing programs. Middleware applications perform the technical negotiations between data requestors and data owners. In this maritime archaeological research project, the focus is on allowing users in widely distributed locations to share data sources and databases.

A commonly used middleware application for e-Research applications is Storage Resource Broker (SRB). This software acts as a ‘broker’ by providing access to distributed resources. SRB is widely used in e-Research and has been made available to James Cook University and other Australian universities by the San Diego Supercomputing Center (<http://www.sdsc.edu/srb>, see also Moore 2001). SRB can be used to link resources from separate databases. Users can access collections via a search engine that queries resources that have been federated via SRB. Once data has been identified to the data federation,

multiple datasets can be queried as if they were in the same location. The holders of the data can set access permissions so that users only see the data that they have been granted rights to.

The application contains a metadata catalogue (MCAT) that lists semantic metadata for each resource that allows the search engine to return better results for searches. Metadata at its most basic level can be considered data about data. One of the MCAT's most vital

[Figure removed due to copyright restrictions]

Figure 2.4 – PARADISEC system for archiving digital media

<<http://www.paradisec.org.au>>

benefits is that it allows data to be held in separate locations, rather than in one central database. The ability of the MCAT to handle data stored in separate locations allows the data to grow as needed, and permits local database administrators to maintain their repositories in a 'as is, where is' manner. This is a particularly important point as it results in maintainability and scalability as the collections grow. Figure 2.5 contains a diagram of a typical use of SRB to create a data federation.

[Figure removed due to copyright restrictions]

**Figure 2.5 – SRB used as a middleware application to provide access to data
(after <http://roadnet.ucsd.edu>)**

2.2.2 Problems sharing large amounts of data

Once multiple datasets are combined, a further problem is presented by the difficulty in searching across a large amount of data. Just as a generic search via a search engine like Google can swamp a user in thousands of links which may have little relation to the data in which the person is interested, a search across a multitude of archaeological data can bury a user in what has been described as a 'data deluge'. Hey and Trefethen predict that the data generated by computer simulations, large instruments, sensors and satellites are likely to soon dwarf the accumulated scientific data collected to date in the history of scientific exploration (Hey and Trefethen 2003). While the size of current archaeological datasets does not approach this magnitude, the total amount of information obtained from systems such as GIS and other data intensive systems is constantly increasing. Conolly and Lake (2006) report that over 90% of the Site and Monument Records in the United Kingdom contain GIS data, although standards have not been put into place to determine the exact format and detail for the data included in reports.

The data from a geographical information system (GIS) can be very useful when it is made available to a federated system as an additional resource. While a GIS uses many types of metadata, in general the system tracks two types of information: *attributes* that describe what is present and *location* that describes where the item is found (Worboys 1995). The combination of attribute and location descriptors provides a method for describing the relative position of objects in relation to each other. This tracking method is key in a discipline like archaeology where the physical relationships between the placements of items in the landscape form a prime source of data. Many recent archaeological projects utilise GIS in the execution of their research. Although none of the sample datasets provided for this research contained GIS data, it is important to not discount this source of archaeological information. The sorts of questions that can be posed to a GIS system are very similar to the sort of questions that can be posed to a sufficiently populated federated data system. Table 2.6 below contains a listing of the types of requests that can be made of a GIS. The same type of data can be made available to a federated data sharing system which is capable managing data from much wider sources than the typical GIS.

Question	Example
Location	What artefacts have been found along the proposed route of the new road?
Condition Trend	Where were Roman coins dating to the second century AD found?
Routing	How does the density of primary debitage change as one moves away from the prehistoric hearth?
Pattern	Are the burial cairns distributed uniformly across the landscape, or do they cluster on SE facing slopes?
Modelling	Where would one expect to find more Mesolithic campsites?

**Table 2.6 – The types of question answerable by GIS system
(after Connelly and Lake 2006)**

The use of data visualisation systems such as GIS has had an effect on the current archaeological method. Models have been developed to attempt to predict where sites will be found by taking into account factors such as soils, climate, water and terrain. (Kvamme 1990, Brandt et al. 1992). In addition, datasets can be used in combination with each other. A GIS has been successfully used to overlay data such as the distribution of metal artefacts discovered with aerial photographic coverage, finds discovered in field walking, magnetometry and resistivity surveys and excavations (Roskams 2001).

An area of increasing interest is the field of computational archaeology that describes computer based analytic methods for the study of long-term human behaviour and evolution. This term is generally used to apply to archaeological research that would be difficult if not impossible to complete without the aid of a computer (Claxton 1995). These types of analyses are rapidly increasing the total amount of data that make up archaeological datasets.

In dealing with the problem of searching across large datasets, two basic models been developed. The first places the burden of narrowing the search onto the user. Generally an advanced search utility is provided, and the user must create a query with a number of highly technical search parameters. An example of this is seen in the Australasian Digital Theses (ADT) program. A screen from their advanced search engine is seen in Figure 2.6 below (see also Suleman et al. 2001).

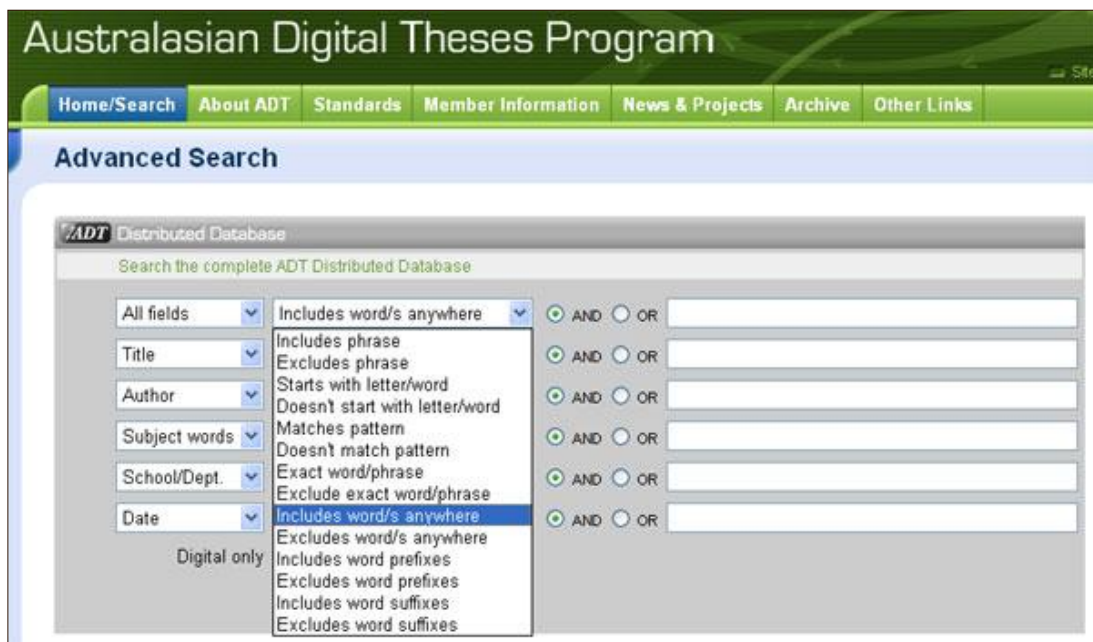


Figure 2.6 – Australasian Digital Theses advanced search screen
 (<http://www.adt.caul.edu.au/homesearch/advancedsearch/>)

The ADT provides a means to search through the combined set of digital theses produced by Australian Universities. As the number of theses per university can be quite large, it is necessary to limit the search results. The user is prompted to create a complicated query via the search engine. This model is extremely common with digital libraries. The assumption is that the user will be skilled enough to create the appropriate query. This may

be a holdover from the 'bricks-and-mortar' library where users who have difficulty finding resources can approach the information desk with a question. Users of digital search systems do not have this option. A second method for targeting search results is to create a search engine which is 'smarter' or more able to determine what data the user is looking for. Some systems are using a technology called the Semantic Web to provide this assistance.

2.3 Semantic Web

Before looking at this technology in detail, a consideration of the limitations of the current Web technology in regard to searching may be of benefit. Antoniou and van Harmelen (2004) suggest that while keyword based search engines are the main tools for using the Web today, these searches have high recall, but low precision. A vaguely worded query can return thousands of links, but few of these links may contain the data that the user is looking for. The search results are also highly dependent on a precise use of vocabulary. An important consideration here is that the returned set of data is a set of links to resources rather than an answer to the questions itself. The following scenario describes an offline version of this. A student approaches a teacher with a question.

Student: How can I solve a quadratic equation?

Teacher: The answer is in one of these books.

The student in this situation comes away frustrated with the lack of answer to his question. The term 'knowledge management' concerns itself with the acquisition, accession and maintenance of knowledge within an organization (Antoniou and van Harmelen 2004). In this environment, knowledge is viewed as an intellectual asset that can be used to create increased productivity and competitiveness within the company. It is especially important for organizations which have geographically dispersed offices to gain access to these knowledge assets.

2.3.1 Semantic Web potential

A difficulty with current Web searches is that rather than an answer to a query, the search engines provide a list of resources which may contain the answer. Due to the difficulty that computers have in ascertaining the intent of the user in a keyword based search, it is hard for a search engine to provide a small set of relevant and complete responses to a user's query. In effect the computer is 'guessing' regarding what sort of data the user is interested in. Lacy (2005) provides a list of reasons for this difficulty:

1. The current Web does not provide enough structure to allow advanced computer processing of content.
2. Not enough of the information on the Web is connected to other information to allow for complex queries or updates of information.
3. There is a need for better information representations on the Web to enable more advanced applications.
4. There is a need for accompanying information that explains the meaning (semantics) of the information.
5. Semantics are required for the efficient automated interpretation of structured Web content
6. The current Web can be improved with structured information and metadata for finding information and explicit semantics.

This leads us to the Semantic Web. The Semantic Web is an initiative of the World Wide Web Consortium (W3C), a group that provides oversight for standards and protocols for the Internet. This technology was first proposed in 2001 as a way to provide a ‘universal medium for the exchange of data and to give computers as well as humans the ability to process and access Web content’ (De Roure et al. 2005). Berners-Lee, the inventor of multiple Web technologies such as the World Wide Web (WWW or Web), HyperText Transfer Protocol (HTTP) and HyperText Markup Language (HTML) is the lead proponent of this concept and gave the following explanation of its benefits:

If HTML and the Web made all the online documents look like one huge book, RDF, schema, and inference languages will make all the data in the world look like one huge database (Berners-Lee 1999).

This echoes the earlier discussion regarding the differences between digital libraries and databases. In concept, the Semantic Web offers the potential to not only eliminate the differences between these two concepts as well as make it easier to find information on the Web. Berners-Lee states:

The Semantic Web is not a separate Web, but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation (Berners-Lee 2001).

Rather than being centralized as a library information system is, the Semantic Web aims at de-centralization, linking sites together through common languages and rules which aid the computer in deciphering the information in order to produce better results.

To illustrate clearly both the need for the new technology and its possible value to shared data systems, a two examples may prove helpful. In the first case, consider a query that might be run on a current search engine such as Google: ‘how many archaeological sites containing Lapita pottery are there the Solomon Islands?’ This simple question returns a list of nearly 300 links, obviously too large of a return set to deal with. Swartz (2001) gives three possible reasons for the difficulty in running this simple query: 1) current language processing and search technologies are not able to process this type of request, 2) there are Web accessible databases that might have the answer to this question, but a simple text-matching query is not sufficient to calculate the answer, and 3) there is no program currently available that is able to count each rail-line’s presence on the Internet, identify them individually and present a final total to the user. In the situation where a Semantic Web is available with a search engine, Swartz (2001) proposes that the query might return the answers such as those in Table 2.7 below.

1	http://www.lapita.org/sites says that the number of sites is over 500
2	There is a database that can provide that number, but you will need to provide an authorization number
3	There is a Web service that can compute that number, but it will cost you \$50 US for the answer
4	I can get you an approximate answer by search and filtering, but it will take many hours to complete.

Table 2.7 – Potential answers to a query question using the Semantic Web

A second scenario, reveals the potential that intelligent software ‘agents’ could provide via the Semantic Web. Agents, as defined in the Semantic Web arena, are pieces of software that work autonomously and proactively for a particular user. The agents receive tasks and preferences fro the person, and seek information from Web sources, communicate with other agents, compare information about user requirements and preferences, select certain choices and give answers to the user (Antoniou and van Harmelen 2004).

In this example, ‘John’ is on his way to work and receives a phone call that a family member has been taken to the hospital. He tells his computerized personal agent to postpone all his meetings and to get directions to the hospital. Without further instructions, his software agent:

- Queries the local hospitals to find which one the family member has been admitted to
- Queries the MapQuest agent to get directions to the hospital and informs the navigation system on John's car
- Contacts the agent of the person John was to have a meeting with this morning, reschedules the meeting for tomorrow using John's electronic diary and informs John of the change in meeting time (Finin and Joshi 2002)

Rather than requiring human input at every stage of the process, the theory is that software agents will be capable of limited decision-making based on the user's preferences. In order to provide software agents with the information to search out the answer to user's queries, a semantically-rich information base must be established. This requires the set up of systems that maintain the architecture. While the above example reflects a social use rather than an archaeological research setting, by extending the case to consider dealing with intelligent querying agents between archaeological datasets, this can clearly be of value to researchers.

2.3.2 Semantic Web Architecture

Berners-Lee (2001) proposed the layered architecture listed below as the technology and protocols necessary for full implementation of the Semantic Web. In order for a data sharing application to provide a more 'intelligent' querying mechanism it should include these components. The WC3 has published the standards required for all but the top two layers; Proof and Trust. A description of the purpose of each layer follows.

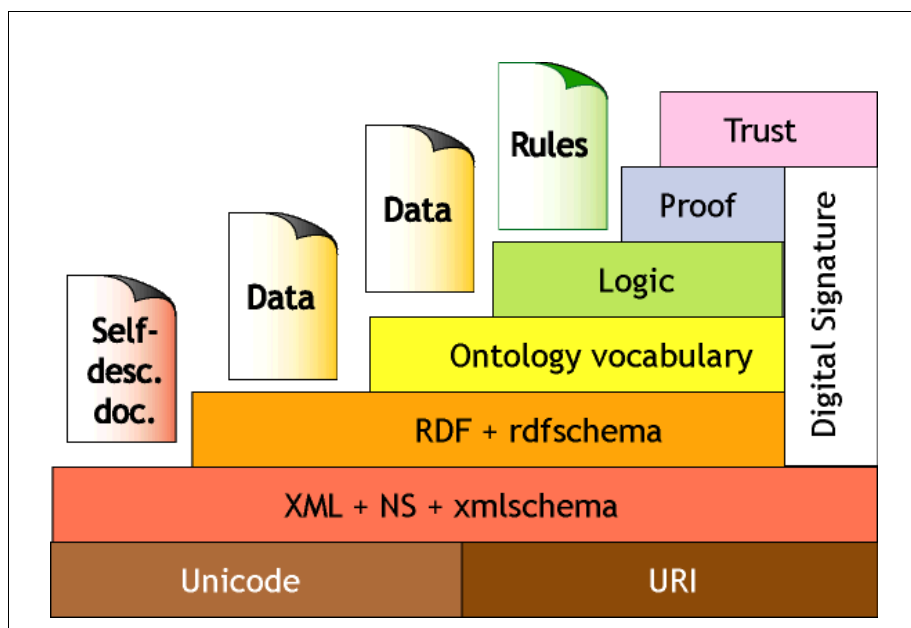


Figure 2.7 - Semantic Web Architecture (Berners-Lee 2001)

Web Interface level – Unicode and URIs

At the most basic level, all Web pages are made up of a combination of text and links. HTML allows document creators to display text, images and other media via a Web browser. Links to other resources are provided via Uniform Resource Locators (URLs). These items also make up the bottom level of the Semantic Web, described by Berners-Lee as Unicode and URIs. Unicode is a system for listing text by assigning each letter or character a separate numerical code. It is an extension of ASCII, and is sufficiently large to handle not only all of the writing systems currently in use, but those from ancient systems as well (see Figure 2.8). Uniform Resource Identifiers (URIs) are a type of URL that identifies resources on the Internet (Figure 2.9). URI creation can be as simple as putting a page on the Internet that lists an item. The creation of URIs is decentralized; i.e. anyone can create one. Due to this lack of control, multiple URIs describing the same resource can be a problem (Swartz 2001). In the case of an e-Research project, the Unicode-URI level can be used to present a search engine to the user that can be used to search across multiple databases.

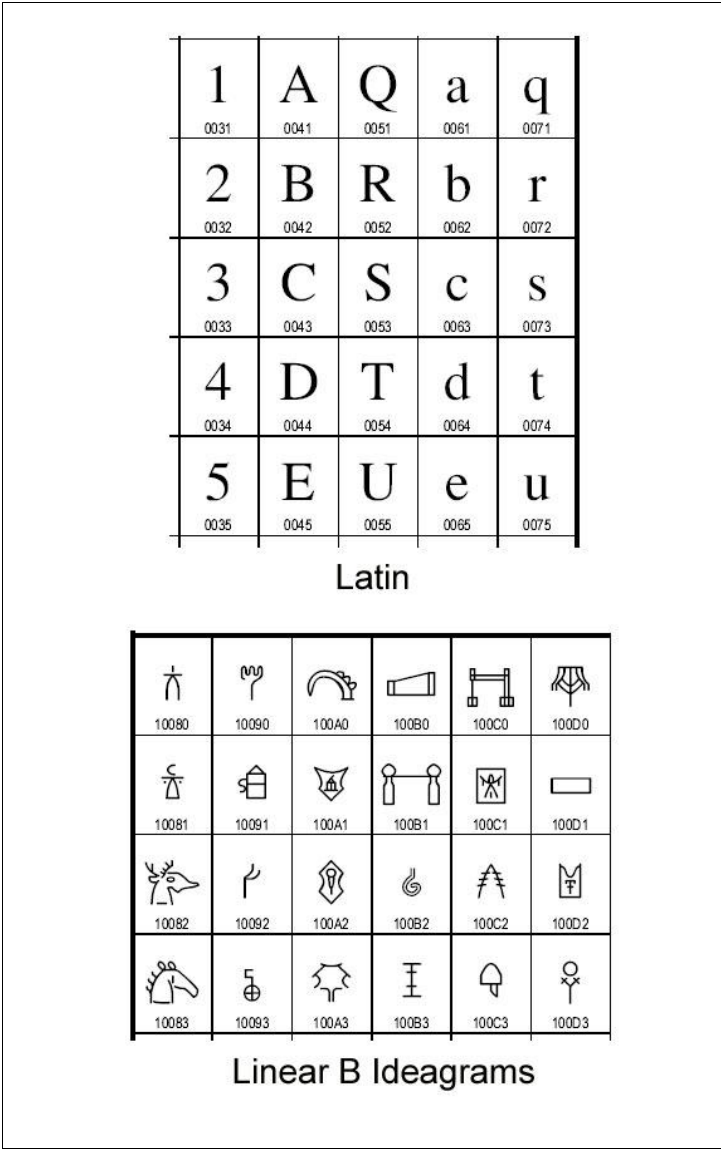


Figure 2.8 – Unicode, a system for listing text on computer systems
 (from <http://www.unicode.org/charts>)

Although the term URL is more commonly used by the general public, in technical literature the term URI is more frequently specified. In either case, URLs are in fact a type of URI, and so the terms can be used interchangeably.

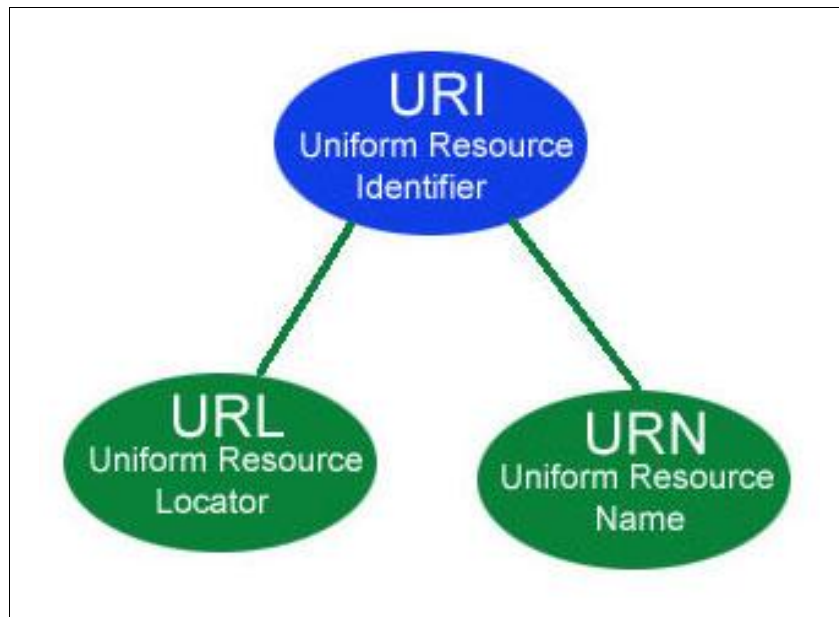


Figure 2.9 – A URI contains the elements URL and URN

XML, Namespace and XMLSchema

XML (eXtensible Markup Language) allows Web page creators to add extra information about the pages through special XML tags. Users can define a simple language that can be used to describe information provided on the page. This information is stored in the XML tags which are defined in the Namespace area that lets the writer specify the meanings of terms used in the tags. For example, 'ship_type' might be an item from a list of ships. XML can be used to separate the data content and underlying structure. XML style sheets are often used to change the formatting of a site design quickly without requiring programmer input. XML supports the exchange of structured information across different applications, and allows the individual Web applications to change the look and feel of the site without changing the data itself.

The XMLSchema describes how the tags and information are used on a particular page. The term schema is often used to describe the structure that makes up an object or database. In the case of a database schema, this would specify all the tables and fields in a database and give information about the type of data stored in them. XML and XMLSchema handle information that gives further detail about the data provided on the page. In data sharing situation, this machine-understandable information could be used to provide extra data about the page to a querying agent (Berners-Lee 1999).

In Figure 2.10 below, sample XML code is given. The first line specifies the version of XML that is being used. The remainder of the text is made up of nested 'elements' or

objects that make up the page. An element has a start tag ‘<recipe>’ and an end tag, ‘</recipe>’, which specifies information regarding an element. In this document there are five elements defined: recipe, title, ingredient, instructions, and step.

```
<?xml version='1.0' encoding='UTF-8'?>
<shipwreck list>
  <Shipname>Mary Rose</shipname>
    <Beam>5 ft </beam>
    <Length>15 ft </length>
    <DateofWrecking>1856</DateofWrecking>
  </Shipname>
```

Figure 2.10 – Sample XML showing defined tags

Within the tag itself is additional information that is available regarding the object. For example, consider the following element:

```
<Shipname>Mary Rose</Shipname>
```

Figure 2.11 – Sample element rendered in XML

The system is defining a tag named ‘Shipname’. Although XML can provide additional information about the content on the page, this information is only understandable to humans. An additional layer, the Resource Description Framework is necessary to provide the translation so that machines such as computers can understand this additional data.

RDF – Resource Description Framework

Powers (2003) states that the Resource Description Framework (RDF) provides a means of recording data in a machine-understandable format, allowing for more efficient and sophisticated data interchange, searching, cataloging, navigation, and classification. A less technical description of this technology would be that RDF allows users to make statements about a resource that are understandable by computer programs. RDF can be used to specify data resources and define the connections between them. It is sometimes called a

data modeling language because a model of the structure of data and interactions between multiple pieces of data can be created using XML.

The key concepts in RDF are defined as resource, property and value. Each resource has a number of properties and values that are associated with that resource. RDF uses XML as the programming language to provide syntactic interoperability. This allows separate applications to use the same ‘syntax’ or word choice in order to communicate with each other. To fully describe a resource in RDF it is necessary to define three properties: subject, property type, and property value. A few potential RDF properties are listed in Table 2.8 below.

Subject	Property Type	Property Value
Ship	Name	Mary Rose
	Built	03/09/1864
	Length	15 ft
	Beam	5 ft

Table 2.8 – Sample RDF properties to define subject ‘ship’

RDF uses triples made up of URIs to define each subject. Since each URI describes a separate resource, this can be used as a mechanism for linking resources. Simple facts can be defined using three specific pieces of information: the subject of the fact, the property of the subject that is being defined, and its associated value. Powers (2003) states that in RDF the subject is the thing being described, a resource identified by a URI, and the predicate is a property type of the resource, such as an attribute, a relationship or a characteristic. In addition to the subject and predicate, the specification also introduces a third component, the object. Within RDF, the object is equivalent to the value of the resource property type for the specific subject. So, a simple statement like ‘the Mary Rose was built in 1864’ could be parsed as seen in Table 2.9 below.

Subject	Predicate	Object
Mary Rose	was built in	1864
Uri='http://maryrose.org	Uri='http://maryrose.org/about	Uri=http://maryrose.org/dob

Table 2.9 – Parsing an RDF tuple

To a computer system not equipped with a semantic translator, this sentence has no meaning. It has no idea who the 'Mary Rose' is, what does ‘was built in’ mean, and what is

the significance of the date? But through RDF each triple can elaborate a definition for each word in the sentence.

The other item at this level on the Semantic Web diagram, RDFSchema (RDFS), is a vocabulary description language that can be used for describing properties and classes of RDF resources. It includes defined semantics that the system uses for generalization hierarchies of such properties and classes (Antoniou and van Harmelen 2004). RDFS is a type specification tool that can be used to arrange each of the triple URI statements in a taxonomic structure that helps computers use the terms and convert between them. For instance, a particular system may define ‘Brigantine’ as a type of ‘Brig’, and state that ‘Brig’ is a subclass of ‘Vessel’. Through simple statements like this, a quite elaborate data structure can be built relatively quickly (Swartz 2001).

RDFS also provides a mechanism for describing specific domains; for example maritime archaeology, geology, or psychology. RDFS contains a primitive ontology language, but in general uses of RDFS are made through more sophisticated ontology languages such as Web Ontology Language (OWL). Ontologies and ontology languages are described in more detail in the next section and in Chapter 3.

Ontology vocabulary

A knowledge management system must have a method of making sense of the multitudes of data contained within its files. The ontology level of the Semantic Web provides a mechanism to create this data organisation. Lacy (2003) proves this definition of ontologies:

An ontology specification is a formally described, machine-readable collection of terms and their relationships expressed with a language in a document file. A conceptualization refers to an abstract model of a domain that identifies concepts.

Based on this definition, Lacy states that an ontology specification (description) must provide the following items:

- A formal description of the ontology (i.e. follows a prescribed syntax)
- The description must be machine-readable – understandable by computers
- It should include a list of terms and their relationships to each other
- The ontology should use an ontology language – such as OWL
- The description should be stored in a document file
- This document is an abstract definition of a real world domain

Ontologies describe relationships between items in a controlled vocabulary list. This provides the basis for information about the terms, and facilitates search results. A controlled vocabulary is defined as a restricted list of terms that have a specific predefined meaning. To put this into basic terms, consider the situation of a car owner who takes his car to a mechanic for service. After a thorough check of the automotive systems, the mechanic presents the customer with a bill of \$250 for repair of the car's propeller. The customer would rightly question the authenticity of this repair due to the fact that a propeller is not on the list of items installed on an automobile. Figure 2.12 contains a simple example of an ontology from the field of maritime archaeology.

To frame the discussion of ontologies in regard to Archaeology, many ontological descriptors already exist that need to be used correctly in order to provide reasonable search results. The research in this project will use descriptors from the English Heritage National Monuments Register as well as terms used in the Australian National Shipwreck Database. An important focus of this project is not just on simple searching, but on interacting with the data and drawing conclusions regarding complex relationships between data sources. The Semantic Web requires ontologies as a basic building block of the system (De Roure et al. 2005). Ontologies can also specify that words are related, or have some relationship to each other. So as the system grows, new terms can be added or extended as necessary.

Class-def <i>ship</i>	% ship is a class name
Class-def <i>brig</i>	% brig is a class name
Subclass-of <i>ship</i>	% it is a subclass of another class
Class-def <i>brigantine</i>	% brigantine is a class name
Subclass-of <i>ship</i>	% also another subclass of ship

Figure 2.12 – Sample light-weight maritime archaeology ontology

Logic, Proof and Trust

A logic statement, as it applies to a computer system is a fact that is irrefutable. At this level of a semantically-aware system, the computer can make assumptions or simple inferences based on these logic statements to generate data. As an example, consider a

four-point academic grading system such as is present in the U.S. The top score, an 'A', is equivalent to 4 points. The lowest passing grade is a 'D' which attracts only 1 point. Each student is assessed a grade for each subject based on these rankings. At one university, any student receiving a perfect 4.0 average score for the semester is sent a letter congratulating them for being placed in on the 'Deans List' for top scholars. The computer can use this simple rule to iterate through the list of students, and send each student who accrued a score of 4.0 at the end of the semester a congratulatory letter. In this situation, the system created new data by computing a list of the top achievers.

Using this academic marking system, assume that a student called the Dean's office to complain about not receiving a congratulatory letter for his grade this semester. Using a proof algorithm, the system could calculate his score, prove that it did not meet the requirements for placement on the Dean's List, and advise the student accordingly. In this way, the system is acting as an information processor and not just a search tool. Although such a system may be easy to describe, it can be quite complex to implement. In some cases proving a scenario may mean searching through thousands or millions of URIs. The W3C has not yet provided a specification for developing Proof in a Semantic Web.

Consideration of data interactions on the Web leads ultimately to a question of trust. If URIs can be created by anyone who published a Web page on the Internet why should we trust them (Berners-Lee 1999)? A way around this is the use of digital signature. Data creators sign a statement electronically that states that their data can be trusted. If you trust that entity, and that person trusts another group of people, eventually a 'Web of trust' is developed. This is key to establishing a safe environment for the transfer of data (Berners-Lee 2001).

Introduction to metadata

Although it is not mentioned specifically in Berners-Lee's Semantic Web hierarchy, metadata is a key component of ontologies. Literally meaning 'data about data', metadata is information about a resource. It documents the administrative, descriptive, preservation, technical, usage history and characteristics of the resource (Hunter 2003). Until the mid-1990s, the term was used primarily by groups that managed access to geo-spatial data through data management systems. In its present usage the term means 'the sum total of what one can say about any information object at any level of aggregation' (Gilliland-Swetland 2000). Metadata is used in digital library systems to describe, locate and track the usage of information resources. In libraries, metadata is used to provide access to content. The metadata may be used as indices, abstracts, and catalog records produced in accordance

to cataloging rules as well as structural and content standards. Directories and search engines, both in digital libraries and on the Web, utilize metadata to aid users in the discovery of resources.

The most commonly used standard for metadata is called Dublin Core (Dublin Core Metadata Scheme). Dublin Core is a small, simple set of fifteen metadata elements that can be used to describe resources to allow searches across a variety of information sources on the Internet. The elements are listed in Table 2.10 below. The metadata terms were defined by a group of librarians, information professionals and subject experts. The guiding principles specify that:

- the elements must be simple and easy to use
- it should not require extensive training to use
- every element is optional and repeatable
- the elements should be international and cross-disciplinary in scope and applicability
- the element set should be extensible to allow discipline or task-specific enhancement
- the primary use of the element set is to be for embedded descriptions of Web resources that can be accommodated within the HTML meta tag (Gill 2000).

There have been a number of large-scale development using Dublin Core metadata internationally. Some of these include the Australian Government Locator Service and the CCTA (Central Computer & Telecommunications Agency) Government Information Service in the UK (open.gov.uk). James Cook University uses Dublin Core metadata that is dynamically added to the header of each page as it is served to the website viewer (Figure 2.13). Further information about metadata is presented in Chapter 3.

While the inter-related components of the Semantic Web architecture can be complex to implement, it should not be considered Artificial Intelligence. In support of this point Berners-Lee (1998) states:

The concept of machine-understandable documents does not imply some magical artificial intelligence which allows machines to comprehend human mumblings. It only indicates a machine's ability to solve a well-defined problem by performing

well-defined operations on existing well-defined data. Instead of asking machines to understand people's language, it involves asking people to make the extra effort.

This begs the question: what benefit will the average user derive from such a system? Based on current research it appears that it should aid in transferring communication for the user between all of the devices that are currently in use from PDA, laptop, desktop and server. Corporate decision information systems should derive the benefit of automating systems that previously required human intervention. Digital libraries and other data combinatory systems should be able to assess the trustworthiness of the persons accessing data, as well as providing better search facilities in order for patrons to find the answers to their questions.

2.3 Specifics regarding a data sharing system

This chapter has detailed the current situation regarding data sharing in maritime archaeology and some technologies that appear to present a method for solving some of the associated problems dealing with data sharing. At this point, a few questions should be

Dublin Core element	Definition
Contributor	An entity responsible for making contributions to the resource.
Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Creator	An entity primarily responsible for making the resource.
Date	A point or period of time associated with an event in the lifecycle of the resource.
Description	An account of the resource.
Format	The file format, physical medium, or dimensions of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Language	A language of the resource.
Publisher	An entity responsible for making the resource available.
Relation	A related resource.
Rights	Information about rights held in and over the resource.
Source	The resource from which the described resource is derived.
Subject	The topic of the resource.
Title	A name given to the resource.
Type	The nature or genre of the resource.

Table 2.10 – Dublin Core metadata identifiers

```
<html lang="en" xmlns=http://www.w3.org/1999/xhtml>
<head>
  <title>
    Welcome to James cook University in Tropical Northern
  </title>
  <meta http-equiv="Content-Type" content="text/html; charset=iso
  <meta name="DESCRIPTION" content="Welcome to the James Cook
  <meta name="KEYWORDS" content="James Cook University, QLD
```

**Figure 2.13 – Dublin Core metadata usage on a university Web site
(<http://www.jcu.edu.au>)**

answered regarding the importance of this research, search engines in general, and the functionality desired in an archaeological data sharing system.

Why is this research necessary?

At the heart of any archaeological investigation is a research plan. The research plan for a project should include all pertinent information regarding the research question. A maritime archaeological researcher should obtain all of the known data regarding previous research and oral histories concerning the research question as well as site information such as location, possible impacts upon the site, ocean currents, weather and maps of the site. All of this data is located in separate places and very often may not be in digital form. The collection of this data adds to the length of time needed to explore any research question. If this data were available online, the time required for this step in the process would be dramatically shortened (Jeffrey pers comm).

Project plan

- Archive research and oral histories
- Site info such as location, impacts on site, currents, weather, maps
- Research question being investigated
- Fieldwork plan
- Resources necessary for survey, excavation, conservation
- Publication plan

Table 2.11 – Typical components of an archaeological project plan

Once the project has been completed the researcher should make available to others all of the previous data plus details concerning the research question under investigation, the fieldwork plan, details concerning the survey, excavation and conservation of data from the site and any publications concerning the project. The data from the project is often placed in a site or artefact database which should be made available to other researchers. The original researcher needs to provide a central storage location for this data, make the data available for research and education purposes and sometimes provide data useful in court

cases pertaining to enforced protection of sites (Jeffrey 2004 pers comm). Table 2.12 below lists a few of the items that may be found in a site or artefact database.

<p>Site/artefact database</p> <p>Purpose</p> <ul style="list-style-type: none">• Provide central storage location for data regarding site and artefacts• Make data available for research and education purposes• Provide data useful in court cases enforcing protection of sites <p>Contents</p> <ul style="list-style-type: none">• Excavation details, forms, artefact listings• GIS data, contour maps, site sketches, artefact sketches• Digital photographs of site and artefacts

Table 2.12 – Contents and purpose of a site/artefact database

Many hours of research time are currently devoted to the process of obtaining information for the research plan and then making the data available to other researchers once the project has been completed. Although a researcher may have moved on to other work, distribution of the data is dependent on that researcher answering requests for data in a timely manner. This places an extra burden of work on the archaeologist that cuts into other projects. Thus dispersal of information becomes limited by the available time of the original researcher. Assuming that the researcher finds the time to distribute this data, a further problem is caused by inability of email to handle large data files. All of these items make a data sharing system necessary.

System functionality

Based on the research needs listed above it become apparent that a search engine connected to a data sharing system is needed in order to make it easy for archaeologists to locate data. This concept can be described as ‘Googling for archaeology’. Once the system has been loaded with the location of the data, it could provide researchers access to the data on an ‘as needed’ basis. This would provide an access point to the data from multiple locations, reducing the time required to locate and obtain data.

As discussed in Chapter 1, in order to allow maritime archaeologists to share data the system should provide the functionality listed below.

ID	Functional requirements for the system
1	User is able to deposit a dataset
2	User is able to set access rights for data
3	User is able to make data discoverable
4	User is able to find data
5	User is able to view data online
6	User is able to download data
7	System is easy to use
8	System allows searches across multiple datasets
9	System returns search results in a Web browser

Table 2.13 – Functionality requirements for maritime data sharing system

This chapter has summarized research and technology issues associated with sharing data. A distributed data model using federation of maritime archaeological data sets has been chosen for this project. This research is using techniques associated with e-Research such as the middleware tool SRB as well as components of the Semantic Web. The following chapter will provide an in-depth look at the specifics of data sharing in regards to metadata and ontologies that are needed in order to support the requirements of a data sharing system.

Chapter 3 – Tools of the semantic trade: metadata and ontologies

The previous chapter provided an overview of the data-sharing situation in maritime archaeology and an introduction to the concepts of federation and semantic search. In this chapter a closer look is taken at two of the key building blocks of these technologies: metadata and ontologies. Metadata provides a means for the user to discover data, and ontologies allow connections to be made between data sets. These ideas are explored in more detail in this chapter, and are examined in relation to how they can be used to create a maritime archaeological data sharing system.

3.1 Metadata

Metadata provides additional contextual information regarding data. This information can be used to help discover resources that match a user's query. As an example, consider the following number: '5038385413'. Without any additional information this is a meaningless piece of data. However, suppose the metadata informs the querying system that this is phone number. In this case, the system might display the number in its usual format: 503 838-5413. The metadata also contains the information that this is the phone number of a pizza restaurant. If the user was interested in ordering a pizza and having it delivered, then this is useful data, otherwise it is meaningless trivia. This is the benefit of metadata; it provides information regarding the usefulness of data. Although metadata was considered briefly in the previous chapter, in the next section it is described in detail due to its importance in data sharing systems.

3.1.1 Metadata tags

The HTML specification followed by all Internet web browsers allows the use of 'tags' which can define and describe the content of a web page or site. The Internet search engine Alta Vista was the first to popularize the use of the meta tag in their search algorithms. The plan was for the keyword 'metadata' in HTML code to be used to specify the content contained in the page in order to provide more effective retrieval and relevance marking in searches. The 'description' tag would be used in the display of search results to provide a more 'user friendly' listing of the contents of the page (Gill 2000). In the financially competitive arena of the web, some site creators exploited these tags in an effort to artificially boost the relevance of their site in the result set. Meta tag 'spoofing'; i.e. repeating keywords hundreds of times became common and search engines were forced to abandon keyword indices as a method of relevance ranking. At the present, the TITLE keyword and the textual contents of a page are the most significant factor in the ranking of result sets.

3.1.2 Why is metadata important?

The use of comprehensive metadata provides many benefits. The primary benefit is increased accessibility of resources via search engines and other finding tools. However, corollary benefits are found in the areas of content management and system maintenance. Through metadata use, system managers can track relationships between the resource item, the host collection, and its users. Multiple versions of an item are often created in digital systems, and metadata can aid in keeping track of these duplicates. Rights access and other legal issues can be controlled and tracked through updateable metadata elements, as well as tracking system usage information which can be used for load balancing and other system maintenance issues. In multiple ways the use of metadata on distributed systems increases the accessibility of resources to a widely dispersed population of users (Gilliland-Swetland 2000).

3.1.3 The purpose of metadata

Metadata can be broadly explained as information which identifies and describes an information object. In general, it defines how an object behaves, its function and use, its relationship to other information objects and how that object should be managed. Gilliland-Swetland describes 5 tasks that metadata performs:

1. Certifies authenticity and degree of completeness of content
2. Establishes and documents context of content
3. Identifies and exploits structural relationship between and within information objects
4. Provides a range of access points for uses
5. Provides some of the information an information professional might have provided in a physical reference or research setting (Gilliland-Swetland 2000).

In this setting, *content* refers to what the object contains or is about (internal characteristics), while *context* describes the who, what, where, why, when and how aspects associated with the object's creation (external to the object). Table 3.1 describes the different categories of metadata that are generated in a digital library system, and Table 3.2 describes the attributes and characteristics of each type of metadata. At each stage in the lifecycle of an information object, metadata elements are added and/or modified. Therefore, metadata continues to accrue during the lifetime of an object (Table 3.3).

Type	Definition	Example
Administrative	Managing and administer information resources	Acquisition information, access rights, location data
Descriptive	Describe or identify information resources	Catalog and other finding aids
Preservation	Preservation management of information resources	Physical conditions, documentation of actions taken to preserve resource
Technical	How the system functions or metadata behaves	Hardware/software documentation, system response tracking, security data
Use	Level and type of use of information resource	Exhibit records, use and user tracking

Table 3.1 – Metadata categories (after Gilliland-Swetland 2000)

Attribute	Characteristics	Example
Source	<ul style="list-style-type: none"> - Internal metadata agent created - External metadata about the object that is added later 	<ul style="list-style-type: none"> - for information object when it is first digitized - registration, access rights, legal information
Method	<ul style="list-style-type: none"> - Automatically created by computer - Manual metadata created by humans 	<ul style="list-style-type: none"> - transaction logs and indices - descriptions, Dublin Core
Nature	<ul style="list-style-type: none"> - Lay metadata - Expert metadata 	<ul style="list-style-type: none"> - not subject experts - subject matter experts/admin
Status	<ul style="list-style-type: none"> - Static: never changes - Dynamic: may change with use or manipulation of an information resource - Long term: ensures that resource continues to be accessible and usable 	<ul style="list-style-type: none"> - Title, Date of Creation - user transaction logs, annotation - technical format, processing information, rights preservation
Semantics	<ul style="list-style-type: none"> - Controlled: conforms to standardized vocabulary - Uncontrolled – does not conform to standardized vocabulary 	<ul style="list-style-type: none"> - AAT, ULAN - Free text note fields, HTML meta tags
Structure	<ul style="list-style-type: none"> - Structured: conforms to a standard - Unstructured: does not conform to a standard 	<ul style="list-style-type: none"> - MARC, TEI, EAD - Unstructured note fields
Level	<ul style="list-style-type: none"> - Collection: relates to the collection - Item: relates to specific resource 	<ul style="list-style-type: none"> - Specialized index - Format information

Table 3.2 – Metadata characteristics (after Gilliland-Swetland 2000)

Phase	Activities and Type of Metadata Created
Creation & Multi-versioning	Object enters information system by being created digitally, or converted into digital format Multiple versions may be created Administrative, Descriptive metadata created
Organization	Objects organized within the structure of the system Registration, cataloging and indexing metadata added
Searching & Retrieval	Stored and distributed objects subject to search and retrieval by users Metadata added to track retrieval algorithms, user transactions, and system effectiveness
Utilization	After retrieval objects may be used, reproduced, and/or modified Metadata added for user annotations, rights tracking and version control
Preservation & Disposition	Objects are subject to refreshing, migration to other areas, and integrity checking to ensure continued availability Inactive items which are no longer needed are discarded Metadata is added to document preservation and/or disposition actions

Table 3.3 – Resource object lifecycle (after Gilliland-Swetland 2000)

3.1.4 Metadata standards

Standards for metadata creation are necessary in order to provide reliable and efficient retrieval of data resources. These standards tend to be geared toward one target community such as EAD (Encoded Archival Description), adopted by the Society of American Archivists in 1999, or sufficiently broad in nature in order to be used by any community (Dublin Core). While establishment of standards that apply solely to one community have been successful, large scale adoption of metadata standards which bridge groups has proven more difficult to achieve. Dublin Core, although widely discussed, has had a very poor uptake by Internet resource creators. According to Gill, less than one percent of web documents utilize Dublin Core (Gill 2000). Approximately 34% of resources available on the web provide metadata of some kind, although it is often highly unstructured. It should be noted that very few search engines utilize Dublin Core metadata, and most ignore the tags entirely (Oppenheim et al. 2000).

Safari (2005) echoed this finding with the report that search engine results are not appreciably affected by the addition of metadata information. Since the original publication of the Dublin Core standards, limitations in the original system have led to a series of extensions and modifications to the format. The Warwick Framework and Interoperability Qualifiers were added in order to stretch the applicability of Dublin Core to diverse

environments. In addition, Dublin Core researchers discovered the 1:1 Principle that states that the most robust system allows the use of separate metadata sets for each item and to describe the relationship types between them using an enumerated list of relationship types. This in theory would allow groups who don't use the same controlled vocabulary to reference resources on each other's systems. Unfortunately, applying the extended Dublin Core element set has proven problematic. The CIMI (Consortium for the Computer Interchange of Museum Information) conducted a three year-long investigation into the utility of using Dublin Core metadata. While they reported that unqualified (original) Dublin Core could be used as an effective tool for discovery of museum information resources in a cross-disciplinary networked environment, CIMI stated that it could not recommend the use of the qualified (extended) Dublin Core for information interchange between museums because it would not allow the museums to specify complex descriptions of the resources. The underlying data model for the Dublin Core element set, which was originally designed for the description of text-based web resources, was inadequate to handle multi-media resources (Gill 2000).

3.1.5 Metadata harvesting

Although metadata as used on the Internet is still in the development stages, it should not be assumed that it presents no value. The original Dublin Core element set provides a robust tool for data discovery. One area where the use of metadata thrives is metadata harvesting. In an effort to navigate the labor-intensive problem of human-monitored metadata generation, the Open Archives Initiative (OAI) suggests that data providers make metadata about their collections available for automated 'harvesting' through an HTTP-based protocol. Data providers are required to supply metadata that complies to a common schema, such as the unqualified (original) Dublin Core Metadata Element Set. Table 3.4 lists the methods used to harvest metadata.

Metadata harvesters gather information pertaining to particular search criteria. Instead of returning one set of results, a metadata harvester could return a different set of data each day. This is frequently seen in news services such as Lexis-Nexus, Yahoo or MSN. The user selects a topic; perhaps low airfare fares to a popular vacation area. Each time a news item is found that meets these criteria, the user receives an email with this data. This method is also used in a technology called RSS (Really Simple Syndication), where news stories or web log entries are distributed to interested subscribers. Software applications called 'aggregators' are used to gather these resources and to provide a convenient way for

the user to retrieve the results. Planet is an example of this type of aggregator (<http://www.planetplanet.org>).

Type	Definition
Automatic document indexing & classification	<ul style="list-style-type: none"> - Assigns document into subject categories - Scans document and analyse the frequencies of the patterns of words - According to a taxonomy assigns the document to a particular category in the taxonomy - Still requires a human to check the categorization, but is 90% accurate
Image indexing	<ul style="list-style-type: none"> - Semantics-sensitive machine and automatic linguistic indexing in which the system is capable of recognizing real-world objects or concepts - Similar system used by Google Images
Speech indexing & retrieval	<ul style="list-style-type: none"> - Speech recognition systems generate searchable text transcript that is indexed to time code on the recorded media - Users can search text and jump to the applicable section of the audio
Natural & spoken language querying	<ul style="list-style-type: none"> - Allows voice recognition systems to query databases using regular language (natural language queries)
Video indexing & retrieval	<ul style="list-style-type: none"> - Systems parse video and segment it into easily searchable database entries - System can extract named entities from transcripts of the video which can be used to produce time and location metadata

Table 3.4 – Metadata harvesting, after Hunter 2003

Another example of metadata harvesting is offered by the Australian Digital Theses (ADT) program, which was mentioned in Chapter 2. The ADT provides a means to search through the collected doctoral and masters theses from universities in Australia. The data is made available to the ADT via the use of a metadata harvester. Each university maintains a web site where the digital versions of these theses are stored. In general each thesis consists of a number of PDFs (Portable Document Format). The host university must add a set of metadata elements to each file in order to allow the ADT to ‘harvest’ or select these documents for storage in its repository. Each evening, a software application on the ADT site examines the web sites of the universities, and searches for new documents to add to their data store. When a new document is found, it is ‘ingested’, or copied into the storage repository for the ADT. In this manner, new data is added to the ADT system without requiring the manual input of a human. Software applications such as EPrints (<http://www.eprints.org>, see also Guy et al. 2004) make it possible for librarians at a

university to add new documents and their associated metadata without requiring technical assistance.

3.1.6 Metadata and multimedia

Audio-visual content requires that the host system provide some kind of interpretation and processing in order to generate metadata without requiring human intervention. MPEG-7 and MPEG-21 were developed in order to standardize the creation of metadata for multimedia resources. MPEG-7 is a multimedia content description interface, while MPEG-21 is a multimedia framework. MPEG (the Moving Pictures Expert Group) designed MPEG-7 as a standard for describing multimedia content. The goal is to provide a set of tools that will enable both humans and computers to generate and understand audio-visual descriptions. This metadata will allow the fast and efficient retrieval from digital archives as well as allow the filtering of streamed content over the Internet. It uses a pre-defined list of Descriptors and Description Schemes that can be used to enable descriptions of multimedia content. MPEG-21 was developed to provide the technology necessary to allow users to exchange, access, retrieve, trade and otherwise manipulate multimedia digital items (Hunter 2003). Multimedia presents a challenge to metadata developers in that many of the formats are either proprietary or self-contained (Hunter 1998, van Beek 2003).

3.1.7 Problems with metadata research

Metadata research as applied to the Internet environment faces a number of problems. One over-whelming difficulty is the sheer volume of resources that are multiplying at an exponential rate. In 2000, the NEC Research Institute calculated that there were over one billion unique, indexable documents on the Internet (Gill 2000), and as of June 2005, Google claimed to have indexed over 8 billion records (<http://www.google.com>). In order to find resources in this vast sea of data, most Internet users access web directories or search engines. Directories can provide good search precision for broad subject areas, but prove costly and labor-intensive if they must be overseen by a real person as opposed to an automated information system. In addition there are granularity issues: should the system index a web site as a whole, or the individual pages? Search engines utilize 'spiders', automated systems that follow links through the Internet to discover resources. As the size of the data increases, keeping the links up to date is a continuing problem.

A growing percentage of the web is provided via dynamic page generation, rather than discrete, separate pages. For example, if you type a book title into a textbox on a commercial site such as Amazon.com, the applicable information concerning that book will

be sent to a page which shows the results of a database query. This allows large companies to avoid having to build an individual page for each item. The data forms part of the 'hidden web' that is not indexed by search engines (Gill 2000). Another issue is search engine partiality. Search engines, often based in the U.S., are more apt to index US sites than those outside the US., and are more likely to index commercial sites than educational ones. However, as additional countries establish more of a web presence this problem is diminishing. Even with those limitations, only about 40% of existing web sites have been indexed. (Laurence and Giles 2001). The Web appears to be too large for any one group to index successfully. Due to the immense size of the World Wide Web there is a need for widespread adoption of standards for metadata structure, content and authentication to make the Web self-documenting (Gill 2000).

3.2 Ontologies

One of the key building blocks of any digital library or data sharing system is the ontology that is used to mandate information structure and organization. The word originated in Greek metaphysics as the study of what kinds of things exist; onto (being) and logia (written or spoken discourse). It has been used for many years by the Artificial Intelligence (AI) information technology community to mean a description of a particular domain (a video store for example), the entities that inhabit it (videos, customers, clerks) and their interactions.

3.2.1 History of ontologies

Thomas Gruber, an AI specialist at Stanford University states that an ontology is 'the specification of conceptualizations used to help programs and humans share knowledge.' (Gruber 1993). To expand this further, it is a set of concepts used to create a true-to-life representation of a domain (Finan et al. 2002). Ontologies are used currently in many disciplines. Examples include the creation of medical guidelines for managing patient health (DMSO-IV for psychiatric evaluations), mapping the genomes of plants and animals, automated exchange of information among commercial trading partners (EDI, electronic data interchange), and in a digital library to provide access to public information resources . Ontologies are increasingly being used as the foundation of classification systems, databases and software applications (Swartz et al. 2001).

Digital libraries use ontologies to structure and classify data. Automatic document indexing in some cases can replace the manual standard bibliographic classifications (Smrz et al. 2003). Another aid to the classification of resources is the use of synonyms,

hyponyms and meronyms. It is quite effective to have a search automatically broadened by a similar term, a more specific term, or a term that names a part of a larger whole when the user is allowed to choose the type of expansion. For example, a hyponymic expansion would be the addition of ‘body of water’ to ‘lake’. In a meronymic expansion, ‘hat’ would be suggested for ‘brim’ or ‘crown’. These types of extensions can be very useful in situations where there are orthographic inconsistencies such as differences in the spelling of a person’s name. For example, the Library of Congress lists over 45 separate spellings for name of the Libyan leader Muammar Quaddafi. Choices range from Moamar al-Kaddafi to Gheddafi MuAmmar. Additional extensions can be provided by listing related-to or see-also links on search results (Smrz et al. 2003). Examples from archaeology might include expansions of artefact to artifact and other spelling differences between commonly used words.

Type of extension	Description of extension
Synonym	Different words with similar or identical meaning (Automobile, car)
Hyponym	A specific word belonging to a broader category (Automobile, vehicle)
Hypernym	A general word denoting a category of more specific words (Fruit, apple)
Meronym	A word denoting that it is a piece of something else (Finger, hand)
Holonym	A word describing a whole which contains other parts (Hand, finger)

Table 3.5 – Possible extensions to words used in search criteria

3.2.2 From Dewey to Google

Ontologies can provide many benefits to a data sharing system, but they do present some potential drawbacks. Ontologies often rely heavily on categorization or classification of data. This entails the organization of the set of concepts by their relationships to other items. The Dewey Decimal system, developed in the 1870s by Melvin Dewey, was one of the first library systems to categorize all books based on a hierarchical cataloging scheme. It assumes that for every book or resource there is a proper numerical classification for it. For instance, consider the categorization of religions of the world according to the original Dewey Decimal system:

Dewey	200: Religion
210	Natural theology
220	Bible
230	Christian theology
240	Christian moral & devotional theology
250	Christian orders & local church
260	Christian social theology
270	Christian church and history
280	Christian sects & denominations
290	Other religions

Table 3.6 (Dewey's Category 200, after Shirky, 2005)

Political and religious issues aside, this indexing scheme with its emphasis on Christianity over other religions reflects the time and culture in which it was created. This illustrates one difficulty with cataloguing schemes; once a new index is created, the rest of the items need to be reshuffled to accommodate the new system. What is being optimized are the number of books on the shelf, and not the data that makes them up (Shirky 2005, Smith et al. 2001).

To consider this point, let us explore the development of two very different search engines: Yahoo and Google. Yahoo was the first substantial effort to bring order to the web. Yahoo created a list of the items available which grew into an ordered hierarchy with categories. As the list became unwieldy the team hired an ontologist to divide the hierarchy into categories and subcategories found today on the site. The trouble with this scheme is that the categories do not always correspond in the way that users expect .

For example, assume a user wants to find a bookshop. A quick click on the subcategory entertainment shows a list of links including '[Books and Literature@](#)'. The ampersand symbol indicates that this is not a link but another subcategory on Yahoo. Entertainment is not the proper place for books, but Yahoo shows the link to point the user in the right direction. Once on the Books and Literature subcategory, the user is presented with yet another link which lists '[Booksellers@](#)', since booksellers are a commercial enterprise and are categorized as a business. After clicking on the Booksellers@ link, the user is provided with a list of over 40 categories of bookstores, and still has not found a page listing a local shop.

In contrast, Google presented a search box and a Go button. The results were returned quickly as an ordered list. The user was not forced to browse through unrelated categories

that have nothing to do with the current search. Due to this benefit, Google has been quickly adopted as the de facto search tool of the Internet. Yahoo later added a search engine to their portal, but by this time the damage (i.e. loss of customers) was already done. Shirky elaborates this difference in search methodology as ‘browse versus search’. In his view, the browse method requires the people doing the categorization to organize the world in advance. The views of the cataloguer override the users needs and the users view of the world. In the search methodology, the reverse is true. A search-based scheme lets the user state what is needed at the moment of the request. The links are presented without a need for a hierarchy. Shirky states:

One of the biggest problems with categorizing things in advance is that it forces the categorizers to take on two jobs that have historically been quite hard: mind reading, and fortune telling. It forces categorizers to guess what their users are thinking and to make predictions about the future (Shirky 2005).

This is not to say that categorization never works. In a situation where the body of data is small, where formal categories have been established and both the cataloguers and users are experts, this scheme has the potential to work well.

Categorization Works	Categorization Doesn't Work
Domain: <ul style="list-style-type: none"> - small corpus - formal categories - stable entities - restricted entities - clear edges Participants: <ul style="list-style-type: none"> - expert catalogers - authoritative source of judgment - coordinated users - expert users 	Domain: <ul style="list-style-type: none"> - large corpus - no formal categories - unstable entities - unrestricted entities - no clear edges Participants: <ul style="list-style-type: none"> - uncoordinated users - amateur users - naïve catalogers - no authority

Table 3.7 – Categorization effectiveness (after Shirky 2005)

3.2.3 Failed ontologies – the metric system

Before turning to the benefits that ontologies can provide to a data sharing system, one example regarding the difficulties presented by ontologies should be examined. The ‘metric system’ or more properly the International System of Units was first proposed by Gabriel Mouton in 1670. It was adopted by France in 1795 and the United States in 1866.

The system gained international status with the signing of ‘The Convention of the Meter’ in Paris, France in 1875. The U.S. was one of the original 17 signatories to the agreement, and is the only industrialized nation that still does not use the system (World Factbook 2006).

The measurements that comprise the metric system make up a basic ontology because relationships between items are clearly defined. It is not clear why such a logically laid out and simple measurement system was not universally adopted, but a resistance to change seems a likely culprit. Harris (2006) states, ‘When an established ontology is challenged by a newer ontology there will be resistance to the new ontology no matter how good it is.’ He goes on to posit that economic protectionism of businesses in the U.S. led to failure in the adoption of the system. Companies were not forced to retool their plants and workshops to meet the new metric standards, and so there was never a substantial enough reason to make the change. This hesitance to adopt a new model has resulted in mechanics needing to maintain two sets of tools, one in each system. An interesting side note is the difference in how the ‘old’ system is described. Outside the U.S., the older system is named ‘imperial’ harking back to England. In the U.S. the old system is named ‘standard’, denoting its usage as the status quo.

3.2.4 Characteristics of ontologies

The sorts of ontologies mentioned in the previous two sections are called ‘monolithic ontologies’ because they attempt to force all users to employ one system. Just as one database for all applications becomes untenable, so does one ontology. A solution to this problem is to allow more than one ontology to be used. Ontologies have been around for a long time. What is new about their use in the Semantic Web is the realization that seemingly different systems can be compared from an ontological point of view. Harris defines an ontology as:

A specification of a conceptualisation used by a community of agents to support the exchange and consistent use of information (Harris 2006).

In this definition he is describing an ontology as a contract shared between agents that intend to exchange information. Agents in this setting are small, semi-autonomous applications that work to exchange data in order to perform some action. The focus of their work is on discovering information rather than just data. An old computer adage states ‘garbage in, garbage out’. The job of an agent is to sift through the ‘garbage’ of extraneous data and find information that is useful. Using a small, but well-defined ontology, agents or

data sharing systems can exchange vast quantities of data and interpret it to gain information. By following the logic and rules in place on the Semantic Web, new information can be inferred. In this environment the value of an ontology can be judged by its usefulness in the exchange of information.

Four characteristics differentiate ontologies: level of description, conceptual scope, instantiation and specification language.

Level of Description

Level of description refers to the precision of the entity descriptions in the ontology. Fully elaborated ontologies may allow distinguishing properties or entities to define new concepts or entities.

Conceptual scope

Conceptual scope refers to whether an ontology describes a domain or upper-level relationships. A domain ontology may describe a specific field of endeavor like auto mechanics, while an upper-level ontology may allow the user to express requests in natural language rather than a specific grammar.

Instantiation

Instantiation describes an ontology's terminological and assertional components. The terminological component defines the terms and structure of the ontology's area of interest. The assertional component populates the ontology with individuals or entities that inhabit the domain. Instantiation is the process where a real-world system of data is created from an abstract ontology.

Specification language

A specification language refers to the computer-understandable language used to describe the ontology in the digital library. There is a definite trade-off between language expressiveness (the number of different concepts that are capable of being described) and the computability of the language. A simpler language will be easier for the computer to use and may result in quicker and more reliable search results (Denny 2004, Juhnyoung et al. 2006).

3.2.5 Levels of ontologies

In order to develop an ontology, a group must come to an agreement as to what components make up the domain of interest. This is achieved by creating a simple controlled

vocabulary which often evolves into a taxonomy, thesaurus and finally into a full ontology as the description of the domain grows. Although each of these terms has a different level of complexity and purpose, they are all considered ontologies. A description of each of these terms follows with examples showing the differences between them.

Controlled Vocabulary

A controlled vocabulary is a list of terms that has been enumerated explicitly. The design goal is to have a list that is unambiguous and non-redundant which is available from a secured, vocabulary registration authority. This may be as simple as a drop-down box on a form that allows the user to choose a keyword, rather than typing in free text. An example of a controlled vocabulary is subject headings used to describe library resources. A controlled vocabulary ensures that a subject will be described using the same preferred term each time it is indexed and this will make it easier to find all information about a specific topic during the search process.

Taxonomy

A taxonomy is a collection of controlled vocabulary terms organized into a hierarchical structure. This is the most common form of ontology. The organizational scheme is normally represented as a tree with a single source node or root and requires that all vocabulary items be placed in parent-child relationships. Lacy (2005) suggests that this structure is easier to understand and navigate than more complex information representation schemes. The scientific periodic table of elements is an example of a taxonomy. Another well-known taxonomy is the Linnaean taxonomy which places organisms into seven major divisions, called taxa (singular: taxon). The divisions are kingdom, phylum, class, order, family genus and species. The classification levels become more specific towards the bottom. Table 3.8 shows an example of this system.

Classification	Human
Kingdom	Animalia
Phylum	Chordata
Class	Mammalia
Order	Primata
Family	Hominidae
Genus	Homo
Species	Homo sapiens

Table 3.8 – Linnaean taxonomy

Thesaurus

Many informal ontologies take this form. This type of ontology adds associative relationships to familial ones. Terms can be linked via similarities rather than a direct hierarchy. Dublin Core is indicative of this type of ontology used extensively in the digital library domain (Antoniou and van Harmelen 2004). Another example is the English Heritage National Monuments and Records Monuments Type Thesaurus (<http://thesaurus.english-heritage.org.uk>, see also Doerr 2001).

Formal Ontology

A formal ontology is a controlled vocabulary expressed in an ontology representation language. This language has a grammar for using vocabulary terms to express something meaningful within a specified domain of interest, enabling the description of explicit semantic regarding relations between terms. Enforcement of an ontology's grammar may be rigorous or lax. Frequently, the grammar for a light-weight ontology is not completely specified, i.e., it has implicit rules that are not explicitly documented. Library science makes extensive use of ontologies, developing upon categorization efforts from the last hundred years. An example of this sort of ontology is listed in Figure 3.1 below.

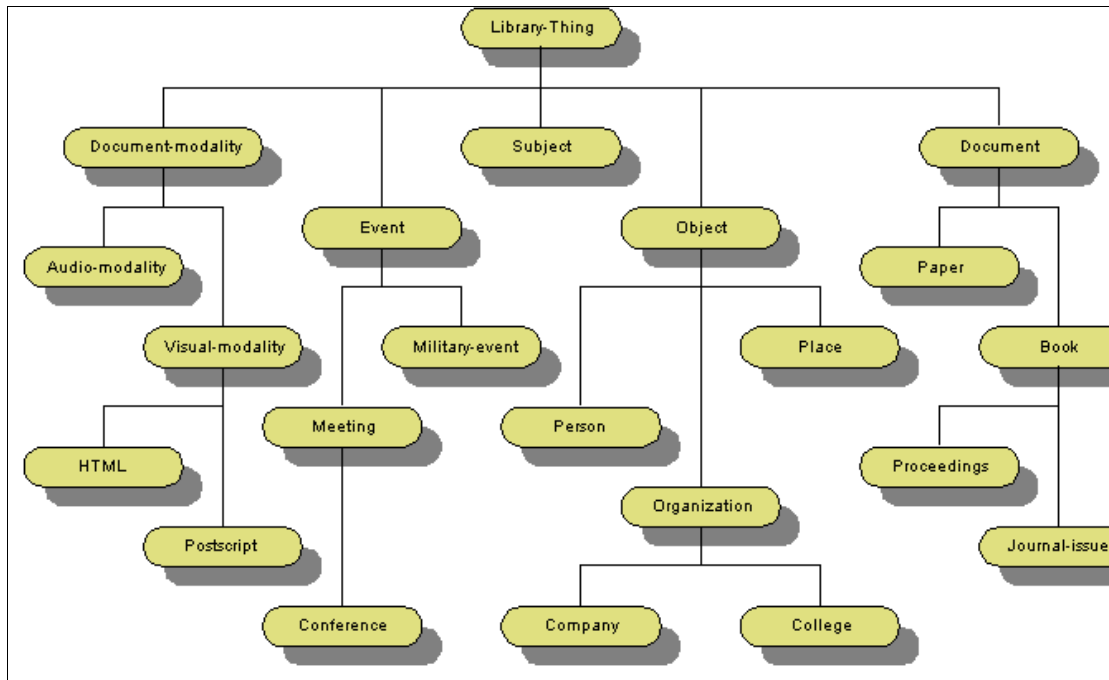


Figure 3.1 – Structure of a proposed ontology for a library information system
(<http://www.cs.vassar.edu>)

3.2.6 Why develop an ontology?

The use of ontologies in digital libraries and for data sharing systems on the World Wide Web is becoming increasingly more common (Harris 2006, Noy 2001). These ontologies range from large taxonomies used to order web sites such as Yahoo to the categorization of products for sale on commercial sites like Amazon. Many disciplines now develop standardized ontologies that experts of a particular domain can use to share information (Noy 2001). In one such scheme in 1998 the United Nations Development Programme (UNDP) and Dun & Bradstreet Corporation (D & B) joined forces to create the United Nations Standard Products and Services Code (UNSPSC), a hierarchical convention that is used to classify all products and services (<http://www.unspsc.org>).

While the motivations for developing an ontology vary from one domain to another, Noy (2001) states that the purpose of a new ontology usually falls under one the reasons listed below.

Sharing common understanding of the structure of information among people or software agents

If different data sharing systems or web sites use the same underlying ontology of terms, then software agents can extract and combine information from these multiple systems.

The agents can then use this data to infer answers to user query or to copy the data into another system.

Enabling reuse of domain knowledge

Rather than ‘reinventing the wheel’ each time a new ontology is needed, researchers can make their ontology available to others, enabling reuse of existing ontologies. This is also useful where several different domains wish to share data. Instead of creating a new ontology that includes terms from both domains, the pre-existing ontologies available for each system can be concatenated into a large system.

Making explicit domain assumptions

Creating a specification for domain assumptions allows researchers to make them available, and allow updates to them as understanding of the domain changes due to new information and technologies.

Separating the domain knowledge from the operational knowledge

By keeping domain knowledge (information about the discipline) separate from operational knowledge (how the data is used) allows the system designer to consider the components of the system separately from the commands needed to assemble them. In theory, a system showing geological information could be re-tooled to show chemistry data if the underlying data sources are changed. In other words, the creation of an ontology makes the data sharing system more generic.

Analysing domain knowledge

Formal analysis of terms in the domain of knowledge is valuable when attempting to reuse existing ontologies and extend them. A clear description of the domain of knowledge assists not only in developing new ontologies but in revealing where gaps in the data exist.

3.2.7 The architecture of an ontology

An ontology is a mechanism that allows agents of other software applications which have been given authorized access to data sources the ability to extract information. Noy (2001) defines an ontology as ‘a formal explicit description of concepts in a domain of discourse’. In order to define concepts in a particular domain or topic the ontology will specify classes, properties, and facets that describe the knowledge base. These terms are described below.

Classes

Sometimes called concepts, these are the main entities in an ontology. Each class is made up of more detailed information about items in the domain. For example in a maritime archaeology ontology, a class called 'ship' could represent all vessels that travel on the water.

Properties

Each class is made up of multiple properties that describe features and attributes of members of the class. For example, in a ship class the properties could be length, body_type or material_type.

Facets

These are restrictions placed on the listing that can be entered for a property. A property may be limited to certain values, e.g. wood or steel, and may be limited to a particular number of values. For example, in the ship class, each ship may be restricted to only a single value for propulsion_type (sail, steam, diesel).

Noy (2001) specifies that an ontology, together with a set of individual instances of classes constitutes a knowledge base of a domain. Once an ontology has been established, it must be integrated into the data sharing system. Harris (2006) describes the architecture for such an ontology as having 5 layers. These layers are described on the ontology definition language, data structures, assertions and constraints, reference data and operational data. He notes that although the first and last layers, ontology data language and operational data, are really elements outside of the ontology, but he has included them because they are necessary for the performance of the ontology. The five layers he specifies for the architecture of an ontology are described below.

Ontology Definition language

This is the language that an ontology designer chooses to define the ontology. Languages such as RDF and OWL are the standard ontology specification languages at the present. These languages are still evolving however. Harris suggests that a well designed ontology may outlive a number of specification languages and so may need to be migrated from one language to another. Although this item is listed here it does not comprise part of the ontology itself, but must be made available to the system in order to create the ontology.

Data Structures

The specification for each ontology must include a complete definition of all of the data structures that it will use. Depending on the specification language selected, these could be: tables containing columns, classes with slots, statements of the form: subject, predicate, object or one of several other basic formats (Harris 2006).

Assertions and Constraints

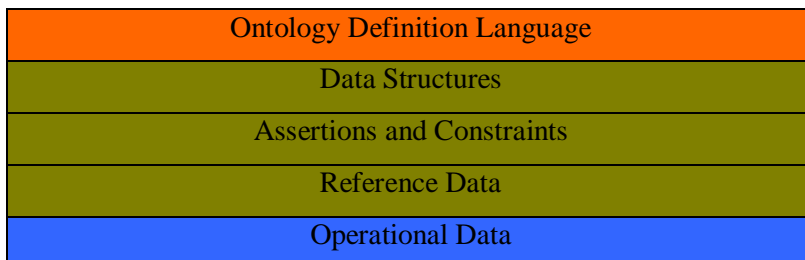
An ontology also contains a set of assertions and constraints that control the structure of the data. These assertions and constraints define the rules concerning the relationships between data structures such as classes and properties and how the data can be accessed or used. Two types of constraints are found in most ontologies: integrity and inference. Integrity constraints limit how operational data can be created and what form it will take. An constraint of this type might mandate that a postal code must have four digits. Inference constraints control how operational data can be used in combination with other data to make inference. Constraints of this type may come into play in situations where privacy concerns are present.

Reference Data

Reference data in this situation describes controlled vocabularies, taxonomies or thesauri stored in the system. Agents that use the same reference data will have a common frame of reference when exchanging data.

Operational Data

This information is sometimes called instance or instantiation data. It supports the ontology but is not part of it. The operational data consists of configuration and activity data. Configuration data is used to allow a certain amount of flexibility in the data sharing system. The system can be ‘configured’ to meet the needs of the users. Activity data consists of the exchanges of messages between agents that agree to use the ontology.



**Figure 3.2 – Five layer model of architecture of an ontology
(after Harris 2006)**

3.2.8 How to define an ontology?

Building a complex ontology requires a level of planning similar to that used in the full-cycle software development process. The first step is to acquire domain knowledge. This requires the assembly all of the information resources and expertise needed to define the terms that will be used to formally describe the environment in the domain. Next, organize the ontology by identifying the overall conceptual structure of the domain. The ontology should then be described in detail by adding concepts, relations and individuals that will inhabit the domain. Any syntactic, logical and semantic inconsistencies between elements should be resolved at this point. Finally, commit to the ontology by publishing the ontology within its intended development environment (Denny 2004).

Noy reports that ontology development is different from designing classes and relations in object-oriented programming (the current standard in software creation).

In object-oriented programming centers primarily around methods on classes – a programmer makes design decisions based on the operational properties of a class, whereas an ontology designer makes these decisions based on the structural properties of a class. As a result, a class structure and relations among classes in an ontology are different from the structure for a similar domain in an object-oriented program (Noy 2001).

With this statement Noy is specifying that although both the programmer and ontologist deal with classes, a programmer makes design decisions based on what an object does, and an ontology designer specifies what an object is, i.e. and what sort of data it contains.

Regardless of the domain, the steps in creating an ontology are the same. These steps are:

1. Define the classes
2. Arrange the classes in a taxonomic hierarchy
3. Define the properties and describe the allowed values for the properties
4. Fill in the values for the properties with instances

Define the classes

First determine the parameters or borders that make up the scope of the methodology. It is important to consider how the ontology will be used by the system, and what sort of questions it should be able to answer. Next, consider reusing any existing ontologies. Now that a concrete picture of the domain has been established, list all of the important terms that will make up the classes for the ontology and specify the properties for each item.

From this information, determine the classes and class hierarchy. This should be specified in a taxonomic structure.

Define the properties and describe the allowed values for the properties

Specify the properties of the classes and determine what are allowable values for these properties. This information should include value types, allowed values and the number of values as well as whether multiple values are allowed. Determine also the data types that are allowed in the properties; i.e. whether a particular property should take a text or numerical value.

Fill in the values for the properties with instances

The last step is to populate the ontology with data. To do this the ontology designer must create an instance for each class in the class hierarchy and fill in the property values. Instantiation of the ontology will allow the system to be tested.

3.3 Ontology tools

As can be seen by the detail in the previous section, constructing an ontology can be a complex, time consuming process. New ontology languages and software applications attempt to solve some of these issues. The next section describes these tools, and their suitability for use on the Semantic Web.

3.3.1 Ontology specification languages

At present there are three ontology specification languages used on Semantic Web applications: RDF, DAML+OIL, and OWL. In the late 1990s separate groups in the U.S. and the U.K. were exploring the idea of creating a web ontology language with more flexibility and expressiveness than that provided by RDF. European researchers favoured Ontology Integration Language (OIL), and American research efforts supported the development of DARPA (Defense Advanced Research Projects Agency) Agent Markup Language (DAML). In 2000, a combination of the features of DAML and OIL was created, hence the name DAML+OIL. OWL (Web Ontology Language) created by the W3C, and which forms part of their Semantic Web specification, has largely superceded the use of DAML+OIL and RDF OIL (<http://www.w3.org>, see also Shadbolt et al. 2006).

OWL is a computer understandable ontology language used to represent web content This language has a specific grammar (set of rules) for using vocabulary terms in a particular domain. The grammar rules may be tight or lax depending on the environment in which

they are used. OWL provides a rich vocabulary description language for describing properties and classes, such relations between classes, cardinality, quality, richer typing of properties, characteristics of properties and enumerated classes. The use of this language allows ontology developers to describe the semantics of knowledge in a machine-accessible way (Antoniou and van Harmelen 2004).

3.3.2 Methods for creating ontologies

Designing an ontology from scratch is an extremely time-consuming process, however there are software tools available to build ontological schemas alone or with pre-existing instance data. There is no automated solution at present however, because a human is required to make the subtle connections between concepts required in an ontology. Ontology editors may wish to use semantic database tools such as WAT, the Word Association Thesauri. This linguistic resource presents the results from word association tests. It lists the frequency of responses to stimulus words (170,000) by individuals of varying age, sex and professional backgrounds. The large sample size of the group ensures the reliability of the data. The thesaurus is available in multiple languages including English, German and Russian (Smrz et al 2003).

Due to the complexity of the task of ontology building, the first step in any ontology project should be to determine if there are pre-existing ontologies that may suit the aims of the new system. Libraries of ontologies are available online; examples of these are listed in Table 3.9. Other sources are integrated vocabularies where one or more individual controlled vocabularies have been combined. An example of this type of resource is the Unified Medical Language System (<http://umlsinfo.nlm.nih.gov>) which integrates 100 biomedical vocabularies and classifications. When using a pre-existing ontology the designer will need to refine it to fit the new domain. Often existing concepts and properties must be edited, and in some cases alternative names for certain classes may prove to be of use (Antoniou and van Harmelen 2004).

An ongoing challenge for data sharing systems is the creation of metadata and ontologies. Although ontological tools have improved in the past several years, manual creation of ontologies is still a labour intensive and expensive process. Antoniou et al (2004) suggests that machine learning techniques allied with knowledge acquisition software is a way to navigate this bottleneck. Machine learning techniques refer to a systems ability to 'learn', i.e. acquire knowledge by ingesting data. Some areas where machine learning may be of assistance are in the areas of extraction of ontologies from existing data on the web, extraction of relational data and metadata from existing data on the Web and merging and

mapping ontologies by analyzing extensions of concepts. This tallies well with Berners-Lee's vision of an 'active ecology of agents' (Berners-Lee 2001). However these semi-automated methods of generating ontologies are still in the early development stages.

Ontology library	Location
Ontolingua ontology library	http://www.ksl.stanford.edu/software/ontolingua/
DAML ontology library	http://www.daml.org/ontologies/
UNSPSC	www.unspsc.org
RosettaNet	www.rosettanel.org
DMOZ Open Directory Project	www.dmoz.org
CancerOntology	www.mindswap.org/2003/CancerOntology/
Art and Architecture Thesaurus	www.getty.edu/research/tools/vocabulary/aat
Union List of Artist Names	www.getty.edu/research/conducting_research/vocabularies/ulan/

Table 3.9 – Online libraries of ontologies, most publicly available (after Noy 2001 and Antoniou et al 2004).

3.4 Overall usefulness of metadata and ontologies

This chapter has focused on the usage of metadata and ontologies in various types of data sharing systems. Metadata provides additional information that can be used to describe the data making it discoverable, and ontologies provide a mechanism for making correlations between data sets. Although these technologies have been available offline for many years, the Semantic Web presents an opportunity to reuse them in novel ways to aid the discovery and retrieval of data. Metadata and ontologies form the core of data sharing systems from commercial sites like Amazon to immense medical databases like the National Cancer Institute. The purpose of this project is to apply these tools to create a data sharing system suitable for the maritime archaeology community. The methodology employed during the course of this project will be presented in the next chapter.

Chapter 4 – Methodology

Developments in information technology have made it possible to share datasets and provide a search engine to promote discovery of this data. Federation of datasets provides researchers a mechanism to gain access to information that was previously unavailable in an offline setting. With the aid of semantic technologies searching tools can be created to allow precise searches across distributed datasets.

This chapter outlines the methodology of the research undertaken in this project. This research seeks to understand the feasibility of federating maritime datasets and to explore the benefits of using the semantic web to target search results. Two different software applications (PGL and ArchaeoView) were utilized in this case study to determine the issues involved with creating a data sharing system. As mentioned in Chapter 1, the focus of the research was on examining the usability of combining federation of data with semantic search to allow researchers access to archaeological data. Towards that end a seven-step process was employed to determine the answers to the research questions. These steps were:

1. Informal survey of data sharing in maritime archaeology
2. Determine whether there are existing ontologies that can be reused
3. Analysis of the data samples
4. Study 1: Personal Grid Library (PGL)
5. Review Process
6. Study 2: ArchaeoView
7. Evaluation

In the first step, an informal survey was taken to determine the current state of data sharing in the maritime archaeology community. Next, three sample databases from maritime archaeological groups were obtained and examined with a view to include them in a data sharing system. Following this, a pre-existing software PGL was used to federate the sample datasets and an analysis of its usefulness was performed. In an effort to more fully understand the complex interplay of metadata and ontologies, a further examination of the strategies used in other domains to share data was completed. Building upon the knowledge gained from the first study, and an examination of other data sharing systems, a purpose-built application named ArchaeoView was designed and investigated. Finally, an evaluation of the information received in these studies was detailed. An in-depth description of the methodology used in this project is presented next.

4.1 Informal survey of data sharing in maritime archaeology

During the initial stages of the research, several Australian archaeologists were consulted regarding the current state of data sharing in archaeology, in an attempt to ascertain the needs of this community regarding data access. Most expressed a desire to move forward with a system to house maritime data, as long as security and intellectual property concerns could be handled. An in-depth exploration of these responses informed the research design and is provided in section 5.1 in the next chapter. The list of questions from the informal survey is located in Appendix D.

4.2 Determine whether there are existing ontologies that can be reused

In any software development project, one of the first steps should be to determine whether there are existing software components that can be reused. In a data sharing environment one of the most expensive and labor-intensive parts of the system to create is the ontology. Keeping this in mind, a literature review was conducted to discover whether there are ontologies or thesauri from the maritime archaeological community that could be reused for this project. It was determined that although there is a move toward standardization (i.e. the English Heritage NMR Monuments Thesaurus), there do not appear to be any explicitly specified archaeological ontologies publicly available which have widespread acceptance.

4.3 Analysis of the data samples

As stated previously, the goal of this project is to link datasets stored in separate locations in order that researchers can query the combined data. Towards that end, datasets were provided by Heritage Victoria (VIC) , Heritage New South Wales (NSW) and the Townsville Maritime Museum (TMM).

To obtain the sample maritime archaeological data for this project, it was necessary for the donor organizations to transfer the datasets to JCU. This caused difficulties from the outset in that many maritime research groups do not have dedicated IT support staff. The data from the Townsville Maritime Museum was stored in a Microsoft Access database, and it was a relatively simple matter to have a researcher at the museum email the Access database as an attachment. However, the other datasets from Heritage New South Wales and Heritage Victoria were embedded in web database applications that are used to provide information regarding shipwrecks to the Heritage organizations web sites (See section 2.1.3). A wait of several weeks was necessary in order for the data to be exported from the host systems. This data was then provided as an attachment via an email. The Heritage NSW data was exported as XML and attached to an email proved difficult to download.

Problems ensued when Microsoft Outlook that attempted to download the XML file automatically. Since the file contained poorly formatted XML, the email application locked up while trying to determine the appropriate manner to display the contents of the file. The dataset was finally downloaded using a web-based email system.

Table 4.1 describes the formats, sizes and content types of the sample data. Due to the data being composed of text only, the datasets are small in size. In fact one of the databases is under 1 MB. This small size should not disguise the fact that each dataset contains thousands of records.

Dataset	File Format	Size	Records	Content Type
VIC	Excel	707 kb	1007	Register of ships
NSW	XML	3211 kb	2113	Register of ships
TMM	Access	2188 kb	4104	Artefact listing

Table 4.1 – Details of sample datasets

The small size of databases in fact aided an in-depth examination of the issues involved in federating datasets. These smaller datasets made it easier to pinpoint the problem areas in combining the data. With a larger file size the sheer volume of the records could obscure important details. An analysis of the three sample datasets revealed three areas of concern in providing access to the data: lack of cohesion between datasets, differences in field names, and the correctness of the data. These three issues and the limitations they placed on the case study will be considered in the following sections.

4.3.1 Lack of cohesion between data sets

As specified in Table 4.1, two of the datasets (NSW and VIC) contain a register of ships. This data is a local copy of data from the Australian National Shipwreck database. The NSW data describes shipwrecks off the coast of New South Wales, and the VIC data contains a list of ships that sank off the Victorian coast. The third dataset from the Townsville Maritime Museum (TMM) however contains a list of accessioned artefacts located in the museum’s inventory.

In addition to differences in content, the three sample datasets were provided in varying formats. While both the VIC and NSW datasets were exported from online database

systems, the VIC data was provided in Microsoft Excel, and the NSW data was exported as XML. To complicate the situation further, the TMM data was stored in a Microsoft Access database. These varying formats of the data are an issue for a system that attempts to span multiple datasets.

The screenshot shows a Microsoft Excel spreadsheet with the following data:

	A	B	C
1	ID Number	Name of Ship	Primary In
2		1 ABSTAINER	Transport
3		2 ACHILLES	Transport
4		3 ADA BURGESS	Fisheries
5		4 ADELHEID	
6		5 ADIEU	Transport
7		6 ADMIRAL	
8		7 AGENORIA	Transport
9		8 AGNES	Transport
10		9 AGNES	Transport
11		10 AGNES AND HANNAH	Transport
12			
13		11 ALBERT	Transport

Figure 4.1 – Dataset VIC, format: Microsoft Excel

```

- <SHIPWRECKS>
  - <ROW>
    <BEAM />
    <BUILDER />
    <CARGO>Timber</CARGO>
    <COMMENTS />
    <CONSTRUCTION>3</CONSTRUCTION>
    <COUNTRYBUILT>13</COUNTRYBUILT>
    <CREW />
    <CREWDTHS />
    <DATEBUILT>1911</DATEBUILT>
    <DATEWRECKED />

```

	A	B	C	D
1	BEAM	BUILDER	CARGO	COM
2			Timber	
3	5.70000		Ballast	
4	5.486	William Wo	Timber	
5	3.90100	David Rob	Maize	Reg
6				Aux
7	5.882	Edward Da	Timber	

Figure 4.2 – Dataset NSW, format: XML and as viewed in Excel

Acquisitions : Table				
ID	Date Entered	Reg Number	Type of Item	Item
23	21/06/1986	986.032	object	compass
22	21/06/1986	986.031	object	gun sight
11	21/06/1986	986.029	other	certificate - nautilus, s.s.
10	21/06/1986	986.027	other	AIMS research fleet, b&W, (29.5x23)cm
24	21/06/1986	986.034	object	vampire, h.m.a.s.

Figure 4.3 – Dataset TMM, format: Microsoft Access

4.3.2 Mapping field names

In order to perform a query against multiple separate datasets to retrieve the data, it is necessary to have at least one field in each dataset that contains similar data. For instance, each dataset has a table with a field specifying the name of a ship. In the VIC table the field is called 'Name of Ship', while in the NSW dataset this is called 'Title'. An examination of these two datasets reveals that there are over 30 fields in each table which contain similar data. In each dataset, the names of these fields are slightly different. To create a query based on this information, the person creating the query would need to know in advance the alternate field names found in each dataset. Although complicated, this method can work satisfactorily with two datasets, but would not scale well enough to search through more a few datasets at one time. Structured Query Language (SQL) code can be used to request data from datasets. An example of the SQL necessary for a simple query is listed below in Figure 4.5. In this case we are requesting the data from the 'Name of ship' and 'Date wrecked' columns of the tables in the NSW and VIC datasets.

While the NSW and VIC datasets are similar in content, the data in the TMM is very different. Being primarily a list of artefacts held at the Townsville Maritime Museum, the data has little relation to the other two datasets. A simple database query such as the one listed in Figure 4.4 can be used to access data similar in content. However when the data

has no relation to any other data in the repository it can be difficult to retrieve it using a standard query.

```
SELECT
NSW.table.Title, NSW.table.DateWrecked,
VIC.table.Name_of_ship, VIC.table.WhenLost
WHERE NSW.table.Title = 'Mary' and
VIC.table.Name_of_ship = 'Mary'
```

Figure 4.4 – Sample SQL query against the NSW and VIC datasets

4.3.3 Correctness of data

In order to perform a query against multiple datasets, it is necessary to make some assumptions about the underlying data at the time the program runs. One of these assumptions is that the datasets meet the normalization restrictions mentioned in section 2.1.2. These requirements must be met in order for the query to return understandable results. Each of the sample datasets fails to meet these normalization rules, making it difficult to run queries against them.

The fourth normalization rule in Table 2.2 states that each piece of data should have a unique identifier. This allows the system to differentiate between similar records. Neither the NSW nor the VIC datasets have unique identifiers for each ship listed. Certain names are 'popular', i.e. occur more than once in a dataset. An example is the ship name 'Mary' which occurs 5 times in the NSW data and 7 times in the VIC data. Frequently a ship name will be reused once a ship is no longer in active service, either through retirement or loss at sea. According to the NSW data, a ship named 'Elizabeth' was lost in 1816, 1844, 1846, 1848, 1860 and 1876. This replication of the ship name without the use of any other unique identifier makes it very difficult to determine to which ship a record refers. This situation is handled in the Australian National Shipwreck database by creating a fixed set of records across all of Australia. Then each ship was assigned a unique number pertaining to that vessel (Green and Vosmer 1992). Although this change was made in the central repository, this identifier was not added to the original local copies of the data held by the heritage organizations.

Another issue presented by the combined datasets concerns the storage of dates. As mentioned previously, the data was provided in three different formats: XML, Microsoft

Excel and Microsoft Access. Each format treats dates in a different fashion. In XML dates are merely a text string as is all of the content. In Excel, dates are stored as a serial number that is then reformatted when they are displayed on the screen in a particular pattern, e.g. yyyy-mm-dd or dd-mm-yyyy ('d' for day, 'm' for month, and 'y' for year). In Access dates are stored as a date/time data type which is proprietary to Access. These dates are also reformatted like the dates in Excel. In order to run a query against date data in these different formats, there needs to be a standard format that the system can expect in order to translate the data into a scheme that the program understands..

Analysis of the three datasets showed that similar kinds of data were often recorded in an inconsistent fashion. As an example, measurements of features of the ships were often stored in a combination of the imperial and metric systems. In some cases various methods of representing imperial measurements were used in addition to metric measurements. The inconsistencies in data formats make it difficult to obtain accurate search results across datasets.

Length formats	Date formats
3ft 6ins	6 July 1888
3' 6'	6-7-1888
3.5 ft	6/7/88
1.0668 M	Around 1888

Table 4.2 – Examples of inconsistent formatting in the sample datasets

4.3.4 Limitations on case study

After an analysis of the issues presented by the data was performed it was necessary to make updates to the data in order to complete the rest of the research. These steps placed limitations on the case study which are considered next.

Basic updates of the data were performed to make the data in each column consistent in regards to the format and data type expected. Therefore, data was placed in a consistent date format to aid searches across a date range. For example, since all of the data in the NSW dataset was provided in XML, all of the original formatting based on data type was missing. These dates were formatted to have a consistent pattern for dates of dd-mm-yyyy. A decision was made to leave the length measurements as is because it was determined that a more complete data cleaning/updating operation was beyond the scope of the project.

1	The data in the datasets were updated to meet format restrictions based on the data type expected
2	Text data was formatted as date where needed
3	Complete data 'cleansing' is beyond the scope of the project

Table 4.3 – Limitations placed on case study of maritime archaeological data

Based on input from the maritime archaeological community, a goal of the project was to allow the users to maintain their data in its present format and not require extensive manipulation of the data to make it suitable for use in the data sharing system. However, in order to be able to perform any queries against the data some updates were required to be made to the datasets. This issue will be considered further in the Results chapter. Once the updates of the data had been completed, the next step was to use Storage Resource Broker to federate the datasets.

4.4 Storage Resource Broker

The technical objective of this project was to provide a 'virtual' interface to the data, meaning that the information sources appear to be grouped together in the same location, even when they are physically held on different sites, by different organizations. This was achieved through the use of Storage Resource Broker (SRB).

Figure 4.5 below depicts a scenario where multiple databases in a variety of formats exist. SRB's Metadata Catalogue (MCAT) contains information regarding the location, format, size, etc of the datasets. Therefore, the SRB server 'knows' how to access each of the datasets. An application is depicted at the top of the diagram that makes a request for data to the SRB server. Using the MCAT, SRB obtains the data from the federated databases and returns it to the application. This method was chosen to allow SRB to maintain connections to federated datasets and to provide a mechanism to return search results.

SRB provides two useful services to this project: access to data across multiple locations including federation of data, as well as metadata storage and querying. The benefit of federating multiple sites, as opposed to just having a single storage system is that federation allows each site to maintain control over its own information through its own administrative domain. Each data holder can keep the data at their location, and grant access to it to appropriate users, thereby permitting controlled access to this information. The metadata service (MCAT) provides every object in the data store to have metadata attached to it.

This allows relevant information, such as the name of the ship, port of origin, or site to be added to the system for individual objects.

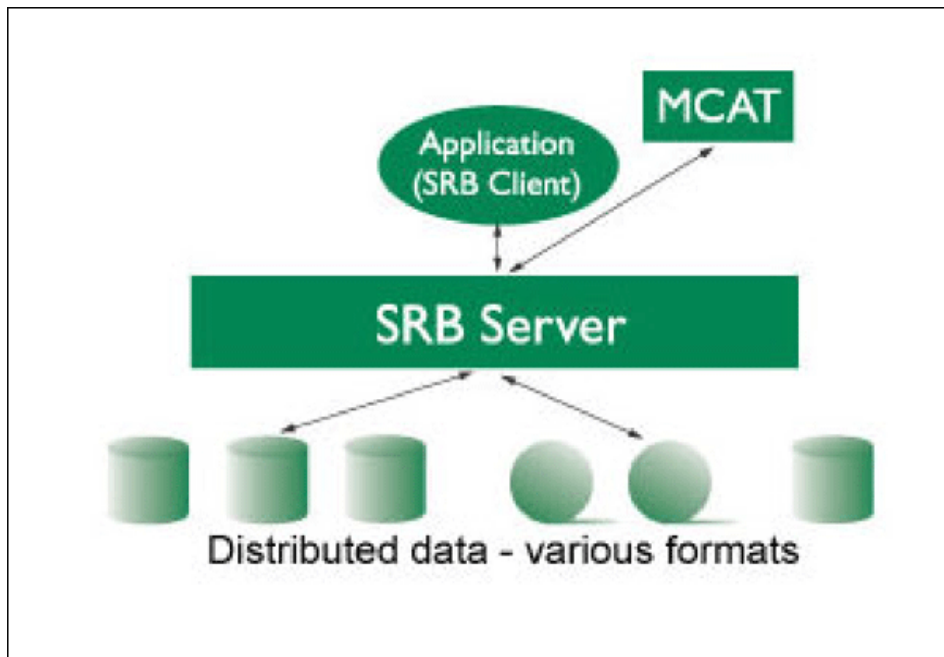


Figure 4.5 – Storage Resource Broker architecture
(after <http://roadnet.ucsd.edu>)

To access datasets held in databases, SRB's Data Access Interface (DAI) was used. This permits standard Structured Query Language (SQL) to be passed to the remote databases, and have the results returned to the search engine. Due to the inconsistency between database schemas, it was necessary to map the database fields to a common ontology which is stored in the system.. SRB provides the ability to combine search results, and allow the user to download the union of the results in a suitable format for import into the researcher's own work. All operations are access controlled, allowing appropriate access to people with different needs. SRB handles the translation and data transfer tasks using the MCAT that keeps track of the location of each piece of data.

4.5 Study 1 – Personal Grid Library

In the first study a previously existing open source software application called Personal Grid Library (PGL) was used in order to explore the problem of federating data and allowing searches using semantic tools. SRB, while a very powerful data federation application is difficult to implement and use without extensive training (Wyatt et al 2007). Therefore various 'front-end' applications such as PGL have been created in order to

reduce the complexity of the tasks necessary to use SRB. PGL acts as a user interface to SRB and hides much of the complexity surrounding the use of SRB from the user. PGL was developed in the Visualisation, e-Research and Grid (VeRG) laboratory at James Cook University. Figure 4.6 shows the general architecture for this application.

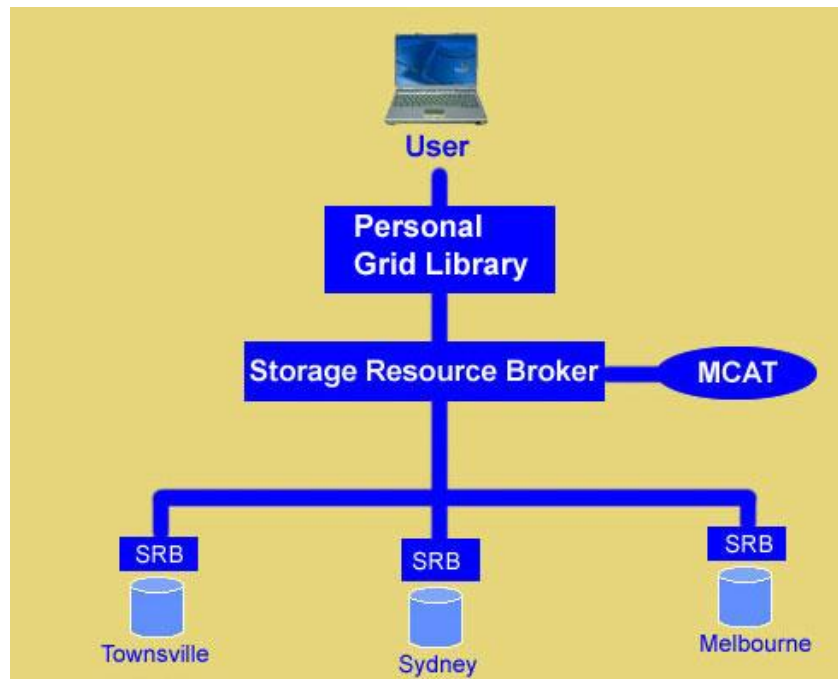


Figure 4.6 – Architecture for Personal Grid Library

PGL allows an administrator to define metadata templates that the users apply to the data in the repository. This aids in the storage and retrieval of the information. PGL runs as a GridSphere portlet. GridSphere is an open source portal framework that enables software developers to quickly create and run third-party portlet applications within the GridSphere portal framework (<http://www.gridsphere.org>, see also Novotny et al. 2004). Portlets are components of a web site that provide access to a specific information source or application. They are often used to combine multiple content items into a single interface. This means that information from several applications can be displayed simultaneously on one page.

In the case of this project, PGL was explored as a means to provide data curators with a way to modify the ontological mappings for new and existing datasets and provide search facilities to the user. This includes the ability to combine search results, and download the results in a suitable format.

Three major tasks that PGL handles are file upload, metadata creation and field mapping. To add a file to PGL, i.e. to give its location and file specification to MCAT, the data holder

clicks the browse button to specify the file. Once this file is known to the system, the user adds metadata regarding the file so that it can be easily found using the search engine. A third task, specific to this research project is to allow searches which combine the datasets and return the results to the user as a table which can be downloaded. In order to handle this duty, a mapping scheme must be created. The user screen used in this task is listed in Figure 4.7 below.

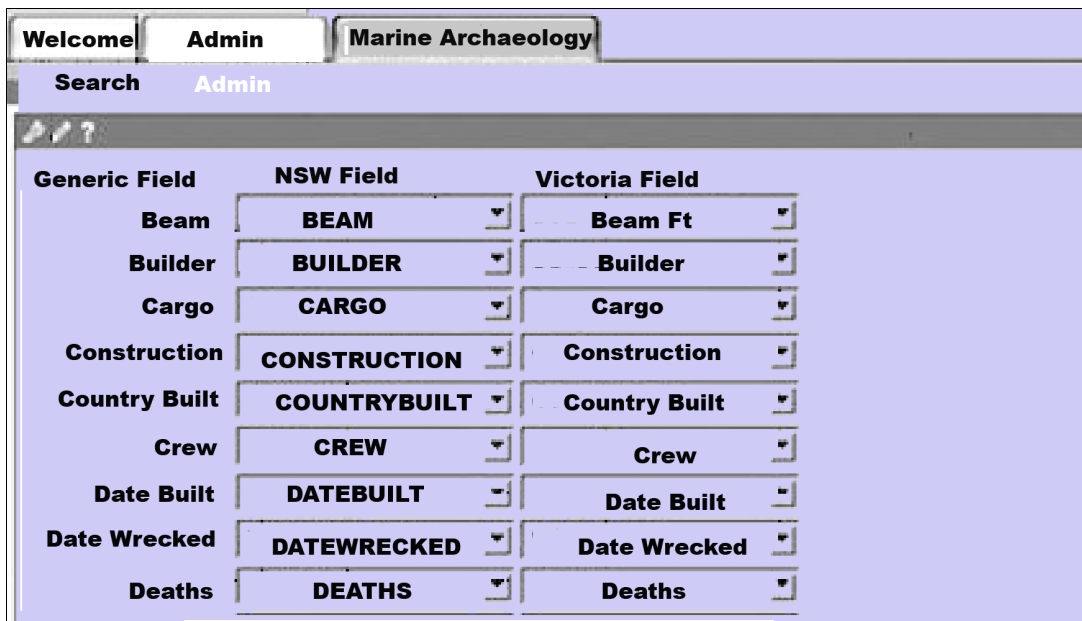


Figure 4.7 – PGL user interface

Pictured in the diagram above are the two datasets ‘known’ to the system. PGL has created a ‘generic’ column which contains a standard term for the column headings in the other datasets. Under ‘NSW field’ are the column headings from the NSW data, and the ‘Victoria field’ contains a drop-down list with headings from the VIC data. The user sets the lists so that this information matches that in the data. For instance, under the ‘Generic field’ one of the field names is DateWrecked. Glancing across at the NSW field shows that ‘DateWrecked’ is also chosen for the NSW data. This means that there is a column in the NSW data also called DateWrecked. In the next column, the VIC data shows that this column is actually called ‘WhenLost’. In order for the search engine to combine the search results, it must know the different names for these columns. When the search runs, even if the user enters a search looking for DateWrecked, the system will understand that this means WhenLost in the VIC data. In effect, the PGL system is using a single ‘light-weight’ ontology. The term light-weight is used because rather than a completely described formal ontology, PGL is updating a thesaurus with mapping information from the maritime datasets.

Although PGL is a quite powerful application, due to its complexity it was found unsuitable for use by the maritime archaeology community for data sharing. These reasons for this finding are detailed in the results section of the next chapter. At this stage in the research, further consideration was given to how the research questions for this project have been addressed in other disciplines.

4.6 Review process

This research used an iterative methodology to conduct research into data sharing for maritime archaeology. Study 1 revealed some of the complexity of the problems associated with data federation and search. While PGL offered benefits over a pure SRB implementation, it still was lacking in terms of usability. In order to pinpoint more closely the methods that should be used to develop a data sharing system for the maritime community, it was necessary to gain a more thorough understanding of the ways that other disciplines have attempted to solve these issues.

4.6.1 Archaeological Data Service

The Archaeological Data Service (ADS) provides the most well populated data sharing system available at the present time in the field of archaeology. On this site, thousands of individual items are available to download. Items are either archived within the system itself (and are maintained by the ADS) or links to the items are provided via a metadata-based search. While the system provides a welcome service for the UK, it does not seek to archive items from outside the United Kingdom. Another difficulty with the system is that it does not allow searches that return the search results in a combinatory manner. The user must click on each link individually and make arrangements to download the datasets one at a time, or to make arrangements to request access to the datasets individually. This requires additional time to review each dataset separately, and determine its suitability as a research source. Any cross-analysis of the datasets must be done off-line. For example, a search on the ADS site for 'hillfort' returns a list of 1553 documents. By clicking on the first link 'Ring Hill hillfort, Ring Hill, Haresfield' the user is taken to a second screen with details concerning this dataset. A reading of the page reveals that this dataset is not available online; the user is directed to contact the data holder directly to request access to the data.

This brings up two issues of special interest to this project. First, one assumes that if the data was listed on the ADS catalogue that the data holder is willing to make this data available. Second, unless the data is made available online, delivery of the data becomes a

problem. If the dataset is too large to deliver via email, the data will have to be placed on some sort of storage media and mailed to the user.

Advanced search

Searches: [Basic](#) | [Map](#) | [Search by resource](#) | [Advanced](#) | [Help](#)

Results

Your search: *what search on hillfort*

Records 1 - 10 of 1553

Pages: 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 > >>

Click on the headings below to view the more detailed records.



- [Ring Hill hillfort, Ring Hill, Haresfield](#)
 Scheduled Monument
Haresfield Beacon & Standish Wood, Haresfield, Stroud, Gloucestershire, England
The National Trust
- [Lewesdon Iron Age Hillfort](#)
 Area of Outstanding Natural Beauty; Scheduled Monument
Lewesdon Hill, Broadwindsor, West Dorset, Dorset, England
The National Trust

Figure 4.8 – ADS search results, list of resources
(<http://ads.ahds.ac.uk>)

Remains of Burgh Walls Camp Iron Age Hillfort, Leigh Woods

Searches: [Basic](#) | [Map](#) | [Search by resource](#) | [Advanced](#) | [Help](#)

[Historic Map](#) | [Street map](#) | [Aerial photo](#) | [Search for other sites in the area](#)

Description
Local Nature Reserve

Location
Leigh Woods; Long Ashton; Bristol; England
Grid ref. OSGB - ST 5624 7299
Grid ref. LL - 002 37 47 W 51 27 13 N
[Switch the map off](#)

Subject type
HILLFORT

Period
Early Iron Age to Post Medieval

Record maintainer (Contact details)
The National Trust

Resource Name (description)
 National Trust SMR

Depositor's Id No.
119086

Accessioned
7 Feb, 2002

Type
Collection

ADS Record ID - NTSMR-NA14271.




Figure 4.9– Archaeological Data Service individual resource listing
(<http://ads.ahds.ac.uk>)

4.6.2 Crosswalks

Many information-rich groups have changed their focus from in-house manual systems to computerized systems. These systems are often designed locally without any reference to community standards for information sharing. In these cases each system maintains its own structural fields, access methods and terminologies. Now that users are requesting access to data from different systems, it becomes difficult to manage incompatibilities caused by format and structure (Woodley 2002). One method of resolving these differences in structure is to create a ‘crosswalk’ between the two systems. This is a similar process to metadata mapping, and allows searches across separate metadata systems. However, it is no easy task to map between very dissimilar systems. Figure 4.13 lists some of the common differences between metadata schemas. In this case schema refers to the underlying structure of the database tables and fields within the datasets. The schema listed as the base is Categories for the Description of Works of Art (CDWA). This schema documents the content of art databases by setting up a conceptual framework for describing information about works of art, architecture, other material culture, groups and collections of works (Baca 2006).

The misalignments specified by Woodley (2002) in Table 4.4 can be summarized as: 1. Differences in metadata between database schemas and, 2. Difficulties in translating from one schema to another. To map one metadata framework to another requires a standardised

schema for all the datasets. This standardisation effort has not been undertaken in the field of archaeology as of yet. Although the English Heritage National Monuments and Records Thesaurus provides an example of steps taken toward this goal, universal acceptance of a standard or set of standards has not occurred. In general, most archaeology datasets do not adhere to a particular structure or format. For this reason, datasets need to be handled ‘as they are’ rather than by relying on a specific schema.

4.6.3 Discovery versus cross-dataset analysis

Once data is created and stored, unless it can be retrieved it is of no use. However, as datasets grow in number and size retrieving a specific piece of data can be difficult. Take for instance the case of retrieving an email from an email system. Once a message has been read, a user will sometimes move an email into a sub-folder to clean up the ‘in box’. However locating a particular email at a later date can be time consuming. Many email system designers realize this and provide a search tool to help users find an email message. The search engine uses metadata about the email along with the user’s search criteria, for example ‘subject = holiday plans’, to retrieve a message.

Schema Issue	Description of Issue
A concept in the original database does not have a perfect equivalent in the target database	CDWA has a field for Creation-Creator-Identity-Nationality/Culture/Race does not have a field with the same exact meaning. Since ‘Subject’ is a much broader category, this is a ‘fuzzy’ match
Data exists in one field in the original schema may exist in separate fields in the target database	The CDWA Creation-Place concept may appear in the ‘Subject’ element in Dublin Core as well as in the DC element ‘Coverage’
Data in separate fields in the original schema may be in a single field in the target schema	In CDWA, the birth and death dates for a ‘creator’ are recorded in the Creator-Identity-Dates, as well as in separate fields – all apart from the creator’s name. In MARC, both dates are a ‘subfield: in the string for the ‘author’s’ name
Information in one schema may reside in a field that is indexed, whereas it is only free-text descriptive information in the other schema	Primary ‘keys’ or indices may not be the same across schemas.
There may be no field in the target schema with an equivalent meaning	Unrelated information may be forced into the same ‘bucket’.
In only a few cases does the mapping work well in both directions.	Often one schema is mapped to another and extraneous information is omitted.

Table 4.4 – Common misalignments between schemas (after Woodley 2002)

The process of finding data that is of interest to the user is called data discovery. Quite often discovery tools consist of a search engine that can be used to locate data. The most common type of system returns a list of links that the user can click to navigate to the data. Once the data is located, it can be downloaded to the users system. This assumes however, that the person conducting the search has a familiarity with the data that is being searched through, and will be able to recognize the data of interest when it is returned. Unless a very detailed description of the data is provided, the user will not know how to differentiate between two similar sets of data. Considering the case where the user is well-informed about the data three or four datasets can be chosen for download.

This scenario however is different from the request of the maritime archaeological community for data access. This group would like to be able to return the results in a combined manner, so that cross dataset analysis is possible. This is very different from discovering individual datasets, downloading them, and then performing analyses upon them. Since the focus of this project is on combining search results while online, returning the combined results online moves the analysis phase earlier in the process. The researcher will not need to download a dataset in order to determine if it meets some criteria.

Following the review phase of the research a further study was completed, this time working with a programmer in the VeRG lab to design a new system capable of federating maritime archaeological datasets and targeting search results. This application forms the basis of Study 2, which is detailed in the next section.

4.6 Study 2 – ArchaeoView

One of the prime goals for the archaeological data sharing system was to make the tool user friendly, so that little administrative or technical input was needed. This requirement formed the focus of the design efforts for the second study. Keeping this idea in mind, development began on a system named ArchaeoView. Figure 4.10 shows the basic architecture for the application. ArchaeoView is also used as a user interface to SRB, and like PGL handles much of the complexity of the data upload details.

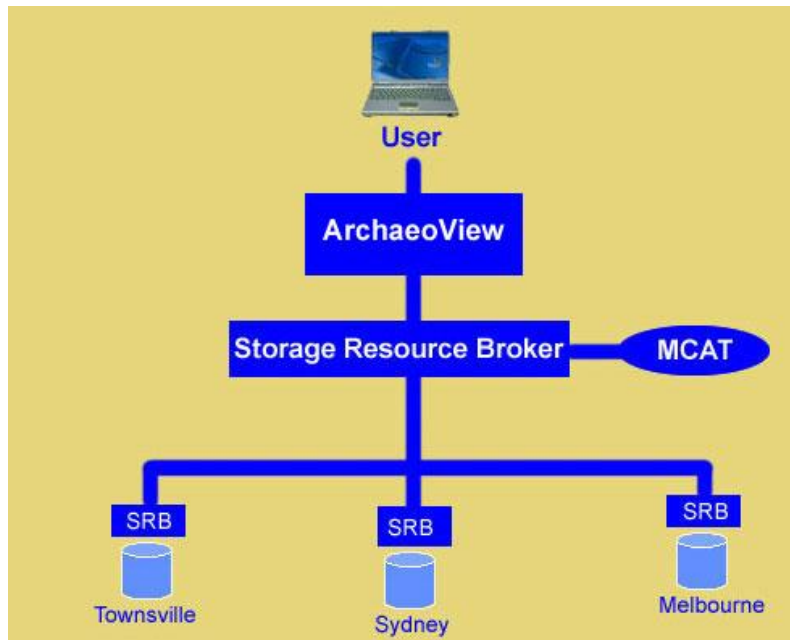


Figure 4.10 – Basic architecture of the ArchaeoView system

The user interface for ArchaeoView has two components: the search engine and the data upload area. The home page for the web application displays the search engine used by the software. The user can select keywords from a drop-down menu that match semantic metadata stored with the maritime archaeological datasets. Figure 4.11 shows the user interface for ArchaeoView, looking at the search engine.

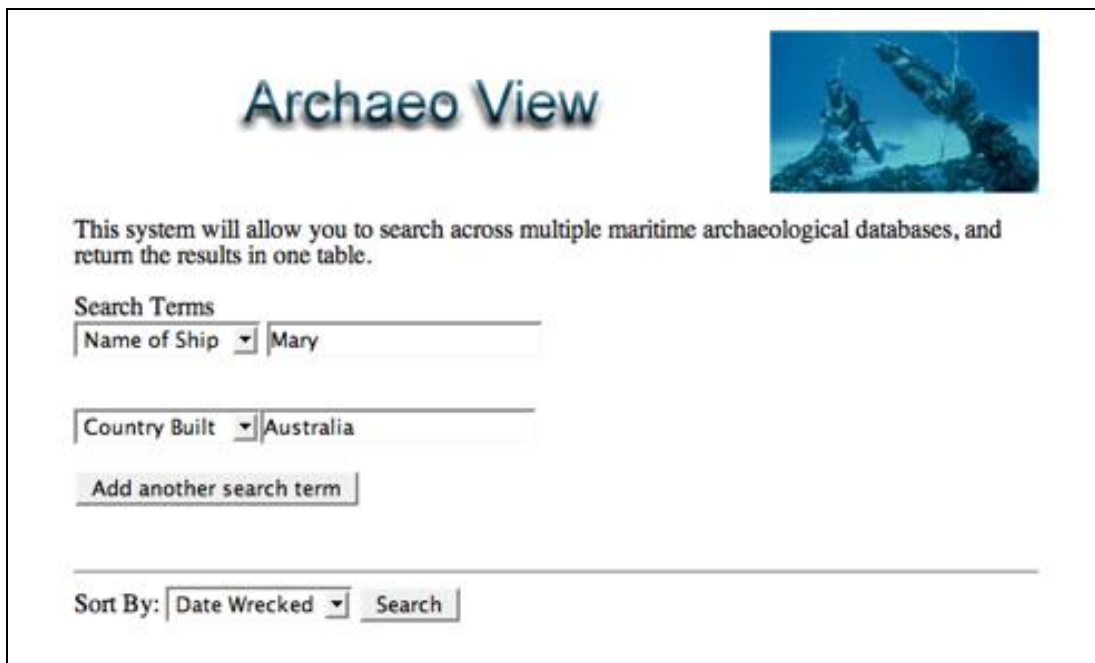


Figure 4.11 – Search engine view on the ArchaeoView application

In the figure above, the user has chosen 'Name of Ship' for the first search term. A free-form text box holds the value for the search. The first criterion for the search is that 'Name of Ship' equals 'Mary'. In this example the user has clicked the 'Add another search term' button to add another keyword field to the search query. For the second item the criteria, 'Country Built' equals 'Australia'. At this point it should be noted that this is a straight equivalency match. Boolean searches such as 'Name of Ship' does not equal 'Mary' are not supported. This is an effort to limit the complexity of the search interface. The user has elected to sort the results by 'Date Wrecked' via the 'Sort By' box. Once the search criteria have been entered, the researcher can click to 'Search' button to run the query. Figure 4.12 shows the results page, displaying the combined set of data returned from the search.

At the top of the results screen are the search criteria that have been run against the combined maritime archaeological datasets. Three pieces of information were sent to the query engine:

- Ship Name = Mary
- Country Built = Australia
- Sort By = Date Wrecked

The search engine is again presented in case the user wishes to revise the search query. The results of search against the federated datasets are displayed below the search engine tool.

Archaeo View



Search Results

Ship Name = Mary & Country Built = Australia

Sort by Date Wrecked

Revise this search

Search Terms

Select one

Add another search term

Sort By:

Ship Name	Country Built	Date Wrecked	Link
MARY ANN	Australia	AUG1850	VIC
Mary Ann	Australia	1851/08/13	NSW
Mary	Australia	1866/02/25	NSW
ELLEN AND MARY	Australia	20JULY1876	VIC
Mary Campbell	Australia	1889/04/28	NSW

Figure 4.12 – Results of the query as returned on ArchaeoView

The results from the query are placed in a single table, and displayed to the user. The data is sorted by 'Date Wrecked', therefore, the data are in chronological order. Under the column link is a URL to the original dataset, so that the user can download the complete data if it proves of interest. An additional benefit of the search is that partial matches are returned as well. For example the search criteria for 'Name of ship = Mary' returns all records with the word 'Mary' in the ship name column. This is helpful for the case where the data may not be entered precisely. The display of this combined view of the data acts as a 'preview' of the data so that time is not wasted downloading data of questionable value. It should be noted that the search is case sensitive, so 'Mary' will return a different set of data than 'mary'. This has bearing on the case where data is formatted in an inconsistent manner.

The second area of the user interface handles the data and schema upload process. This area is reached via a link that is only visible once the user is logged into the system with

administrative authorisation. Figure 4.13 shows the screen used to add datasets to ArchaeoView. A three step process is required in order to upload data. These steps are:

1. Federate the datasets by identifying them to SRB and the MCAT using ArchaeoView
2. Specify the database schema
3. Map the new dataset to the generic schema in ArchaeoView

The ArchaeoView application allows the user to upload a file into SRB. In the research for this project, once the data cleansing operations described in section 4.2 were completed, the datasets were uploaded to a server on which SRB and the associated MCAT were installed using ArchaeoView.



Archaeo View

Add data to the system
To add a new file to the system, choose the resource type from the drop-down box, and select the file using the browse button.

File Type:

File name:

Schema:

Figure 4.13 – Data upload screen for ArchaeoView

In the first step of the upload process the user specifies the file type from the 'File Type' drop-down menu. The file is uploaded by selecting the dataset using the Browse button for 'File name'. In step two of the process the user specify a database schema. A browse button is offered to choose the schema from the users machine. In the example in Figure 4.13, the user specifies an XML dataset and provides a schema which is in XMLSchema format (see section 2.3 for more information regarding this format). Once these fields have been completed, the user clicks the 'Upload' button.

The third and final step in the process is to map the new dataset to the existing data schema used by ArchaeoView. Figure 4.14 shows the mapping utility used by the system. In order to translate between datasets that are in multiple formats and have differing structures such as table and column names, the system must have a mechanism to handle these inconsistencies. This is handled by mapping the field names of the new datasets to existing 'generic' field names in the ArchaeoView system. For example, in ArchaeoView there is a generic field name called 'Crew' which contains the number of Crew who sailed upon a particular ship. This field is also called 'Crew' in the NSW data but is named 'Crew No' in the VIC data.

Archaeo View

Mapping your data
 This section will allow you to update the Archaeo View search engine to look for data in your data sets. The Generic pull-down menu lists the keywords that the system uses.

To tie your data into the system, you should add a mapping for each field in your data set that has similar data. You may add as many fields as you like, the more fields you add, the more likely it is for your data to be included in the query.

Your schema:

System term	Your term
<input type="text" value="Name of Ship"/>	<input type="text" value="Title"/>
<input type="text" value="Date Wrecked"/>	<input type="text" value="Date Lost"/>

Figure 4.14 – Mapping a new dataset using ArchaeoView

In order to map the data to the ArchaeoView generic schema, the user first must specify the schema and thereby the dataset to which this mapping applies. The 'Browse' button is used to select the file, and then the user clicks 'Update' in order for ArchaeoView to display the appropriate field names for the dataset. The user selects the system term from the drop-down box that matches the column heading for the new dataset found under the 'Your term'

menu. The user should add all the fields that hold matching data. In this case, the user would continue to click the ‘Map another term’ button in order to keep adding matching fields. Once all the appropriate fields have been specified, the user clicks the ‘Submit’ button at the bottom of the form. Now that the user has completed the three steps necessary to make the data visible to the system, it can be discovered using the semantic search engine.

This chapter has detailed the methodology that was used in the research for this project. An iterative methodology was chosen, which is inline with the current model for agile or rapid application development (Cohen 2004). In this process a cyclical model of design, creation and testing is implemented. An analysis of the sample datasets was conducted, then two different software applications were reviewed to determine their suitability for use in an archaeological data sharing program. Once these tasks were completed, an evaluation of the results was documented. Details regarding this evaluation are provided in the following chapter. Each of the applications examined used a version of a light-weight ontology rather than a fully optimized one. The bulk of data sharing applications being designed for the Semantic Web take this format due to the less stringent requirements for placed upon thesauri versus ontologies (Powers 2003). A more thorough consideration of the benefits and liabilities of the two systems as weighed against the use cases are provided in Chapter 5.

Chapter 5 – Results and Discussion

The preceding chapter detailed the iterative methodology followed in conducting the research for this project. This chapter presents the research results and a discussion of the implications of these results. The results are centered in three areas: an informal survey of data sharing in maritime archaeology, an analysis of the two data sharing systems used in the case study, and an analysis of data federation and semantic search as associated with maritime archaeological datasets.

Before turning to the results obtained from the research, a review of the research question may be useful. Each archaeological project results in the creation of data sets containing information regarding that research. This thesis seeks to determine whether this data can be made available in a data-sharing environment. In consideration of this research focus, two further questions emerged:

1. Are there tools available to federate or combine these data sets?
2. How can the search results be appropriately targeted when searching across a variety of data sources?

5.1 Informal Survey of Data Sharing in Maritime Archaeology

In order to determine the data sharing needs of the maritime archaeology community, an informal survey of maritime researchers was implemented. This survey was conducted at James Cook University and at the AIMA/AAA annual conference in 2005. Eight archaeologists were interviewed, six maritime and two terrestrial (land-based). Of the maritime archaeologists one was a early career researcher (less than 2 years experience), three were mid-career (5-8 years experience) and one was mid-career but very experienced (more than 10 years experience in the field). The two terrestrial archaeologists were also mid-career level with more than 8 years experience. These proportions of researchers are fairly representative of the field in that most archaeologists have approximately 8-15 years of experience, with smaller numbers of practitioners at either end of the scale. As mentioned in the first chapter, there are a very small number of maritime archaeologists working full-time in the field at present, making these numbers reasonable for an informal survey. The interviews with the archaeologists took the form of a free-flowing discussion regarding the state of information-sharing in their field. The basic questions used in the survey is available in Appendix D.

The information gathered in this survey was evaluated by grouping the responses based on the five areas listed below. The responses of each archaeologist were given equal merit as bona fide members of their group rather than separating the comments based on status.

The findings from this survey fell into five broad areas:

- What is the current situation regarding data sharing in this community?
- What does this community want to achieve regarding data sharing?
- What problems currently exist concerning data sharing?
- What problems must a data sharing system handle concerning maritime archaeological data?
- What human factors will have an impact on the data sharing system?

The findings regarding these areas are listed in the remainder of section 5.1.

5.1.1 What is the current situation regarding data sharing in this community?

There are three major groups that make up the maritime archaeological community in Australia: museums, state heritage agencies and universities. Each group has a varying focus, but all members of these groups maintain paper and digital records concerning their research. The maritime community is quite small with less than 40 members working in Australia. Due to the large geographic distances separating researchers and the lack of a cohesive network, publication has become the main method for sharing data. However, due to high publication costs in Australia, much of the data remains unpublished, and makes up a vast amount of 'grey literature'. This data is sometimes made available through word of mouth at conferences, but is generally inaccessible to most of the community.

Much of this unpublished legacy data is stored in out of date formats. These data sets are often ad hoc databases that originated from spreadsheets or other textual storage means. In some cases the original data has been migrated or transferred from one data storage format to other. Added to this existing data is a tremendous amount of new data produced from advanced methods of research such as GIS, ground penetrating radar, magnetometers, and computational archaeology. The records are not generally shared with other researchers and are often maintained solely by the original researcher who conducted the project.

Although some maritime heritage organizations make portions of their data available online via the Internet, the data are often generic in nature, and are targeted towards the general

public. The Archaeological Data Service (ADS) in the UK has attempted to provide a solution to this problem by allowing researchers to upload portions of their datasets to an online repository. This central storage facility is managed by the ADS, and datasets from outside the UK are not accepted. The Western Australia Maritime Museum maintains a central database of shipwreck information called the Australian National Shipwreck database, but entries in this collection are limited to basic information concerning the physical and geographical location of shipwrecks off the Australian coast.

5.1.2 What does this community want to achieve regarding data sharing?

The most often stated request from the maritime archaeology community regarding data sharing is to be able to integrate datasets from various projects in order to perform cross-site analysis. Currently this is a difficult process requiring access to individual data sets in geographically dispersed areas. The actual content of these data sets is often in question, requiring a substantial amount of effort to contact the holder of the data, negotiate access to the data, and arrange for transport of the data to the researcher's location.

5.1.3 What problems currently exist concerning data sharing?

As mentioned previously, there currently does not exist a method to provide access to distributed datasets. Transport of the data from one place to another remains an issue. In addition, there is often little dedicated IT support that is available for solving this problem. A common concern among members of this community is that there is no funding model in place capable of handling the creation of a nation-wide data sharing system.

Currently, no data standards have been put in place regarding the creation, maintenance and archiving of archaeological data. Each individual researcher formulates a system that meets the requirements of a particular project. The use of GIS has become increasingly popular, but there has been no standard established to determine what data should be captured and stored. Due to the lack of a central system, there is no method available to supply the location of data sets that may be of interest to a research project, handle the multiple file formats and structures of the varying data sets, and determine access rights to a particular group of data.

5.1.4 What problems must a data sharing system handle concerning maritime archaeological data?

Little standardization of data recording mechanisms and formats has been initiated in the field of archaeology. While there are general guidelines in place for the acquisition and recording of archaeological data, the absence of a unifying standard has led to a lack of

organization in the structure of the resulting data sets. An additional complication is created by the interdisciplinary nature of archaeology. The same basic data may need to be examined by cognitive archaeologists, forensic archaeologists, environmental archaeologists, ethno archaeologists, industrial archaeologists, landscape archaeologists or urban archaeologists.

5.1.5 What human factors will have an impact on the data sharing system?

At present, the data is generally maintained by the group that originally conducted the research. When another researcher needs access to this data, a request must be sent to the data holder. The response to this request relies on the generosity and availability of the owner of the data. Each request must be handled individually, and the outcome is not always positive. Requests for data are often denied (Roe pers conv 2005). While the reasons for this hesitance no doubt vary, in order to federate datasets the data holder must make the data available to the system. Varying access levels can be set up which limit data access to a set of users with whom the data owner has a trust relationship. This would remove the need to 'beg' for data that is part of the archaeological record. Since the data holder would place the data in the distributed system, and set appropriate access restrictions upon it, many of the trust issues involved would be resolved and would not need to be re-evaluated on a case by case basis. Some data may fall under the jurisdiction of intellectual property contacts and due to legal restrictions would not be available for a data sharing system.

There are currently various concerns regarding the ownership of the data that must be resolved. Many researchers feel a certain ownership over data that often is the result of months if not years of dedicated work. A further reason for holding on to data is the need to report on the research in refereed publications. Many researchers are hesitant to reveal their data without completing the remaining work required to publish. Recent governmental funding models have increasingly required that researchers make this data available to others based on the fact that the research was financed by federal grants. However, ownership issues remain a significant issue.

In summary, the informal survey of the maritime archaeology community revealed that a large amount of legacy data from archaeological research exists, as well as a rapidly increasing volume of data from highly technical specialist explorations. Maritime archaeologists have expressed a desire to gain access to this data via some kind of a data sharing system, but have concerns regarding monitoring and limiting access based on legal

considerations. In addition, although maritime archaeologists are highly skilled in their particular discipline, they may require assistance to utilize complex computerized systems.

In the next section, the findings obtained from the analysis of the two data sharing systems are reported.

5.2 Analysis of Data Sharing Systems: PGL and ArchaeoView

During the course of this research, analysis of two data sharing systems was performed. The first system, Personal Grid Library (PGL), provides a generic interface that is not targeted toward the use by any one discipline. In contrast, the second system, ArchaeoView was designed specifically for sharing archaeological datasets. The results for this research were obtained by evaluating each system regarding the benefits and detriments of using each application in view of the research questions and use cases defined in Chapter 1. Both applications were examined in view of the following criteria: background details regarding the system, requirements for depositing datasets, data sharing effectiveness and overall functionality against the use cases.

5.2.1 Study 1: Personal Grid Library (PGL)

Background details regarding system

- Application is capable of creating digital libraries quickly
- Data is federated through the use of SRB
- Application is multi-platform (runs in Windows, Macintosh, Linux and others)
- A separate metadata template can be created for each library
- Searches metadata attached to objects in order to target results
- Authorized users can set access permissions for resources

As mentioned in Chapter 2, digital libraries are increasingly becoming an important component of e-Research. These systems enable researchers to efficiently store, retrieve and manipulate data and documents. For the federation of data, many groups have turned to middle-ware applications such as Storage Resource Broker (SRB). However, custom application development is required to make data in digital libraries or other repositories available to the researcher. Personal Grid Library (PGL) is a Web-based system capable of semi-automating the creation of digital libraries within SRB. The application provides access to data via a Web interface that draws its data from SRB. PGL uses the metadata and replication systems of SRB and a Web browser-based delivery system to allow researchers to develop their own digital libraries.

All data communications between PGL and SRB are handled by the Jargon library, a Java interface to SRB. Jargon contains programming code that allows PGL to control SRB. PGL runs in Java, and therefore can be installed onto any Java-compatible system such as Windows, Macintosh, or Linux. The use of SRB provides PGL with an integrated metadata catalog system, allowing metadata attributes to be associated with any type of digital object or file. The metadata entered through PGL directly updates the SRB metadata tags in its metadatabase (MCAT). This means that it is possible to use PGL to create a lexicon or controlled vocabulary specific to a particular domain that can be used for cataloguing any digital object. These terms can then be searched in a generic manner, in order to discover objects of interest within the library.

An XML descriptor file placed within an SRB directory is used to generate a view of the data within PGL. Information about the digital library is held in a metadata descriptor file. This file includes library information, as well as digital object metadata templates. The description of the digital library uses Dublin Core metadata tags, a primary international standard for document metadata. For each digital object type held in the digital library, a flat and unlimited set of metadata tags can be defined. Each tag has a label and description for use in the interface, data type, a set of possible values, and a default value.

Metadata templates form the core of the PGL system. They are set up on a per-library basis. PGL metadata templates are editable by a user who has appropriate access to the metadata descriptor. Metadata can be added to the system by selecting the file, and updating the metadata text boxes on the screen. Searching using PGL is designed to be straight forward. By default only one search field is available, and it searches the main user and system metadata fields. Authorized users are able to restrict access to resources that can be returned in a search.

Requirements for depositing datasets

In order to make the data sets available to PGL, they must be uploaded to SRB. PGL is able to handle this action, and provides a browse button to allow the user to upload individual files. At this point the contents of the file cannot be searched, only the metadata describing the file. Therefore, unless the metadata is extremely descriptive this file may not be retrieved via the search engine.

To make the contents of a database visible to the system, PGL uses a query tool within SRB called Data Access Interface (DAI). This tool allows PGL to send queries to SRB that request portions of data within the databases that match a particular search string. DAI

requires that all databases stored in SRB be formatted as PostGres, an open source database. This requirement means that any database obtained from donor organizations must be transferred into this database format. The PostGres software is free to download, however there is not a utility available to automatically save, for instance, a Microsoft Access database into PostGres. A new PostGres database must be created, and then using a command line interface, the contents of the original database must be copied into the new PostGres database. Any data that does not match the stated data type for a particular field in PostGres will cause the operation to halt.

Data sharing effectiveness

- A system administrator must set up the system
- A system administrator must migrate the data into PostGres format prior to uploading into SRB
- Data can be federated using PGL
- The data can be maintained at its current location via SRB
- Searches are limited by access restrictions, but this information is not passed on to the user
- An experienced user can operate the system successfully
- The system was able to return data from two of the three sample datasets

Through the use of PGL, multiple datasets can be quickly added to the system. If content-based searches are necessary, technical assistance must be sought in order to transfer the databases into PostGres format prior to uploading them to the system. The application is able to federate data, maintaining it in its current position once that location has been made know to SRB’s MCAT. Access to the data is provided through two mechanisms: a browse screen and a search engine. Although the user interface offers few on screen prompts, an experienced user can retrieve data from the system. Due to the system’s inability to handle data that is not associated in some way, only two of the three sample datasets were accessible via the PGL system. The search utility was able to return data from two of the three sample datasets and the third database could be accessed via the browse utility.

ID	Use case description
1	User is able to deposit a dataset
2	User is able to set access rights for data
3	User is able to make data discoverable
4	User is able to find data
5	User is able to view data online
6	User is able to download data

7	System is easy to use
8	System allows searches across multiple datasets
9	System returns search results combined in one table in a web browser

Table 5.1 – Use Cases for this research

Analysis of PGL functionality against defined use cases

PGL was analysed against the use cases that were defined in Chapter 1. These use cases are presented again in Table 5.1. The analysis revealed that although PGL can be operated by an experienced user, it does not meet all of the defined use cases. A description of the results obtained from using PGL are listed below.

- A system administrator is required for many duties
 - Set up SRB on local host machines and the central server
 - Set up PGL on central server
 - Reformat databases into PostGres format
 - Create metadata templates for data resources
- Application is difficult to use by a untrained user
- Requires an in-depth understanding of the structure of data uploaded to system as well as the system metadata schema
- Usability of system is hampered by a complex and sometimes cryptic user interface
- Few on screen instructions are provided for the user
- Search interface is difficult to use with few on screen prompts
- Search results are cryptic, no clear message to the user if no results are returned
- Keywords or metadata associated with a file are not made known to the user

In use case 1, the functionality requirement states that a user should be able to upload a dataset. As mentioned in the previous section, due to the fact that PGL is using SRB as the underlying data repository, all databases must first be stored in PostGres format. This requirement places the burden upon the user to call upon the services of an IT technician to reformat the database. Based on the analysis of the maritime archaeological community, it is likely that this step would be onerous due to the lack of technical support reported by the archaeologists.

Another use cases states ‘The system is easy to use’. While an in-depth examination of computer usability is beyond the scope of this thesis, PGL was examined concerning its general ease of use. PGL is capable of providing federation of maritime datasets and targeting search results based on a metadata search, however it is limited in its usefulness to the public due to the complexity of the system. This complexity is revealed in two areas: initial setup of the system, and the user interface. The prime goal of PGL as well as SRB, is to federate datasets. In the federated data model, the datasets remain at their original location. The MCAT in SRB understands where the data is located, and when the data is requested, the system performs a query against the data to retrieve it to the user’s Web browser. This process requires that each location maintain a copy of SRB, running on their host machine. Since SRB requires relatively few computer resources, an older computer is capable of hosting this application. In the scenario discussed in this research, this organization scheme would take the form depicted in Figure 5.1 below. In order to set up SRB on each of the host machines, the assistance a system administrator who had been trained in the use of SRB is required. This is not unusual in that most Web based applications require some sort of ‘Webmaster’ or other authority that manages the system.

The second usability concern with PGL is in regards to the user interface. There are two areas of the application that cause difficulty for an untrained user: mapping dataset fields, and conducting a search against the stored datasets. To map fields between two datasets, the user must have a precise knowledge of the structure of the datasets regarding field names. While the users may have a good understanding of the format of the data on

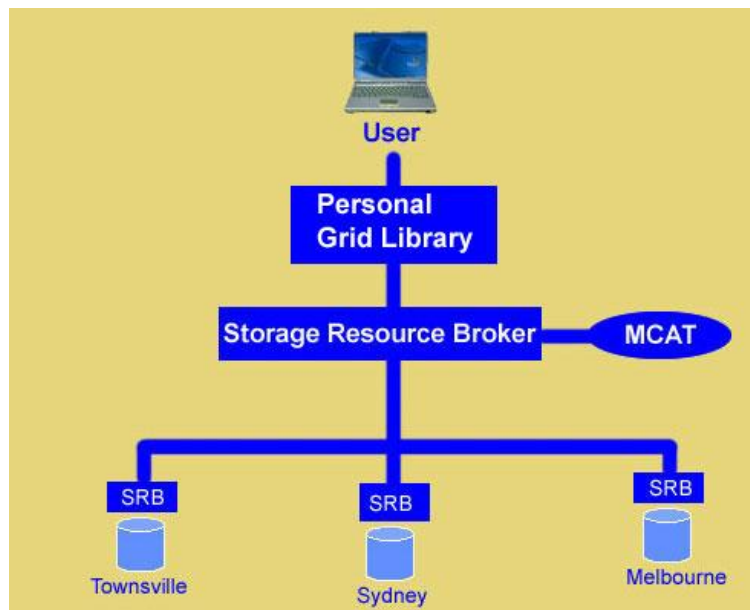


Figure 5.1 – Architecture of PGL using SRB

their host servers, it is not likely that they would understand the format of datasets with which they are not familiar.

Thus, the mapping of individual datasets against the generic fields known to the system would again require the intervention of an administrator. Metadata templates for multiple datasets can be created, but this would necessitate the assistance of someone trained in XML and XMLSchemas. Another area of difficulty with the user interface concerns the search utility within PGL. The search interface that the researchers would use to discover data is very cryptic. There is no instruction on the screen to inform a user of the steps necessary to perform a search. The keywords or metadata used to perform the searches are not made available to the user. Since the user is not prompted with suggestions for keywords, many searches return no records. It is not obvious to the user whether no records were returned because the keyword is not found, or whether there is no data in the system that matches the search criteria. It is also possible for a search to not return any records due to restrictions placed on a resource at the data access level. Currently there is no way to inform the user that additional information is available, and the process through which the user may request access to the data. An issue of concern in regard to returning usable search results hinges upon the format of the data in the sample datasets. Many inconsistencies between data within the datasets made it difficult to return data matching the search criteria. Since this is not only an issue for PGL, but all data sharing systems in general, this topic will be addressed further in the discussion section.

Although PGL is a powerful application that is capable of allowing users to create digital libraries quickly, based on the issues mentioned here, PGL does not appear to meet the present needs of the maritime archaeological community for sharing data. Even if training was provided in order to overcome the complexity issues, a central administrator would still be required to oversee the application. After further analysis of the research problem, a second study was implemented using a system that was purpose-built to handle the issue of maritime data sharing. The results of Study 2 are detailed in the next section.

5.2.2 Study 2: ArchaeoView

Background details regarding the system

- Application designed for the maritime archaeological discipline
- Application is multi-platform (runs in Windows, Macintosh, Linux and others)
- Data is federated through the use of SRB

- System has main two components: a data upload facility and a search engine
- Much of the complexity of relating databases to each other occurs behind-the-scenes
- Each dataset requires a separate schema document
- Application can read the contents of a database and return search results regarding the internal information

The ArchaeoView system was designed to meet the needs of archaeology researchers, keeping in mind the previously defined use cases (see Table 5.1). The prototype system is a Web based application that works with SRB in a similar fashion to PGL, but attempts to present a more user friendly interface to the researcher. The system has two components: a data upload facility and a search engine. Each piece works together to handle the system requirements for data federation and data retrieval through semantic search. A description of the user interface was provided in section 4.6.

Like many computer applications, a considerable amount of work is done behind the scenes before the system is ready to handle a user request. In the ArchaeoView application when the system starts, it loads schema files stored in the application. These schema files hold the individual database structure of each dataset. Next, it uses this schema information to run a query against the combined datasets present in the data store. It uses information from the MCAT and SRB's Data Access Interface (DAI) to extract a snapshot of the collective data. This snapshot is sent to the Query Generator. Now the system is ready to handle a user request.

When a user enters a text string using the search engine, the Query Generator uses the underlying schema information to scan through the combined data and return everything that matches the user's search request. This combined information is returned to the user on the results page. Through this process it is possible for the user to build a quite complicated search query from simple atomic components such as keywords.

ArchaeoView is similar to PGL regarding data retrieval. In PGL, the only way to allow searches inside databases is to use the mapping interface to form a connection between fields in the databases which may have similar data. For instance the NSW data has a field called 'Name of Ship' and in the VIC data this field is called 'Title'. The mapping utility allows the system to expand the query at run-time to include these fields. This requires that the data depositor have an understanding of which fields in a database match those in the

central system. Although much of the complexity has been behind the scenes by using a separate database schema for each dataset, the user is still required to map the fields in the dataset to a master set of field names stored in the ArchaeoView database.

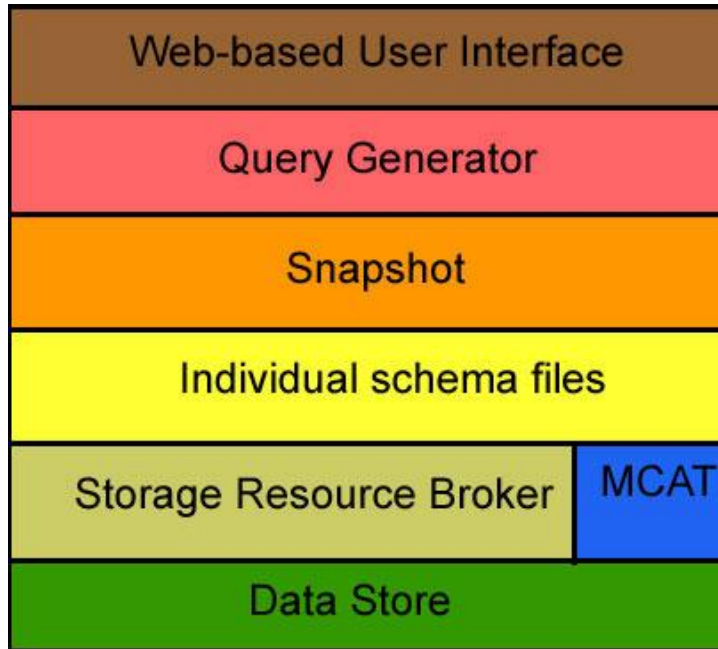


Figure 5.2 – Data structure within ArchaeoView

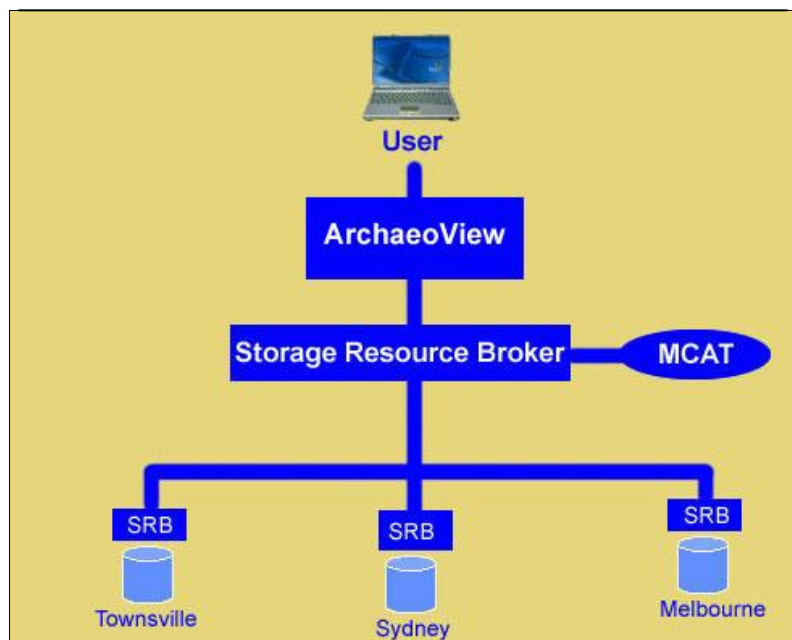


Figure 5.3 – System Architecture of ArchaeoView

Requirements for depositing datasets

To notify ArchaeoView of the location of a dataset, the user clicks a browse button to select the file on the data upload screen. Next the researcher must specify a database schema that describes the structure of the database that is being uploaded. This database schema is written in XMLSchema format and lists the field names, data types and order of the fields within the dataset. This is needed by the system for querying purposes. In the final step, the user maps the names of the fields in the dataset to those listed in the master schema for the application.

Data sharing effectiveness

- Data can be federated using ArchaeoView
- The data can be maintained at its current location via SRB
- In order for the system to find the data, the MCAT in SRB must be updated
- The system uses a copy of the data rather than the original data
- Replication ensures that no updates are made to the original data
- The user must provide a schema of each dataset formatted in XMLSchema
- The system is capable of creating very complex queries based on a simple user query using the database schema
- The system was able to return results from 2 of the 3 sample datasets

ArchaeoView provides a method for users to federate their datasets. Each database must be made known to SRB so that its location details are specified in the MCAT, SRB's metadata catalogue. When the ArchaeoView system starts, it is provided with a copy of the set of data from SRB. This process is called replication, and it ensures that the original data is unharmed by the users interaction with it.

In order to provide a search utility capable of sifting through the multitude of data available through federation, ArchaeoView uses semantic metadata and schema information that is provided for each of the datasets. When a data provider adds a database to the system, a schema of the underlying data is also uploaded that gives the search utility the information necessary to find the correct data. If this schema were not available to the system, the user would have to create the complex queries necessary for data extraction at the time that the search request was initiated. In a similar result to PGL, ArchaeoView was able to return the internal contents of two of the three datasets. Due to data formatting issues and a lack of correlation between the data, contents of the third dataset were not accessible.

ID	Use case description
1	User is able to deposit a dataset
2	User is able to set access rights for data
3	User is able to make data discoverable
4	User is able to find data
5	User is able to view data online
6	User is able to download data
7	System is easy to use
8	System allows searches across multiple datasets
9	System returns search results combined in one table in a web browser

Table 5.2 – Review of use cases for this study

Analysis of ArchaeoView functionality against defined use cases

- ArchaeoView contains two separate user interface views
- One interface handles the data upload process
- Another screen provides the semantic search tool
- The pairing of federation with semantic search introduces benefits as well as limitations
- Data must be correctly formatted, and uploaded into in SRB using application
- The user must provide an XMLSchema document for each dataset, this would require technical assistance
- The user must manually map the fields in the database that contain similar data to the system's generic database schema
- Mapping of data fields requires previous understanding of the system's generic schema, and most users would require training in this area.
- The reasons for mapping fields is not immediately obvious to a new user

As with PGL, ArchaeoView was evaluated against the use cases presented in Chapter 1. These use cases are provided again in Table 5.2 as an aid to the reader. Based on the results as presented above, ArchaeoView meets all of the requirements except for items 1 and 7: User is able to deposit a dataset, and System is easy to use. These functionality and usability results are identical to those obtained from the evaluation of PGL.

As mentioned previously, ArchaeoView consists of two user interface views. The first handles the process of uploading data and the second facilitates a search against the combined datasets. Each component reveals both the strengths and limitations of federation paired with semantic search. In section 4.6 it was noted that three steps are necessary to upload a dataset. These steps are reviewed below:

1. Migrate the datasets to a format suitable for SRB and the MCAT
2. Upload the datasets using ArchaeoView
3. Map the new dataset to the generic schema in ArchaeoView

Step one notes that the datasets must be appropriately formatted for upload purposes. This process requires that the databases be transferred into a format understandable by the Data Access Interface (DAI) in SRB. The DAI is the querying agent within SRB that returns the datasets. In this study the three sample maritime datasets were modified into PostGres format and were then loaded into SRB using ArchaeoView. Due to the lack of documentation and suitable tools for this task, it is likely that the user would require assistance from an IT professional in order to perform this action. Once the datasets are loaded in ArchaeoView, a further complication arises when the user interface requires that the data provider upload a suitably formatted data schema pertaining to that dataset. The data schema must be provided in XMLSchema, a technology unfamiliar to most users. Although there are open source tools available through the W3C capable of scanning a dataset and creating a schema with limited user interaction, this information is not well known to the general population, and the user would likely require additional IT assistance. One area of the interface that reinforces the difficulty of providing machine understandable connections between data sources is revealed in the data mapping section of the application. The data provider must manually select any and all fields from the ArchaeoView generic schema which match fields in his or her dataset. However, the reasoning behind this action may not be obvious. To a human it may seem reasonable that the fields 'Ship Name' and 'Name of Ship' contain the same type of data, but a computer system will not make this connection. When the differences are more abstract such as 'Name of Ship' and 'Title', this becomes even more difficult. If the user is unable to choose appropriate fields to map to the generic schema, the new dataset will not be included in any searches.

Although ArchaeoView presents a simpler, more user-friendly software interface to the researcher, the complexity of data federation tied with semantic search make both PGL and ArchaeoView difficult for an untrained researcher to use. The reasons for this difficulty lie

in both the archaeology and information technology domains. In the following discussion section this finding will be explored in more detail.

5.3 Use of Data Federation and Semantic Search with Maritime Archaeological

Datasets

In Chapter 1 the research question ‘Can existing archaeological databases be made available in a data sharing environment’ was posed. To answer this question, research was conducted using 3 sample datasets. The bulk of the research effort entailed the use of data federation and semantic search techniques. The previous section described the results that have been obtained from this research. While neither study was able to produce a system which meets all of the original criteria for use cases, system usability and functionality using the sample datasets, in the course of the research, additional information regarding the use of data federation and semantic search was revealed. This section considers the results obtained in the three areas of data federation, semantic search and system usability separately. The conclusion of the section addresses issues and barriers to implementation associated with combining the technologies of federation and semantic search and offers suggestions regarding how these obstacles can be avoided.

5.3.1 Data federation results

- Federation offers the ability to remove the ‘gatekeeper’
- Data can be made available to a restricted set of users
- Audit tracking can be used in cases of intellectual property concerns
- Implementation issues regarding existing archaeological datasets
 - Most archaeological datasets are not designed for use by a central system
 - Column headings are not descriptive enough
 - There is a high overhead associated with reformatting the data
 - Data that is not related is ignored and is ‘invisible’

The discovery of existing archaeological data is a primary step in many research projects. Ascertaining the location of data as well as requesting access to the files currently requires weeks if not months of effort. The federation of archaeological datasets provides a method of making the data available and managing access to it. Currently, the originator of the data acts as a ‘gatekeeper’, restricting access to data. Frequently requests for data are denied pending the original researcher publishing the results of his research. As publication of data sometimes lags by as much as 10 years after excavation, this situation is untenable. Data federation systems can remove the gatekeeper by allowing researchers to make certain

portions of their data available to a restricted set of individuals. Audit tracking software can be used to monitor access to materials in cases where intellectual property issues may be of concern (Atkinson 2006, pers conv). Once individuals become linked in a web of trust, issues regarding control of the flow of data can be alleviated.

However, implementing a federated data sharing system that returns a combined view of multiple datasets is not without its drawbacks. Most archaeological datasets are not designed in a manner that will allow them to be easily used by a central system. For example, many databases consist of a multiple tables. In order for the querying system to obtain all of the information that it may need to query each table, a precise database schema describing the format of the tables must be provided. The column headings for each table may not be sufficiently descriptive to allow a system to determine the sort of data held in the fields. There is a high overhead associated with formatting these datasets to meet the standards required for a data sharing system. In the case of the sample databases used in this project, at a minimum the databases were formatted in different database software, column headings were updated to meet the standards required for the database management application, and in some cases wholesale re-formatting of data such as dates was necessary. The adjustments to the data were time consuming and required the assistance of personnel with IT training.

Once the data is available in the system, a further difficulty arises when attempting to combine datasets which may not have relational data. Modern 'relational' databases make connections between data that is related in some way. In both PGL and ArchaeoView, in order to return a combined list of data, all of the databases queried must have data with similar content. For example, the NSW and VIC data are made up of a listing of ships wrecked along the Australian coast. The TMM data however, consists of a listing of artefacts obtained from the excavation of shipwrecks in Australia. There may well be artefacts listed which belong to a particular ship in the NSW or VIC data. However, since there is not a unique identifier in the shipwreck data that matches a similar identifier in the artefact data there is no way to make this connection. To illustrate this point, take the case of an artefact listed in the TMM data as being from a ship named the 'Mary Elizabeth'. Since there are multiple entries for 'Mary Elizabeth' in both the NSW and VIC datasets, there is no way to determine to which ship the artefact belongs.

In both of the systems used in the two studies, non-related data is ignored. Therefore the data from the TMM dataset cannot be included and thus becomes invisible or undiscoverable to the user. Since the only way that data can be retrieved is through the

search utility, this is a serious drawback. Figure 5.4 depicts the connections between the sample datasets. Section A, the intersection of the NSW and VIC circles, indicates that there are fields in the two datasets that contain related data. Analysis of these two databases showed that there are 34 related fields between these two datasets. The TMM circle is completely disconnected from the other two circles, indicating a lack of data that can be associated with the NSW and VIC datasets.

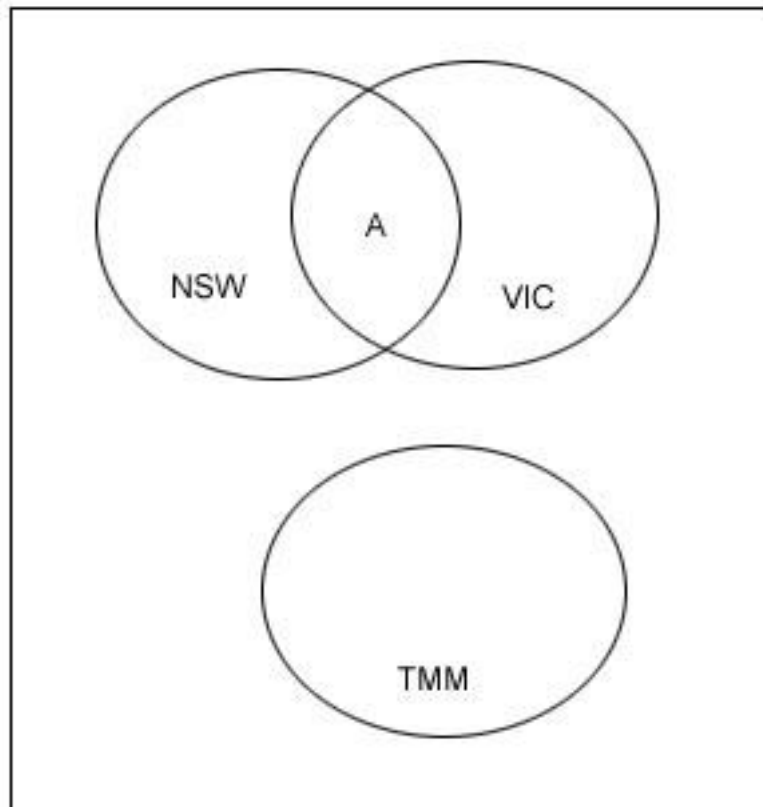


Figure 5.4 –Intersection of datasets

5.3.2 Semantic search results

- Semantic search offers the ability to limit search results appropriately
- Data searched must contain metadata and additional information used by the semantic search engine
- There is a high overhead involved with preparing datasets for use on system that provides semantic search
- Currently there is a lack of semantic tools available for the user
 - The user must create a schema representation of the data
 - W3C has not yet finished the requirements for the XMLSchema specification
- Standardization of the datasets would assist in semantic search
- The burden of work to add metadata and format the data falls on data providers

- Data entered in the field may have a different schema format than the schema of the target system where it will ultimately reside
- Due to differences in systems it is difficult if not impossible to create a universal schema
- Current middleware is unable to handle all these issues
- Front end applications specific to a particular discipline are sometimes created

The second part of the research question, i.e. targeting search results, presents another challenge. While data distributing systems have a long history in the field of information technology, search mechanisms have not kept pace. Only in the last few years has searching for data been a skill that the user has had to develop. In the past, the user generally only needed access to local datasets, for which he had a great familiarity. With the advent of the data deluge and the rise of collaborative work projects these datasets have become too numerous and complex to be searched through easily (Wyatt et al 2007).

Searches across large volumes of data using traditional search engines often retrieve data unrelated to the users research query. In order to implement semantic-enabled searching, the data being searched must contain metadata and other additional information that can be used by the search engine. There is a high overhead associated with preparing datasets for use in systems that provide federation and semantic search. A large part of the extra work needed to prepare datasets for use in such systems is caused by the lack of semantic tools. In order to add a new dataset to either of the systems studied in this thesis, the user must create an XMLSchema representation of the database. Although this schema can be created in any text editor, the knowledge necessary to complete the task would put it beyond the reach of most users.

This is a common problem that has been handled adequately in other areas by creating applications which create the necessary code without excessive user input. Take for example the creation of HTML pages. When Web pages were first used on the Internet, each page was created in a text editor and the writer was required to understand the intricacies of HTML in order to make content which was suitable for display via web browsers. Multitudes of books were written to teach the average person to create web pages. However many users found this process cumbersome and desired an application that would handle the HTML creation tasks and let them simply enter text, in a method much like the functioning of a word processing application. The creation of simple HTML editors such as HotDog, AOLPress, FrontPage and more recently Dreamweaver have

solved this problem. There are currently tools which are capable of creating an XMLSchema for a database, but they are not widely known. In addition the W3C is still in the process of completing the requirements for the XMLSchema specification and until that is complete software developers may choose to not bring out new applications which handle the schema creation task (Murata et al. 2005).

This leads to another issue regarding preparation of datasets for the semantic web. Unless a person is trained in semantic data management, it is unlikely that the researcher will have the mindset to 'think in schemas'. This is an ongoing problem with search engines and the data presented on the WWW. The HTML specification allows the use of 'tags' which can define and describe the content of a web page or site. The plan was for the keyword 'metadata' in HTML code to be used to specify the content contained in the page in order to provide more effective retrieval and relevance marking in searches. The 'description' tag would be used in the display of search results to provide a more 'user friendly' listing of the contents of the page (Gill 2000). However, as this system relied on the page creator being to add suitable metadata to the page this system failed. Users were not sufficiently trained in the use of metadata and the necessity of its use. A similar situation occurs with the semantic web. In order to add schemas, metadata, annotations or other supportive information to the data to aid searching, the burden of the work is placed on the data provider. In a situation where the data owner is already hesitant to make this data available this overhead can seem unnecessarily high.

Another detail, which hampers the usability of the semantic web, is the lack of adaptability between schemas. A schema developed for a particular dataset for use in a specific system may not be usable on a different system. Therefore a data gathering system that a researcher uses in the field to track data may have a different schema format from a federated system where the data will be placed at a later date. These differences make it difficult for a universal schema to be developed. Middleware tools such as SRB are currently unable to provide the level of assistance necessary to handle these three issues. While middleware can provide the link between users and their data, quite often front-end applications that are specific to a particular discipline must be created. This development requires a considerable amount of technical ability, time and resources that may not be available to an archaeological research unit.

5.3.3 System usability results

- A semantically enabled search engine allow targeting of results
- Burden is placed on the data provider to suitably format and update the datasets
- Complex process involved in preparing and uploading the data
 - Migrate the data into an appropriate format and upload it
 - Manually set mapping between fields in the data sets
 - In-depth understanding if the data and the host system is required
 - Time is necessary to format the data appropriately
 - Training is necessary for the users of the system
- Difficulty in creating an easy to use, generic application

As mentioned previously, many benefits are gained through the federation of datasets. Placing datasets in a federated system speeds up the process of providing access to the data, but also eliminates much of the discussion between data owner and data requestor needed to facilitate the data transfer process. The addition of semantic data to the federated system allows the targeting of search results, but places an additional burden of work on the data provider to suitably format and update the data.

The first step in the process of adding data to the federated systems is to make the data available in a format suitable for SRB. Since the only database format currently accepted by SRB is PostGres, all databases must be translated into this format. This requires the user to seek outside intervention to upload the data. Many archaeological units may not have access to this type of assistance, thus hampering their ability to use the federated system. In the next step, the data provider must notify either PGL or ArchaeoView of the location of the datasets by uploading them. In PGL, this step is may be difficult due to the complexity of the user interface. As mentioned in Chapter 4 there are no on-screen prompts to lead the user through this process. In ArchaeoView, the process is fairly explanatory, but the user must still provide and upload an XMLSchema representation of the format of the dataset. In the final step, both systems require the user is required to manually set the mapping between fields in the new dataset and the fields contained in the system's generic database. This requires an in-depth understanding of both the data being deposited and the data in the host systems. To gain this familiarity with the data the user will require training. Some users may make this connection easily, while others may consider this process too problematic.

Usability concern
Time is necessary to format data to place in system
Training is necessary to use either system
It is difficult to develop a generic system

Table 5.3 – Usability concerns for Studies 1 and 2

Three overall concerns provide the basis of the results regarding the overall usability of the systems examined in Studies 1 and 2. These concerns are listed in Table 5.3 above. Based on conversations with archaeologists in the field, it is expected that data providers would not have the time or resources necessary to format the data and upload it into the federated system. This work would need to be carried out by a technician familiar with the system. As specified in Chapter 1, archaeologists have differing access to IT support as well as training. In order to use either of the systems described in this thesis, users would need additional training specifically as it related to the process of mapping their data to the generic schema used by the system. Users would also require assistance in the creation of a schema which matches the structure of the dataset, but it is expected that this could be created by the technician who held responsibility for administering SRB.

The third concern is associated with the generic versus specific nature of the systems. As mentioned previously, SRB provides the middleware that enables communication between the users and their data. Due to the complexity of SRB, various front-end applications are created to facilitate this communication. PGL and ArchaeoView are similar in that they both fulfill this requirement. In order to provide a user-friendly interface, a system that targets a specific community and uses schemas developed for that particular group will lose the generic aspects of development that may make it usable in more than one discipline. Therefore a data sharing system developed in archaeology may not be usable by users wishing to upload medical or geological data.

5.3.4 Pairing data federation with semantic search

- Current barriers to implementation of data federation
 - Inherent complexity of the technology
 - Data federation is a complicated endeavor and user interfaces reflect this
 - Front-end applications have been created to provide a more user-centered interface
 - Systems are often limited to the domain for which they are developed
 - Generic (complicated) versus targeted (simple)

- Requires administrator with technical expertise
- Funding model to support repositories and administrators is an issue
- Current barriers to implementation of semantic search
 - The data provider must create and upload a database schema showing the structure of the database in XMLSchema
 - Technical assistance is required
 - Need for abstraction to make system easier to use, so the user doesn't need to know where the data is located

The research for this thesis has focused on the feasibility of combining a federated data sharing approach with semantic search techniques. During the course of this research two systems were evaluated as to their suitability for this task. In general the results have shown that although these systems can be used, extensive training would need to be provided to users in order for the systems to be of value. The reasons for the complexity of these user interfaces are directly related to limitations of the two technologies under discussion. In this section, the barriers to implementation related to data federation and semantic search are examined.

The difficulties in utilizing federation and semantic search are compounded by the inherent complexity of the individual technologies involved. Some of the challenges presented in this research are due to the processes required to federate data. As mentioned in the previous section, the task of federating data is not a simple one. Due to the multitude of tasks that must be handled by an application that handles the federation of data, the user interfaces for these middleware applications are correspondingly complex. To work around this complexity, additional applications such as PGL and ArchaeoView have been created with a view towards providing a more user-centered application. However, these applications are highly dependent on the area for which they were developed and have limited value to those outside the domain. This has the effect of causing a generic system to appear complex and thereby require additional training of the users in order for it to be feasible. Therefore, any software that is developed using federation techniques must weigh the benefits of a generic interface versus a user-friendly one.

Any federated system, due to its technical nature, must be managed by an administrator. The administrator would need to have familiarity with all of the datasets within the system, and have the technical ability to upload data and set access requirements as determined by the data providers. In a situation such as the maritime archaeologists who may have limited access to technical assistance, this scenario may not be feasible. Currently no individual

group has the means to set up such a system and to maintain it once it has been built. A main concern addressed in conversation with archaeologists at the 2005 AIMA/AAA conference in Freemantle, Western Australia regarding the creation of such a system, was the funding model for participation.

One of the prime issues associated with the federation of data is the concept of abstraction. The process of abstraction refers to the act of hiding extraneous detail concerning an object. This is done to reduce the complexity of a system and to increase efficiency. In the case of a federated system, the original datasets are stored in separate, often geographically distributed locations. Rather than requiring the user to know the actual location and name of a data file, the system handles access to this data. In SRB this is handled by the MCAT. The data file is given a logical name which is understandable to the user, but which may be different than the actual file name. When the user needs the file he requests the logical file name, and wherever it is store it is retrieved. The user doesn't need to know what type of file system it is stored on, or even where the files are located. In federated systems files are often replicated or copied in order to provide faster access to the data. This also backs up the file in case there is damage to a portion of the system. Data abstraction is of benefit to both the data provider and the researcher attempting to gain information, however it does add to the overall complexity of the system.

While federation of datasets can be problematic, with the limited assistance of a technician, the use of federated system can provide many benefits. Semantic search however has several issues that pose barriers that may be more difficult to overcome. Each of the applications examined in the studies require that the data provider create and upload a database schema showing the structure of the database. For PGL and ArchaeoView, the schema must be formatted in XMLSchema. Other systems may require more complex schemas as well as ontologies which use Resource Description Framework (RDF) or Web Ontology Language (OWL). In either case technical assistance in creating an ontology or schema is necessary in order for a semantically enable search engine to retrieve the data. Proponents of the semantic web such as Berners-Lee suggest that the RDF specification contains enough flexibility to handle various domains, but at the present there are few production systems in place with use semantic tools. Many groups are working on the problem of making the semantic web accessible to the general public, but as of yet appears to be no conclusive solution to this problem (Lacy 2005). Data federation on its own is a fairly complex domain, but adding semantic search to it makes the problem even more complicated. It currently requires a specially trained technician to create database schemas and ontologies and to upload them along with the associated datasets in order to make

semantic search available. The creation of application that is capable of handling these tasks will make semantic search more accessible to researchers. However, at the present these tools are either quite expensive, or not generally available.

5.4 Research Implications

The results from the research indicate that although there are difficulties inherent in the process of federating maritime archaeological datasets, this process would be of benefit to the maritime archaeological community. However, several issues currently exist that make the sharing of data problematic. The remainder of this chapter focuses on the steps that would need to be taken by each community (maritime archaeology and information technology) in order to make data sharing feasible.

5.4.1 Maritime archaeology concerns

- Resolving data consistency issues
- Handling legacy data
- Creation of a data standard
- Ontology development
- e-Research infrastructure

The sample maritime archaeological datasets provided for this research have many inconsistencies in the manner in which data is formatted and stored (see Chapter 4 for details). These dataset discrepancies make it difficult for a central data system to query their contents. The ‘traditional’ IT methodology for handling data formatting issues is to restructure the datasets to meet a standard, and then deposit them in the system. The difficulty imposed by reformatting data to meet a standard is that this is very time consuming (i.e. expensive) to implement for large amounts of data. To follow this process with all of the existing maritime archaeological datasets would require a mammoth investment of time and resources from the maritime archaeology community. Based on the input received from the survey of archaeologists, it is unlikely that this option would be feasible.

Another method of handling this issue would be to implement a two-pronged solution: 1) upload the older unformatted data with sufficient metadata to allow searches and 2) develop standards for datasets created in the future so that their contents can be easily searched. Both systems evaluated in this study have the potential to allow single documents to be uploaded, and to be searched based on metadata entries for that resource. This places the burden of creating metadata on the person uploading the data, but other digital libraries

have used this technique without great difficulty. This method is currently in use by the ADS in the UK for British archaeological datasets. To implement the second part of this option, i.e. develop standards for the creation of future datasets, the archaeological community would need to come to an agreement regarding metadata terms and the basic structure of maritime archaeological datasets. The European archaeological community has moved toward this type of standardization by establishing the Core Data Standard for Sites and Monuments (ICOM-CIDOC 2005) and the English Heritage National Monuments Record Thesauri. Although these standards have been implemented for the cultural heritage arena, they could be extended to include other archaeological endeavors. The geosciences have employed a much more precise data standard called SO 19115. This standard specifies mandatory and conditional metadata sections, metadata entities, and metadata elements required to support data discovery, access and transfer. The creation of a similar system for archaeological data would eliminate many of the difficulties involved in sharing archaeological data via a federated system. A corollary of this issue is the creation of data schemas or ontologies for use with archaeological data. Once a standardized way creating data is specified, it will be must easier to create an ontology that mirrors this standard. This will aid tremendously in the development of semantic search tools for searching across archaeological datasets.

One issue that cannot be avoided in this discussion is the need for an infrastructure to support e-Research efforts in archaeology. At present the archaeological community is not able to fund the entire development effort for tools to support the archaeological portion of e-Research. The NCRIS research projects (see Chapter 3) regarding the use of e-Research tools in the humanities has the potential to provide access to the infrastructure, hardware and the personnel needed to engage in such technologically challenging projects (NCRIS 2004). Participation in this scheme as well as other forthcoming governmental funding opportunities should be considered.

5.4.2 Information technology concerns

- Use combination of search techniques
- Metadata search
- Internal content search
- Schema creation tools
- Better documentation of middleware tools such as SRB
- Handling multiple table datasets in search utilities

A particular issue with direct bearing on this research was the choice to return a subset of data from the combined datasets. This combinatorial method, while it showed promise, was not feasible given the existing state of the maritime data. Data that could not be related to another data set in a query was essentially ‘invisible’ and thus could not be made available for download. In order to work around this issue, the two part solution suggested earlier would be of value. In that scenario, the end result would be a system that allows two types of searches: 1) a metadata search and 2) an internal contents search. Documents that are not capable of being searched in a combinatorial manner due to data incompatibility could be accessed via a metadata search. Returning a combined subset of data that provides links to enable download of the original dataset moves the evaluation of the suitability of the data earlier in the research process. This has the potential to save time, but is ultimately a limitation if vast amounts of data are eliminated from the search due to formatting difficulties. Therefore a combination of metadata with internal contents search seems the most reasonable method to follow.

The current status of data federation and semantic search technologies involves a substantial amount of manipulation of the data in order to make it available for input into the system. Even if the databases are formatted correctly as to type and content, the creation of schemas to describe the datasets are a continuing issue. While it is design choice to decide whether a schema creation tool should be embedded in the target system, access to such a tool will need to be available to either the users or the system administrators. The set up of federation systems such as SRB requires a large amount of training on the part of system administrators. A complete set of documentation regarding the process of setting up SRB, connecting front-end applications to it, and the provision of metadata and schema creation tools would be of great benefit.

An area that will require further research regarding data federation and semantic search is the issue of accessing data that resides in multiple tables within a single database. In the NSW datasets, quite often the data needed to complete a query is found in separate tables. For instance, in order to list the ‘CountryBuilt’, ‘Name of Ship’ and ‘Type of Ship’, the query would need to access three different tables within the same dataset. In a traditional database application, the user would use a query or reporting tool capable of accessing each of the tables and create any new queries as needed. This would require however, that the user have a thorough knowledge of the structure of the data in the tables. In the case of a data sharing application, the user searching for a particular type of information would not

know the structure of the data, and would most likely not even know from what dataset the data was being retrieved.

In the next and final chapter, a summary of the conclusions derived from this research is detailed as well as a discussion regarding the direction of further research in data federation and semantic search.

Chapter 6 – Conclusions and Further Research

The previous chapters in this thesis have detailed research conducted to examine the effectiveness of combining federation of maritime archaeological data with semantic search to enable archaeologists to share data. This final chapter focuses on the conclusions that can be drawn from this research and looks at the directions that future work may take. In Chapter 1, it was stated that various archaeological datasets exist. The central research question for the thesis has been to discover whether this data can be made available via a data sharing mechanism. Two additional research questions have also been considered:

1. Are there tools available to federate or combine these datasets?
2. How can the search results be appropriately targeted when searching across a variety of data sources?

In the process of conducting this research two studies were implemented using sample datasets provided by three maritime archaeological research groups. Each study considered the effectiveness of an individual software program in regard to the research questions. The applications were evaluated on the basis of providing a mechanism for the federation of data and permitting semantic searching of data. The results of the two studies were provided in Chapter 5. This chapter provides a summary of conclusions, and recommendations for areas of future research.

6.1 Summary of Conclusions

- Federation and semantic search of datasets is possible
 - Support infrastructure needed
 - Development of new tools required
- Issues regarding effectiveness of combining results for search
- Three existing methods for search
- Suggested model for search in future data sharing systems
- Specific improvements needed

The results of the research indicate that it is possible to allow access to maritime archaeological datasets using a data federation system that also employs semantic search. However, the usability of such a system can be problematic given the current state of maritime archaeological data and the relatively low level of access to IT infrastructure and technical support by archaeologists. As detailed in Chapters 4 and 5, archaeological datasets are often created in an ad hoc fashion. This results in the production of files and

databases without a uniform structure. Due to this lack of consistency across datasets, it is very difficult to create a centralised data sharing system that is capable of handling this multitude of formats and internal organizational schemes. In addition, for an archaeologist to use a data sharing system, significant infrastructure (i.e. underlying networks, software and hardware) must be put in place. At the present these support mechanisms are still in the development stage, especially as they pertain to research in the social sciences.

A particular feature of this thesis was concerned with providing a preliminary, combined view of the datasets without requiring that the complete datasets be downloaded to the user's computer. Each of the systems evaluated allow the user to perform a cross-dataset query that searches through each dataset, and returns a combined set of data matching the query. This allows the user to view data derived from multiple datasets at one time, compare the results, and make evaluations regarding whether the original dataset is worth downloading. If the user is interested in a particular piece of data, the original dataset can be downloaded. However, a drawback to this method is revealed when a dataset does not contain data that is related to another dataset on the system. Due to this lack of relation a combinatorial search may not work, and therefore some data may not be accessible via the search engine.

Others systems which provide access to federated resources encounter similar problems, but have avoided this issue by treating individual files as separate and unrelated entities. The ADS and Google are examples of this. The ADS provides access to a variety of files dealing with archaeological reports and data from the UK. When a user enters a query, the search engine returns a list of available resources. Rather than searching inside each individual resource, the system scans through metadata associated with each file. It should be noted that the data depositor requires the assistance of technicians who provide advice regarding the correct metadata information for the files.

A slightly different approach is taken by the search engine used by Google. This system maintains an index that lists location and textual content for each document. When a site is 'spidered' or discovered by Google, the search engine imports content from each page into its database. The system then creates a list of links to resources that it returns to the user. The Google search engine scans through the content that it has stored for each page in order to provide the search results.

While these processes work well for single items, they are not useful for structured data such as databases. As with most other search engines the ADS and Google treat each

database as a single item, in essence a ‘black box’. The search engine cannot peer within the database and return a selection of data. Although the single item model is not effective for internal database queries it does have the benefit that unrelated items are available to the system at all times.

The use of the ‘single item’ model as demonstrated by ADS and Google is inadequate to handle the wide variety of resources provided in the maritime archaeological datasets. These files consist of three types of information: unstructured data such as text files, structured databases like Access and FileMaker, and data unrelated to other items in the data store. The research has revealed three methods for obtaining information using present search engine technology: simple metadata search (ADS), content scanning of single documents (Google), and query based interrogation of databases (ArchaeoView and PGL). In order to offer the capability of accessing all three types of data, the system needs to be able to use each of the three search techniques in tandem. This requires that a system treat unrelated datasets as unstructured data. Table 6.1 describes data in this type of system.

A resource can be made discoverable by the addition of metadata concerning the item, or by mapping the columns in a structured database to a generic schema provided by the application. The system should have the ability to choose which search method will be of the most benefit. This flexibility would handle the situation where a dataset ‘disappears’ because it is unrelated to other data in the system. The use of this type of system would allow the unstructured legacy data from archaeological explorations to be made available along with more structured data such as that from a GIS.

Type of Data	Description of Data	Search Method
Databases	Related to other datasets	Treat as structured data Use combinatorial query
Databases	Unrelated to other datasets	Treat as unstructured data Use metadata query
Unstructured data	Individual files such as text, images, video	Treat as unstructured data Use metadata query

Table 6.1 – Suggested methods for use of combined search technologies

6.1.1 Implementation Issues

In order to implement federated data paired with semantic search, changes will need to be made to the process of maritime archaeological data creation, modification and curation. These suggested changes are described below.

- Data standards for archaeological datasets including contextual data
- Contextual information regarding the meaning of data
- Technical support in dataset creation, upload and maintenance
- Additional software tools to aid in data federation and semantic search

Many archaeological databases are created solely with the needs of the original research team in mind. Little consideration is given to how this data may be used in the future by other archaeologists. However, as funding models change and mandate the sharing of resources, mechanisms must be put in place to allow the transmission of archaeological information. This process will require both the creation of data standards and adherence to those standards by researchers. Along with standards for file creation, greater consideration should be given to contextual information. Any metadata, or additional provenance details, needed by a researcher to make sense of the data should accompany the resource.

The archaeological community will require IT support regarding the data life cycle of databases. Currently researchers do not receive adequate support regarding dataset creation and curation. Access to this support will be needed if researchers wish to share data. The use of a data sharing system would aid in the delivery of data and reduce the time necessary to discover datasets of interest to an archaeologist's research question. Although some initial concerns have been raised by individual researchers regarding data access rights, the federation of data is perceived to be of great benefit to the archaeological community by resolving data access issues and providing online search mechanisms for data retrieval. Restriction to data access can be handled via authentication schemes that set specific data sharing rights.

To enable data sharing, additional software tools will be required to assist archaeologists to prepare resources for deposit in semantically enabled systems. Currently, to upload a dataset into an application that uses a semantic search engine, the user must have in-depth knowledge regarding the structure of the data as well as an understanding of metadata creation and schemas. The development of software tools capable of handling the creation of metadata and associated semantic documents would greatly simplify this process.

However, it is likely that archaeologists would still require training in order to learn to use these semantic tools. Thus the continued need for access to IT support is a matter of importance.

6.2 Recommendations for future work

Field research is an integral part of the data gathering process for terrestrial and maritime archaeology. Progress is currently being made to transition from pen and paper recording methods to the use of digitized tools such as PDAs, notebook computers and GIS systems. New types of archaeological investigation offer the potential to radically increase the amount of data that form the existing archaeological knowledge base. To develop a usable system for collaborative data sharing it will be necessary to establish a consistent structure for archaeological data and to develop the tools necessary for the deposition and sharing of that data. This will require the following steps:

- Develop a method to handle existing data
- Develop extensible standards for archaeological data
- Establish a data repository for archaeological data
- Develop schemas for archaeology
- Establish a repository for archaeology schemas

Develop a method to handle existing data

As mentioned in the results chapter (Chapter 5), it would be an extremely expensive and time consuming process to manually reformat all existing datasets to meet some arbitrary data standard. A more reasonable method of handling the data inconsistencies found in the current archaeological datasets is to allow them to be uploaded in their current state, and assist the data provider in the creation of appropriate metadata for each dataset so that it can be discovered using a semantically enabled search engine.

Develop extensible standards for archaeological data

Going forward with new datasets however, it is imperative that an extensible data standard be established for use in the creation of archaeological data. Standards can provide interoperability between software systems, define the formats that a dataset can be stored in, describe the resource so that it can be discovered and specify other details regarding how the data is managed or archived. It is important to create data standards that specify the use of non-proprietary or open source formats because these have a greater chance of remaining accessible after an extended period. Rigid standards often have difficulty scaling or

growing as technology changes, therefore, it is important that standards remain capable of being extended as the data sharing environment changes and matures. Due to the complexity and breadth of archaeological investigations it is not likely that the creation of one over-arching standard is feasible. Currently there are over 50 suggested standards for data creation listed on the ADS site. These could form the basis of data sharing standards and formats for archaeological data worldwide. The difficulty does not appear to lie in the creation of standards, but in persuading the members of a community to adhere to them.

Establish a data repository for archaeological data

In order to provide a data sharing solution for archaeological data, it will be necessary to establish a data repository system. Many disciplines have created similar repositories, particularly in the fields of medicine and the biosciences. This thesis has focused on the technical requirements for such a data store, but the social considerations are also numerous. Some of the questions that will need to be answered before such an application can be built are listed below:

- Who will own the data?
- What researchers and projects will have access to the data?
- How will intellectual property issues be addressed?
- Should this be an international effort or country by country (e.g. the ADS)?
- How will the repository be funded? Subscription? Government funded?
- How will the infrastructure and support staff be funded?

Although the creation and maintenance of a data repository is not a trivial effort, there are many benefits to be obtained from such a system. Once data standards have been put in place, a consistent set of identifiers can be used across the combined datasets making cross-dataset analysis much easier. In addition, it is likely that once intellectual property and data access restrictions are negotiated that collaborative partnerships between researchers can be more easily established. The data that has been deposited in the system can be monitored and curated with much less effort due to the data all meeting a common standard. Data from multiple research projects can then form a rich resource for future researchers.

Develop schemas for archaeology

The first steps toward archaeological schemas can be found in the creation of the English Heritage National Monument Register and other similar thesauri. These schemas can be used by data sharing applications to inform the search engine as to data structure,

formatting, and content. Chapter 3 described the use of schemas and ontologies to inform semantic search applications. Once standards for archaeological datasets have been agreed upon, standard schemas for archaeological datasets can be created. The results from this research have shown that while XMLSchema is the current industry standard for the creation of schemas; tools that automate the creation of these schemas are not readily available. If these tools can be developed this will allow researchers to easily create a schema based on the structure of their data that can be used by a semantically-enabled search engine to provide access to their files via a federated system.

Establish a repository for archaeological schemas

One outcome of making schema development tools available would be the creation of schemas of interest to a larger community through the creation of a schema repository. By sharing existing schemas, the work necessary to modify a schema for a particular data sharing environment is reduced. In addition, schema repositories can provide an authoritative version of a schema that should be used by data providers, creating consistency in data access methods. This will aid in the creation of common tools across a data repository and better management of metadata.

6.3 Implications for archaeological methodology

The use of e-Research tools such as on and off-line data sharing systems has the potential to improve data gathering techniques used in archaeological investigations. Currently, data gathering techniques for field archaeology tend to follow the workflow listed below:

1. Data is created and recorded in a field notebook or other paper form
2. Data is entered into a temporary data file by another member of the team
3. Data is returned the office
4. Data is later entered into a permanent storage file by yet another researcher
5. Copies of the data are disseminated to other researchers on a per-request basis

Based on this general workflow, at least three versions of any dataset are created. Due to the lag time in entering the data from one system (a notebook or PDA) to another (database or spreadsheet), data are often lost, overwritten or entered incorrectly. In addition, members of a team often spend time determining which version of data is the most correct. The use of a digital data-capturing tool in the field that is connected to a data repository would remove the chance of creating multiple erroneous versions of data. Once the researcher has returned from the field the validity and completeness of the data could be verified before making it available to other members of the team and then to other researchers who have

interest in the data. This would let the researcher focus on data gathering rather than data entry.

e-Research has the potential to streamline archaeological research strategies. For example, rather than devoting days searching through off-line archives and resources, a researcher could address a query to a federated and semantically trained data sharing system and receive the results in a matter of minutes. Other examples include the use of digital photography, GIS systems and online web cams make available 3-dimensional views of remote objects via a Web browser. While due diligence will always be required in obtaining data for project plans and other reports, making data easily available to others will result in tremendous time savings over the lifetime of an archaeological project. In addition, the funding model for many disciplines is undergoing a change due to the requirements of funding agencies. It becoming increasingly common that grants contain provisos that mandate that the data obtained from work funded by the grant be made available to the funding body and the community in general (Sargent 2005).

Although a data sharing system can answer many questions, it can only access the data that has been deposited in it. In the last chapter, much emphasis was placed upon the concept of removing the 'gatekeeper' from the data sharing process by placing the data into a federated system. In essence, through the creation of a federated and suitably secured data store, the archaeologist would no longer have to personally oversee every request for data, and could pass on this chore to the data system. By setting up access rights to the data the researcher can still specify a very limited group of people to be granted access to the archaeological data. However, if this select group of users is too restricted, the gatekeeper is still in place, only in a technologically different guise. This brings us back to the basic issue of convincing archaeologists to share their data (by whatever means).

6.4 Conclusion

The research for this thesis has shown that while the creation of a data sharing system to provide access to existing maritime archaeological data is possible, it will require significant input from both the archaeology and the information technology arenas. The analogy of the Internet as an 'Information Superhighway' is apt and applies to broader context of e-Research as well. Just as standards for roads and highways are established and certain requirements are put in place for the use of them, standards and infrastructure will need to be established in order to allow researchers to share data in a productive and safe manner. The benefits obtained in creating such systems will not only be found in the discipline of archaeology but in other fields as well.

References

Adams, MD & Venter, JC 1996, 'Should Non-Peer-Reviewed Raw DNA Sequence Data Release Be Forced on the Scientific Community?', *Science* 25, vol 274, No 5287, pp. 534 – 536.

Antoniou, G & van Harmelen, F 2004, *A Semantic Web Primer*, MIT Press, Cambridge.

Archaeological Data Service, Strategies for Digital Data: A Survey of User Needs, viewed 15 June 2005, <<http://ads.ahds.ac.uk/project/strategies/4.htm>>.

Australian Digital Theses Program, Advanced Search, viewed 10 July 2006, <<http://www.adt.caul.edu.au/homesearch/advancedsearch/>>.

Australian National Shipwreck Database, National Shipwreck Database Home Page, viewed 23 August 2005, <<http://eied.deh.gov.au/nsd/public/welcome.cfm>>.

Baca, M, 2006, *Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images*, Visual Resources Association, Chicago.

Berners-Lee, T 1998, What a Semantic Web can represent, viewed 30 June 2006, <<http://www.w3.org/DesignIssues/RDFnot.html>>.

Berners-Lee, T 1999, *Weaving the Web. The original design and ultimate destiny of the World Wide Web by its inventor*, Harper, San Francisco.

Berners-Lee, T, Hendler, J & Lassila, O 2001, 'The semantic web', *Scientific American*, viewed 23 May 2006, <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2>.

Borgman, CL 1999 'What are Digital Libraries: Competing Visions', *Information Processing and Management* 35, pp. 227-243.

Brandt, R, Groenewoudt, B and Kvamme, K 1992, *An experiment in archaeological site location: modeling in the Netherlands*, *World Archaeology* 24: 268-82.

Codd, EF 1970, 'A Relational Model of Data for Large Shared Data Banks', *Communications of the ACM*, Vol 13, No 6, pp. 377-387.

Cohen, D, Lindvall, M & Costa, P 2004, 'An introduction to agile methods', *Advances in Computers*, pp. 1-66, Elsevier Science, New York.

Claxton, JB, 1995, 'Future enhancements to GIS: implications for archaeological theory' in GR Lock and G Stancic (eds.) *Archaeology and Geographic Information Systems, a European Perspective*, CRC, London.

Conolly, J and Lake, M 2006, *Geographical Information Systems in Archaeology*, Cambridge University Press, Cambridge.

Delgado, J (ed) 1997, 'Maritime archaeology'. *Encyclopaedia of Underwater and Maritime Archaeology*, pp. 259-260, British Museum, London.

De Roure, D, Jennings, NR and Shadbolt, N 2005, 'The Semantic Grid: Past, Present and Future', *Proceedings of the IEEE*, vol 93 no 3, pp. 669-681.

Doerr, M 2001, 'Semantic Problems of Thesaurus Mapping', *Journal of Digital Information*, vol 1 no 8, article 52.

Finin, T & Joshi, A 2002, 'Agents, Trust, and Information Access on the Semantic Web', *SIGMOD Record*, vol. 31, no 4, pp. 30-35.

Gethin, P 2001, 'Why the Bath Profile Makes 39.50 Work', *Liber Quarterly*, vol 11, pp. 372-381.

Gibbs, M 2004, 'Maritime archaeology in Australia', *Archaeology*, pp. 36-54, Australian Scholarly Publishing, Melbourne.

Gill, T 2000, Metadata and the World Wide Web in *Introduction to Metadata: Pathways to Digital Information*, edited by Murtha Baca. Available at <http://www.getty.edu/research/conducting_research/standards/intrometadata/metadata.html>, accessed 30 June 2005.

Gilliland-Swetland, AJ 2000, Setting the Stage in *Introduction to Metadata: Pathways to Digital Information*, edited by Murtha Baca. Available at http://www.getty.edu/research/conducting_research/standards/intrometadata/setting.html, accessed 30 June 2005.

Goble, C & De Roure, D 2002, 'The Grid: An Application of the Semantic Web', *iSIGMOD Record*, vol 31, no 4, pp. 65-70.

Green, J & Vosmer, T 1992, *The Australian shipwreck database; an interim report*. Western Australian Maritime Museum, viewed 15 June 2005, <http://www.museum.wa.gov.au/collections/maritime/march/shipdb.asp>>.

Griffiths, J M 1998, 'Why the Web is not a library', B.L. Hawkins and P. Battin (eds), *The Mirage of Continuity: Reconfiguring Academic Information Resources for the Twenty-First Century*, Council on Library and Information Resources, Washington, D.C.

Gruber, TR 1993, 'A translation approach to portable ontologies', Knowledge Acquisition, vol 5 no 2, pp. 199-220.

Guy, M., Powell, A. and Day, A. 2004 'Improving the Quality of Metadata in Eprint Archives', Ariadne 38, January 2004.

Hammer, S and Favaro, J, 1996, 'Z39.50 and the World Wide Web', D-Lib Magazine,.

Harris, J, Judging the Likely Success of an Ontology, viewed 23 October 2006, <http://www.virtualtravelog.net/entries/2004/01/judging_the_likely_success_of_an_ontology.html>.

Henderson, G1986, Maritime archaeology in Australia, University of Western Australia Press, Nedlands.

Hey, AJG and Trefethen, AE, 'The UK e-Science Core Program and the Grid,' Future Generation Computer Systems, vol. 18, no. 8, Oct. 2002, pp. 1017-1031.

Hey, AJG and Trefethen, AE 2003, 'The Data Deluge: An e-Science Perspective', Berman, F, Fox, GC, and Hey, AJG, (eds), Grid Computing - Making the Global Infrastructure a Reality, pp. 809-824. Wiley and Sons.

Heyworth, M. P., Ross, S., and Richards, J. D. (1995). Internet Archaeology: An international electronic journal for archaeology. *The Field Archaeologist* 24: 12–13.

Hosty, K & Stuart, I 1994, 'Maritime archaeology over the last twenty years', Australian Archaeology, vol 3, pp 45-54.

Hunter, J & Iannella, R, 'The Application of Metadata Standards to Video Indexing', Second European Conference on Research and Advanced Technology for Digital Libraries ECDL 1998, Crete, Greece, viewed 20 June 2005, <<http://www.itee.uq.edu.au/~jane/jane-hunter/ECDL2/final.html>>.

Hunter, J 2003, 'Working Towards MetaUtopia – A Survey of Current Metadata Research', *Library Trends, Organizing the Internet*, Torok, A (ed) 2003, viewed 23 June 2005, <http://archive.dstc.edu.au/RDU/staff/jane-hunter/LibTrends_paper.pdf>.

International Committee for Documentation of the International Council of Museums (ICOM-CIDOC), 'Introduction to the International Draft Core Data Standard for Archaeological Sites and Monuments', viewed 23 May 2006, <<http://www.willpowerinfo.myby.co.uk/cidoc/arch0.htm>>.

Internet Archaeology, 'Internet Archaeology Electronic Journal Home Page', viewed 12 March 2006, <<http://intarch.ac.uk/>>.

- International Organization for Standardization (ISO), International Organization for Standardization (ISO) homepage. Viewed 23 May 2006 < <http://www.iso.org/>
- JCU Picture Archive, viewed 23 September 2006, < <http://digitalarchive.jcu.edu.au/>.
- Juhnyoung, L and Goodwin, R 2006, 'Ontology Management for Large –Scale Enterprise Systems', *Electronic Commerce Research and Applications*, vol 5 no 1, Spring 2006, pp. 2-15.
- Kent, W 1983, 'A Simple Guide to Five Normal Forms in Relational Database Theory', *Communications of the ACM*, vol 26, pp. 120-125.
- Kvamme, K 1990, 'The fundamental principles and practice of predictive archaeological modelling', in A. Voorrips (ed.) *Mathematics and Information Science in Archaeology: A Flexible Framework* (Studies in Modern Archaeology 3), Bonn, Holos-Verlag, pp. 297-305.
- Lacy, LW 2005, *OWL: Representing Information Using the Web Ontology Language*, Trafford Publishing, Victoria.
- Lawrence, S and Giles CL, 1998, Searching the World Wide Web, *Science*, vol: 280, pp. 98-100.
- Maritime Heritage Online: New South Wales, viewed 15 April 2006, <http://maritime.heritage.nsw.gov.au/public/welcome.cfm>.
- Moore, RW in R. Boisvert, P. Tang, 'The Architecture of Scientific Software,' pp. 273-284, Kluwer Academic Publishers, 2001.
- Murata, M, Lee, D, Mani, M and Kawaguchi, K 2005 'Taxonomy of XML schema languages using formal language theory', *ACM Transactions on Internet Technology (TOIT)*, vol 5 no 4, pp. 660 – 704.
- Oppenheim. C and Smithson, D 1999, 'What is the hybrid library?', *Journal of Information Science*, vol 25 no 2, pp. 97-112.
- Oppenheim, C, Morris, A, McKnight, C, Lowley, S 2000, 'The Evaluation of WWW Search Engines', *Journal of Documentation*, vol 56 no 2, pp. 190-211.
- Queensland Museum: Pandora Artefact Database, viewed 13 May 2005, <<http://www.qm.qld.gov.au/features/pandora/search/index.asp>>
- Muckelroy, K 1978, Maritime Archaeology, Cambridge University Press, Cambridge.*
- National Research Council 1999, *Funding a Revolution: Government Support for Computer Research*, National Academies Press, Washington D.C.
- NCRIS 2004, 'National Collaborative Research Infrastructure Strategy Strategic Roadmap', viewed 25 June 2006, <<http://www.pfc.org.au/wiki/pub/Main/Documents/NCRISStrategicRoadmap.pdf>>.
- Novotny, J, Russell, M and Wehrens, O 2004, GridSphere: a portal framework for building collaborations, *Concurrency and Computation: Practice and Experience*, Volume 16, Issue 5, Pages 503 – 513.

Noy, NF & McGuinness, DL 2001, 'Ontology Development 101: A Guide to Creating Your First Ontology', viewed 3 September 2005,
< <http://ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html>>.

PARADESIC, 'Services', viewed 26 October 2005.
< <http://paradisec.org.au/services.html>>.

Pidock, W 2003, 'What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology, and a meta-model?', viewed 20 April 2006,
< <http://www.metamodel.com/article.php?story=20030115211223271>>.

Powers, S 2003, Practical RDF: Solving Problems with the Resource Description Framework, O'Reilly and Associates, Inc, Denver.

Richard, JD, 1997, 'Preservation and re-use of digital data: the role of the Archaeology Data Service', *Antiquity*, vol 71 no 274, pp 1057(3).

Roskams, S 2001, *Excavation*, (Cambridge Manuals in Archaeology), Cambridge University Press, Cambridge.

Safari, M 2005, 'Search Engines and Resource Discovery on the Web: Is Dublin Core an Impact Factor?', *Webology*, vol 2, no 2, pp. 21-35.

Sargent, M 2005, 'An e-Research Strategic Framework: A Discussion Paper', viewed
<http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/e_research_consult/documents/discussion_paper_pdf.htm>.

Schopf, J M and Newhouse, S J 2004, State of Grid Users: 25 Conversations with UK eScience Groups', viewed 10 July 2005, http://www.nesc.ac.uk/technical_papers/UkeS-2004-08.pdf.

Schroeder, R & Fry, J 2007, 'Social Science Approaches to e-Science: Framing an Agenda', Journal of Computer Mediated Communication 12(2), article 11.

Shadbolt, N, Berners-Lee, T, Hall, W 2006, 'The Semantic Web Revisited', IEEE Intelligent Systems, vol 21 no 3, pp. 96-101.

Shirky, C 2005, 'Ontology is Overrated: Categories, Links, and Tags', viewed 23 September 2006, < http://shirky.com/writings/ontology_ouerrated.html>.

Smith, B and Mark, DM. (2001) 'Geographical categories: an ontological investigation', *International Journal of Geographical Information Science*, 15:7, 591 – 612.

Smrz, P, Sinopalnikova, A and Povolny, M. (2003) Thesauri and Ontologies for Digital Libraries, Proceedings of the 5th Russian Conference on Digital Libraries RCDL2003, St. Petersburg, Russia, viewed 15 July 2005,
< <http://rcdl2003.spbu.ru/proceedings/D3.pdf>>.

Suleman, H, Anthony Atkins, A, GonÁalves, MA. France, RK., Fox, EA, Chachra, V, Crowder, M and Young, J 2001, 'Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 1: Mission and Progress, and Part 2: Services and Research', in *D-Lib Magazine*, Vol 7, No 9, September 2001.

Swartz, A & Hendler, J 2001, The Semantic Web: A Network of Content for the Digital City, Proceedings Second Annual Digital Cities Workshop, Kyoto, Japan, viewed 25 June 2005, <<http://blogspace.com/rdf/SwartzHendler>>.

Thieberger, N 2004, Building an interactive corpus of field recordings, in Carson-Berndsen, Julie, (Eds.) *First steps for Language Documentation: Computational linguistic tools for morphology, lexicon and corpus compilation.*, pages pp. 88-89. Paris: ELRA.

Troll, D & Moen, B 2001, 'Report to the DLF on the Z39.50 Implementers' Group', viewed 15 February 2006, < <http://www.diglib.org/architectures/zig0012.htm>>.

UNESCO 2001, 'Convention on the Protection of Underwater Heritage', viewed 12 December 2006,
< http://www.unesco.org/culture/laws/underwater/html_eng/convention.shtml>.

van Beek, P., Smith, JR Ebrahimi, T, Suzuki, T, Askelof,, J 2003 'Metadata-driven multimedia access', *IEEE Signal Process. Mag.* 20(2) (March 2003) pp. 40–52.

Waters, D J 1998, 'What are digital libraries?', *CLIR (Council on Library and Information Resources)* Issues, No 4, viewed 10 January 2006,
<<http://www.clir.org/pubs/issues/issues04.html>>.

Wells, A, Pearce, J, Groom, L & Lee, B. 'Connecting and Sharing: the Emerging Role of Z39.50 in Library Networks', paper presented at the VALA Conference, Melbourne, 28th-30th January 1998, National Library of Australia, viewed 20 December 2005, <<http://www.nla.gov.au/nla/staffpaper/awells2.html>>.

Wise, AL 1997, 'The Archaeology Data Service', *CRAFT 15*, CTI Centre for History, Archaeology and Art History, University of Glasgow, UK.

Woodley, M 2002, Crosswalks: The Paths to Universal Access, in *Introduction to Metadata: Pathways to Digital Information*, edited by Murtha Baca. Available at, <http://www.getty.edu/research/conducting_research/standards/intrometadata/path.html>, accessed 30 June 2005,

The World Factbook 2006, 'Washington: Central Intelligence Agency, Appendix G', viewed 2006-08-08, < <https://www.cia.gov/cia/publications/factbook/appendix/appendix-g.html>>.

Wyatt, M, Sim, N, and Atkinson, I 2007, 'YourSRB: A Cross Platform Interface for SRB and Digital Libraries', 4th Australasian Symposium on Grid Computing and e-Research, Ballarat.

Appendix A - Glossary

1:1 principle

A term used in Dublin Core where related but conceptually different entities (like a painting and a picture of a painting) are described by separate metadata

Abstraction

A programming concept whereby the interior details of an object or component are hidden from view in order to reduce complexity and aid efficiency.

Access Control

The ability to selectively control who can view, update or otherwise manipulate information

Accession

The process of adding an item to a collection in a museum. Generally includes indexing and description of the object.

Administrative metadata

Metadata used in managing and administering information resources, especially location or access control information

ADS

Archaeological Data Service. A division of the Arts and Humanities Data Service in the UK. Archives and maintains a catalog of archaeological resources from Great Britain.

ADT

Australasian Digital Thesis program. A program in Australia that provides access to digital theses from universities.

Aggregation

The process of gathering data of a particular topic from multiple sources and returning it to the user.

AIMA

The Australasian Institute for Maritime Archaeology. The main maritime archaeological association in Australia and New Zealand.

Archaeological record

A term used in archaeology to describe all archaeological evidence including the physical remains of past human activities which archaeologists seek out and record in an attempt to analyse and reconstruct the past. Also includes all data gathered in these research endeavors.

Artefact

Any object made, modified, or used by humans.

ASCII

A code used by computers to represent English characters. This code makes it possible to transfer data from one computer to another.

Authentication

A machine process that verifies that an individual computer or object is who it claims to be.

Back-end database

A database that contains and manages data for an information system, separate from the user interface or presentation view of the data

Bio-informatics

The application of computer technology to the management of biological information. Specifically, it is the science of developing computer databases and algorithms to facilitate and expedite biological research

Browser

Software that allows the user to view information on the Web

Centralized storage

Data is held in a central location, often in a single database

Client

Any program that uses the services of another program. On the web, a web client is a program, such as a browser, editor, or search engine that reads or writes information on the web

Conservation of artefacts

A series of steps taken to stabilize and preserve an artefact from decay or further damage after it has been removed from an archaeological site

Content Management System (CMS)

A system used to manage the content of a Web site. Generally contains templates or forms which allow non-technical users to create, modify or delete documents on a Web site without the assistance of a Webmaster. Often used in an Intranet scenario to share documents not accessible outside of the private network.

Controlled vocabulary

A carefully selected list of words and phrases which are used to tag units of information so that they may be more easily retrieved by a search.

Cross-site analysis

The process of comparing data obtained from two or more archaeological sites.

Crosswalk

A chart or table that represents the semantic mapping of fields or data elements in two different databases or standards. Also called field mapping.

Cultural heritage

Physical artefacts and intangible attributes of a group or society, often inherited from past generations that are maintained in the present for the benefit of future generations.

Curation

The care, management and use of items in archaeological collections. May be extended to include data and other digital resources.

Database

A collection of information that is organized so that it can be easily accessed, managed and updated. Often contains similar types of data; i.e. sales data. Common applications are Microsoft Access, FileMaker, MySQL

Data deluge

Concept that data from instruments and other scientific and computerized calculations is creating such a volume of data that it will swamp current storage and retrieval devices.

Data discoverability

A property that describes whether an object is able to be found and accessed.

Data Life Cycle

The concept that data will progress through stages during its existence: data collection, deposit, processing and dissemination.

Data set

A collection of data, stored in a single format. For example a table stored in Microsoft Access. Also called dataset.

Data store

A collection of data accessed by a system.

Data visualisation

The process of representing abstract data as images that can aid in understanding the meaning of the data

Datum

A reference point from which measurements are made. In an archaeological setting, a datum point is established and measurements of other items in the site are described in reference to their distance from this point.

Default values

Values that are stored in a field when no other value is specified. For instance, an opening bank balance is assumed to be \$0.

Delimited

A file is considered delimited when certain characters are chosen to signify separations in the text. For instance a text file may use commas to separate columns.

De-normalization

Intentionally violating one or more of the data normalization rules in order to aid system operability or efficiency. (See normalization).

Digital library

A collection of documents in organized electronic form, available on the Internet or on CD-ROM (compact-disk read-only memory) disks. Depending on the specific library, a user may be able to access magazine articles, books, papers, images, sound files, and videos.

Digital Library

A collection of texts, images, and other data resources encoded so as to be stored, retrieved, and read by a computer, generally on the Internet, but can be offline storage mechanisms such as CD-ROM or other media.

Digital Library Federation

A consortium of libraries and related agencies who are actively developing electronic information technologies to extend their collections and services.

Digital Signature

A very large number created in such a way that it can be shown to have been created only by somebody in possession of a secret key, and only by processing a document with a particular content. It can be used for the same purposes as a person's handwritten signature on a physical document.

Distributed storage

Data is held in separated locations with little or no connections between data sources.

Dublin Core Metadata Element Set

A minimal set of metadata elements that creators or catalogers can assign to information resources, regardless of the form of those resources, which can then be used for network resource discovery, especially on the World Wide Web.

EAD

Encoded Archival Description, a document type definition that represents a structured method for grouping archival or manuscript records so that they can be retrieved more easily

EDI

Electronic Data Interchange. A pre-web standard for the electronic exchange of commercial documents

Element

A discrete or separate component of data or metadata. There are 15 metadata elements in the Dublin Core specification.

Enumeration

The process of defining all of the values that are acceptable in a certain situation.

e-Journal

A publication whose distribution is solely over the Internet.

Enterprise

Used to describe a company or group of people.

e-Research

Research that is carried out in highly distributed network environments, often using immense data sets that require high performance computing (HPC), grid computing or advanced visualization to interpret.

e-Science

Large scale science projects that are carried out through distributed global collaborations enabled by the Internet. A branch of e-Research.

Excavation

The exposure, processing, recording and sometimes removal of archaeological remains. Sometimes called a 'dig'.

Federation

A collection of distributed databases that can be searched as if they were grouped within a single system. Data appears to be located in a central data store, but is actually located in distributed locations.

Field mapping

A chart or table that represents the semantic mapping of fields or data elements in two different databases or standards

Format

The structure that the data is stored in. Example: data formatted in Access or Excel.

Functionality

The actions that a system is capable of executing. Often called functional requirements. Examples of functionality are: saving a file, searching for a file, etc.

Generalization hierarchies

A system of organizing concepts in which similar concepts are grouped together and placed 'under' more general concepts. For example the concepts 'apple' and 'pear' might fall under 'fruit'.

GIF

Graphics Interchange Format. A format for pictures transmitted pixel by pixel over the Net. Created by CompuServe, the GIF specification was put into the public domain, but Unisys found that it had a patent on the compression technology used. This stimulated the development of PNG.

GIS

Geographic Information System. Refers to computer programs for capturing, storing, checking, integrating, analysing and displaying data about the earth that is spatially referenced through maps.

Granularity

The level of detail at which an object is viewed or described.

Graphics

Two or three dimensional images, typical drawings or photographs. See GIF, PNG

Grid computing

The application of the resources of many computers in a network to a single problem at the same time - usually to a scientific or technical problem that requires a great number of computer processing cycles or access to large amounts of data

Grid section number

Used in archaeological excavations to denote a certain square of ground.

Hidden Web

Web pages which are generated from a search query rather than individual pages. Listings on commercial web sites like Amazon.com form a large part of the hidden web. These pages are not accessible from an external search engine like Google.

Heterogeneous data

Datasets which contain data that is dissimilar in content and format from other systems. Refers to multiple datasets, often stored in separate locations.

Homogenous data

Datasets which contain data that is similar in content and format. Often refers to a single database.

HTML

Hypertext Markup Language. A computer language for representing the contents of a page of hypertext; the language that most Web pages are currently written in.

HTTP

Hypertext Transfer Protocol. The standard protocol that enables users with Web browsers to access HTML documents and other data on the Web.

Hyperlink

An abbreviated reference to a "hypertext link".

Hypertext

Text with links to other text or files. Documents written as hypertext contain text that when "clicked on" by the user with a mouse, links to other documents.

Hyponyms

A word that is more specific than a given word. For example, "Golden Retriever" is a hyponym of the more general word "dog".

Ingestion

The process whereby data systems are provided data. This may be a completely unmonitored process.

Information object

A digital item or group of items referred to as a unit

In situ

In its original place

Internet

A global network of networks through which computers communicate by sending information in packets. Each network consists of computers connected by cables or wireless links.

Interpretation (museum)

The process of presenting museum collections and information to the visitor in a form so that the person can see, read, experience and understand the items.

Intranet

A part of the Internet or part of the Web used internally within a company or organization

ISP

Internet Service Provider. The party providing one with connectivity to the Internet.

Iterative

In software development, the term is used to describe a planning and development process where an application is developed in small sections called iterations. Each iteration is reviewed and critiqued by members of the team; insights gained from the critique of an iteration are used to determine the next step in development.

Java

A programming language developed by James Gosling of Sun Microsystems. Designed for portability and usability embedded in small devices, Java took off as a language for small applications.

JPEG

Joint Photographic Experts Group. This group defined a format for photographs that uses fewer bytes than the pixel-by-pixel approaches of GIF and PNG, without too much visible degradation in quality.

Load balancing

Dividing the amount of work that a computer has to do between two or more computers so that more work gets done in the same amount of time, and in general all users get served faster. Also attempts to ensure that one computer is not overloaded by too many data requests.

Light weight ontology

Usually a taxonomy consisting of a set of concepts and the hierarchical relationships between the concepts.

Link

A reference from one document to another (external link), or from one location in the same document to another internal link, that can be followed efficiently using a computer. The unit of connection in hypertext.

MARC record

A standard for machine-readable library catalogue cards

Maritime archaeology

The study of human interaction with the seas, lakes and rivers through the archaeological study of manifestations of maritime culture. (Delgado 1997).

Markup language

A formal way of annotating a document or collection of digital data using embedded tags to indicate the structure of the document.

Meronym

A word that names a part of a larger whole. For example, “brim” and “crown” are meronyms of “hat”.

Metadata

Data about data on the Web, including but not limited to authorship, classification, endorsement, policy, distribution terms, IPR, and so on. A significant use for the semantic web.

Metadata harvesting

The action taken by small software applications that navigate the Internet looking for specific content. When it is found, the source document, or a link to the resource is retrieved and ultimately passed on to a user. See ADT for more information.

Meta tag

An HTML tag that enables metadata to be embedded invisibly on Web pages.

Middleware

Software applications that perform the technical negotiations between users and their data. Form the “glue” that holds complex systems together.

Migration

In an IT context, migration involves the transfer of data from one system to another. Often uses when a legacy system become out-of-date.

Monolith

A term describing a single database or storage structure. Requires data to meet very stringent requirements regarding content and format.

Multimedia

Digital resources with a combination of text, numeric data, still and moving video, animation and sound. Often used as shorthand for non-textual files.

Namespace

An area in an XML document where custom tags are defined.

National e-Science Centre

A consortium composed of the Glasgow and Edinburgh Universities to develop advances in scientific data curation and analysis and to be a primary source of top quality systems and repositories that enable management, sharing and best use of research data.

National Shipwreck Database

all known shipwrecks in Australia and allows users to search for those historic shipwrecks protected by Commonwealth or State/Territory legislation.

Normalization rules

Database rules that attempt to ensure that duplicate records are not created, and that a change to one record does not negatively impact other related records.

Open source

Software whose source code is freely distributed and modifiable by anyone. W3C sample code is open source software

Ontologies

A set of concepts - such as things, events, and relations - that are specified in some way (such as specific natural language) in order to create an agreed-upon vocabulary for exchanging information.

Ordered taxonomy

The arrangement of items into ordered classes. For example, 'manx' is a type of 'cat', which is a type of 'animal'.

Oxford e-Science Centre

Conducts research in computational and information science in multidiscipline collaborations.

PNG

Portable Network Graphics. A format for encoding a picture pixel by pixel and sending it over the Net. A recommendation of the W3C, replacing GIF which is a proprietary format.

Protocol

A language and a set of rules that allow computers to interact in a well-defined way.

Prototype

A research and development term meaning a first system designed to test a particular framework or theory

Provenance

In archaeology, this term refers to the geographic origin of an object. Also used to describe data relating to the site where the object was discovered, GIS coordinates, and location data.

Query

A request for information from a data source.

QuickTime

Apple's propriety standard for handling video, audio, animation, graphics, text and music.

RDF

A general framework for describing any Internet resource such as a Web site and its content.

RDFSchema

Defines a set of terms that establish relationships between entities to make it easier for systems to locate data

Repository

A storage system for data. Can be distributed over many locations or a single location.

Research plan

A project plan describing the research method, potential inputs, previous research, data gathering techniques, and archival plans

Rescue salvage

A quickly mounted archaeological excavation implemented when a resource is in danger of loss or damage.

RSS

Really Simple Syndication. An XML based technology which allows websites to publish a text version of the content suitable for import into other systems.

Scaling

The ability of a dataset or system to grow naturally as the system begins to contain more data. A system which cannot scale must be rewritten or redesigned when the data grows too large for it to handle easily.

Schema

A document that describes an XML or RDF vocabulary. Any document which describes, in a formal way, a language or parameters of a language. Also used to describe the structure of a database.

Search engine

A software program that collects information taken from the content of files available over the Internet.

Semantic interoperability

The ability to search for digit information across multiple sources.

Semantic Web

The web of data with added meaning so that a computer program can learn enough about what the data means to process it

Server

A program that provides a service (typically information) to another program, called the client. A Web server holds Web pages and allows client programs to read and write them.

Site ID number

A unique number denoting a certain portion of an archaeological site.

Software agents

A complex software entity that is capable of acting with a certain degree of autonomy in order to accomplish tasks on behalf of its user. An agent is defined in terms of its behavior.

Spider

A search engine software that follows all of the links in each site to gather information regarding the location and content of each page.

SRB

See Storage Resource Broker

State based society

A group of people whose governing structure is based around a centralised bureaucratic entity.

Storage Resource Broker

Software created by the San Diego Supercomputing Centre, Storage Resource Broker (SRB) is client-server middleware that provides a uniform interface for connecting to heterogeneous data resources over a network.

String

A sequence of alphanumeric characters, for example “holidays in Byron Bay”.

Stylesheets

In IT usage, a stylesheet is an XML based document that describes how content on a website should be displayed. Common descriptors are type font, color, size.

Supercomputers

A computer that performs at or near the currently highest operational rate for computers. A supercomputer is typically used for scientific and engineering applications that must handle very large databases or do a great amount of computation (or both).

Syntactic interoperability

The ability for two different systems to communicate even though they use a different wording or syntax for terms.

Tags

Short bits of text used to indicate metadata or other structural elements in a document.

Targeting (search results)

The ability of a search engine to precisely narrow down a broad search into a more narrow one that will return a more useful set of data.

Taxonomy

The science of classification according to a pre-determined system whose resulting catalogue is used to provide a conceptual framework. Generally consists of a hierarchical grouping of concepts.

Thesaurus

A light weight ontology that contains a list of terms and similar terms from other systems. In data sharing systems, thesauri are used to handle the mapping between systems with similar content but different column names for the data.

Treasure hunters

Divers or companies who retrieve items of value from submerged sites, often shipwrecks, without the authority of the governments. These groups often attempt to pass off their efforts as archaeological investigations.

Type specification tool

A software used to aid in the creation or modification of taxonomies or ontologies.

Unambiguous

In IT usage, a unique name or ID.

UNESCO

United Nations Educational, Scientific and Cultural Organisation. A specialized United Nations agency associated with creating universal agreements on ethical standards.

Unicode

A very large character display system which supercedes ASCII. Capable of displaying and referring to character sets from any language.

URI

Universal Resource Identifier. The string (often starting with http) that can be used to identify any resource on the Web.

URL

Uniform Resource Locator. A term used to define the current location of a certain item.

URN

Uniform Resource Name. A location independent identifier for a file. No matter where the file is located, this name should be constant.

Usability

A term referring to a system's ease of use. "User friendly" systems have high usability, and are easy for a new user to learn without extensive training.

Use case

A methodology used in system analysis to identify, clarify, and organize system requirements. A use case is made up of a set of possible sequences of interactions between systems and users in a particular environment.

User interface

The part of a computer application that is visible to the user. Sometimes called "graphical" user interface, to imply that images are visible, not just text

Utility

A small software application created to solve a particular problem.

Virtualization

A process of providing a unified view of hardware and other resources regardless of their location on the network.

Visual Resources Association

A multi-disciplinary organization dedicated to furthering research and education in the field of image management within the educational, cultural heritage, and commercial environments.

W3C

World Wide Web Consortium. An international consortium which develops interoperable technologies (specifications, guidelines, software, and tools) for the Web.

Web

Abbreviated version of the term World Wide Web.

Web log

Sometimes called a 'blog', essentially an online journal.

Web Ontology Language

Abbreviated as OWL. A language used to define structured, Web-based ontologies designed to allow data contained in documents to be processed by applications.

WWW

World Wide Web. The set of all information accessible using computers and networking, each unit of information identified by a URI.

XML

Extensible Markup Language. Recommended by the W3C as a generic language for creating new markup languages. Markup languages (such as HTML) are used to represent documents with a nested, treelike structure.

XML Schema

XML Schemas express shared vocabularies and allow machines to carry out rules made by people. They provide a means for defining the structure, content and semantics of XML documents

X, Y, Z coordinates

The three coordinates necessary to fully describe the location of an object in a space. Often used in archaeology to describe the placement of an artefact within an excavation site.

Z39.50

A data transfer interface protocol used in digital libraries to transmit information.

Appendix B – Sample Ontology: XMLSchema for NSW dataset

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance">
  <xs:import namespace="http://www.w3.org/2001/XMLSchema-instance"
schemaLocation="xsi.xsd"/>
  <xs:element name="SHIPWRECKS">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" ref="ROW"/>
      </xs:sequence>
      <xs:attribute ref="xsi:noNamespaceSchemaLocation" use="required"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="ROW">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="BEAM"/>
        <xs:element ref="BUILDER"/>
        <xs:element ref="CARGO"/>
        <xs:element ref="COMMENTS"/>
        <xs:element ref="CONSTRUCTION"/>
        <xs:element ref="COUNTRYBUILT"/>
        <xs:element ref="CREW"/>
        <xs:element ref="CREWDTHS"/>
        <xs:element ref="DATEBUILT"/>
        <xs:element ref="DATEWRECKED"/>
        <xs:element ref="DEATHS"/>
        <xs:element ref="DECADEWRECKED"/>
        <xs:element ref="DRAFT"/>
        <xs:element ref="ENGINE"/>
        <xs:element ref="F1"/>
        <xs:element ref="FOUND"/>
        <xs:element ref="FROM_PORT"/>
        <xs:element ref="GPSDATUM"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

```

<xs:element ref="HOWWRECKED"/>
<xs:element ref="INDUSTRY1"/>
<xs:element ref="INDUSTRY2"/>
<xs:element ref="INSPECTED"/>
<xs:element ref="JURISDICTION"/>
<xs:element ref="LATMAX"/>
<xs:element ref="LATMIN"/>
<xs:element ref="LOA"/>
<xs:element ref="LONMAX"/>
<xs:element ref="LONMIN"/>
<xs:element ref="MASTER"/>
<xs:element ref="MAX1"/>
<xs:element ref="MAX2"/>
<xs:element ref="MIN1"/>
<xs:element ref="MIN2"/>
<xs:element ref="MONTHWRECKED"/>
<xs:element ref="OFFNO"/>
<xs:element ref="OWNER"/>
<xs:element ref="PASSDTHS"/>
<xs:element ref="PASSENGERS"/>
<xs:element ref="PORTBULT"/>
<xs:element ref="PORTREGISTER"/>
<xs:element ref="PROTECTED"/>
<xs:element ref="PROTECTIONNOTES"/>
<xs:element ref="REGION"/>
<xs:element ref="REGNO"/>
<xs:element ref="SHIP_TIMESTAMP"/>
<xs:element ref="SIGNAGE"/>
<xs:element ref="SITE_ID"/>
<xs:element ref="SOURCES"/>
<xs:element ref="STATE"/>
<xs:element ref="STATEBUILT"/>
<xs:element ref="STATUS"/>
<xs:element ref="TITLE"/>
<xs:element ref="TONA"/>
<xs:element ref="TONB"/>
<xs:element ref="TO_PORT"/>

```

    <xs:element ref="TYPE"/>
    <xs:element ref="URL"/>
    <xs:element ref="WHERELOST"/>
    <xs:element ref="YEARSSINC"/>
    <xs:element ref="YEARSSINCEWRECKED"/>
    <xs:element ref="YEARWRECKED"/>
  </xs:sequence>
</xs:complexType>
</xs:element>
<xs:element name="BEAM" type="xs:decimal"/>
<xs:element name="BUILDER" type="xs:string"/>
<xs:element name="CARGO" type="xs:string"/>
<xs:element name="COMMENTS" type="xs:string"/>
<xs:element name="CONSTRUCTION" type="xs:integer"/>
<xs:element name="COUNTRYBUILT" type="xs:integer"/>
<xs:element name="CREW" type="xs:integer"/>
<xs:element name="CREWDTHS" type="xs:integer"/>
<xs:element name="DATEBUILT" type="xs:year"/>
<xs:element name="DATEWRECKED" type="xs:date"/>
<xs:element name="DEATHS" type="xs:integer"/>
<xs:element name="DECADEWRECKED" type="xs:year"/>
<xs:element name="DRAFT" type="xs:decimal"/>
<xs:element name="ENGINE" type="xs:string"/>
<xs:element name="F1" type="xs:integer"/>
<xs:element name="FOUND" type="xs:boolean"/>
<xs:element name="FROM_PORT" type="xs:string"/>
<xs:element name="GPSDATUM" type="xs:string"/>
<xs:element name="HOWWRECKED" type="xs:string"/>
<xs:element name="INDUSTRY1" type="xs:integer"/>
<xs:element name="INDUSTRY2" type="xs:integer"/>
<xs:element name="INSPECTED" type="xs:boolean"/>
<xs:element name="JURISDICTION" type="xs:integer"/>
<xs:element name="LATMAX" type="xs:double"/>
<xs:element name="LATMIN" type="xs:double"/>
<xs:element name="LOA" type="xs:decimal"/>
<xs:element name="LONMAX" type="xs:double"/>
<xs:element name="LONMIN" type="xs:double"/>

```

```
<xs:element name="MASTER" type="xs:string"/>
<xs:element name="MAX1" type="xs:double"/>
<xs:element name="MAX2" type="xs:double"/>
<xs:element name="MIN1" type="xs:double"/>
<xs:element name="MIN2" type="xs:double"/>
<xs:element name="MONTHWRECKED" type="xs:string"/>
<xs:element name="OFFNO" type="xs:long"/>
<xs:element name="OWNER" type="xs:string"/>
<xs:element name="PASSDTHS" type="xs:integer"/>
<xs:element name="PASSENGERS" type="xs:integer"/>
<xs:element name="PORTBULT" type="xs:long"/>
<xs:element name="PORTREGISTER" type="xs:string"/>
<xs:element name="PROTECTED" type="xs:integer"/>
<xs:element name="PROTECTIONNOTES" type="xs:string"/>
<xs:element name="REGION" type="xs:string"/>
<xs:element name="REGNO" type="xs:string"/>
<xs:element name="SHIP_TIMESTAMP" type="xs:NMTOKEN"/>
<xs:element name="SIGNAGE" type="xs:string"/>
<xs:element name="SITE_ID" type="xs:integer"/>
<xs:element name="SOURCES" type="xs:string"/>
<xs:element name="STATE" type="xs:string"/>
<xs:element name="STATEBUILT" type="xs:integer"/>
<xs:element name="STATUS" type="xs:integer"/>
<xs:element name="TITLE" type="xs:string"/>
<xs:element name="TONA" type="xs:integer"/>
<xs:element name="TONB" type="xs:integer"/>
<xs:element name="TO_PORT" type="xs:string"/>
<xs:element name="TYPE" type="xs:integer"/>
<xs:element name="URL" type="xs:anyURI"/>
<xs:element name="WHERELOST" type="xs:string"/>
<xs:element name="YEARSSINC" type="xs:string"/>
<xs:element name="YEARSSINCEWRECKED" type="xs:integer"/>
<xs:element name="YEARWRECKED" type="xs:year"/>
</xs:schema>
```

Appendix C – Use cases describing functionality and usability requirements

ID	Use case description	Actor(s)	Stakeholder(s)
1	User is able to deposit a dataset	Data provider	Researcher Data provider
2	User is able to set access rights for data	Data provider	Researcher Data provider
3	User is able to make data discoverable	Data provider	Researcher Data provider
4	User is able to find data	Researcher	Researcher Data provider
5	User is able to view data online	Researcher	Researcher Data provider
6	User is able to download data	Researcher	Researcher Data provider
7	System is easy to use	Researcher Data Provider	Researcher Data provider
8	System allows searches across multiple datasets	Researcher	Researcher
9	System returns search results combined in one table in a web browser	Researcher	Researcher

Appendix D – Informal Survey of Maritime Archaeologists

1. What is your background as an archaeologist, where and when did you train?
2. How many years have you worked in the field?
3. Are you full or part-time?
4. What is your current position?
5. Do you share data with colleagues? How is this done?
6. Do you find any difficulties or constraints on how you can collaborate?
7. Is there anything that you would like to be able to do regarding sharing data that is difficult to do at the present?
8. What do you see as the problems concerning data sharing right now?
9. Are there any difficulties such as intellectual property that would make it difficult to share data?
10. What sort of things would be important to you in a data sharing system?