

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 Information derived from published or unpublished work of others has been acknowledged in the text and a list of references is given.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