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 Information derived from published or unpublished work of others has been acknowledged in the text and a list of references is given.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

List of Figures

List of Tables