Breaking Down National Barriers: ARENA – A Portal to European Heritage Information

Claus Dam[1], Tony Austin[2] and Jonathan Kenny[2]

[1] Kulturarvsstyrelsen, Danmark
(chd@kuas.dk)
[2] Archaeology Data Service, University of York, UK
(afa2@york.ac.uk)
(jk18@york.ac.uk)

Abstract. This paper outlines the recent work of the ARENA (Archaeological Records of Europe: Networked Access) project. In particular it looks at the development of a portal allowing users to search sites and monuments index data from six European countries. The paper looks at the portal architecture and the use of Z39.50 and OAI protocols in tandem, making participation in such portals available to a wide variety of data providers. The paper also considers the search interface and some of the problems and limitations that such a portal encounters, especially in terms of the amount of data potentially available. Finally the paper considers the future for the ARENA partnership as its members consider how to keep a network alive after project funding comes to an end.

1. Introduction

This paper describes a portal that is still under development. It is being developed by the ARENA project; this is a network of six European heritage organisations working together with the support of the European Commission through the Culture 2000 programme. ARENA (Archaeological Records of Europe: Networked Access) (http://ads.ahds.ac.uk/arena/) is committed to the public sharing of archaeological information. Working on preservation of and access to digital archaeological data the ARENA network is developing a portal that will allow the interoperable searching of Sites and Monuments index type data from six different data sets in different countries.

An overview of the project is given followed by a detailed description of the process of creating a portal architecture for a network of partners holding different data bases in different languages. The description of the architecture includes the portal structure, the technologies used, the importance and limitations of thesaurus and term mapping processes and the role of standards in making this possible.

Following a description of the portal architecture the search interface developed to sit on top of the portal is described. Following these descriptive elements the paper concludes by discussing the successes and pitfalls encountered by ARENA in its portal development. In particular confronting the possibility of breaking down national boundaries, considering other projects working on European interoperability and lastly what directions future projects may take.

Fig. 1. The ARENA partners represent 6 countries.

2. The ARENA Network

ARENA is a three-year project that started operation in early 2002. The partnership consists of six organisations from separate European nations:

- The Archaeology Data Service, University of York, UK
- The National Agency for Cultural Heritage, Copenhagen, Denmark
- The Museum Project, University of Oslo, Norway
- CIMeC, Institute for Cultural Memory, Romania
- FSÍ, The Institute of Archaeology, Iceland
- Poznan Archaeological Museum, Poznan, Poland

The ARENA project has four main objectives designed to address digital data preservation and access issues:

- Organisation of Initiatives for Exchange of Experience and the Further Training of Professionals. This has been addressed through the ARENA series of workshops, seminars, web site and publications.
- Promoting Elements of the Archaeological Heritage Concerned. Each partner in ARENA has developed online resources using digital archaeological data from archives of international importance.

- Organising Research Projects, Projects to Raise the Public's Awareness and to Teach and Disseminate Knowledge. Research projects have taken place at each partner allowing the technical development required to create the resources above.
- The Adapted and Innovative use of New Technologies, to the Benefit of Participants, Users and the General Public. This paper outlines the development of the ARENA portal, the product of the fourth ARENA objective.

The development of a portal allowing users to search data sets from the six ARENA partners was the most ambitious of the project objectives. The concept of searching for heritage data across national boundaries was a vital component of a European project.

ARENA sought to address the difficulties inherent in understanding the past whilst bounded by national borders, borders that are too often imposed on interpretations of past "pre national" cultures. ARENA set out to promote a geo-spatial approach to searching for data.

The Archaeology Data Service online catalogue ArchSearch (http://ads.ahds.ac.uk/catalogue/index.cfm) includes geo-spatial attributes within its implementation of the Dublin Core. This allows data to be consulted through a clickable map interface that can eliminate the influence of national boundaries. ARENA aims to develop a clickable map interface through:

- Sharing experience and expertise in the development of map-based interfaces to archaeological records.
- To implement map-based searching at several of the project partners.
- To investigate the implementation of a common map-base for searching at a trans-national scale, including copyright and coordinate system issues.
- To develop a system of interoperable map-based searching which allows users to cross the archives of several partners, with an easy to use and intuitive user-interface, for public use.

Behind the clickable map interface there would be a great deal of technical work to enable the simultaneous searching of partner's data sets. These data sets were in different languages, they had different structures and they had varied fields (such as period or monument type) that have local meaning. The work required to create the portal was both technical, using the International Standards Maintenance Agency Z39.50 and Open Archives Initiative (OAI) protocols and thought provoking, requiring the mapping of local terms and meanings to a generalised ARENA core.

3. The ARENA Technical Architecture

At the time that this paper was delivered ARENA had just achieved cross searching of heritage data held by organisations in separate European nations. The Archaeology Data Service data set could be searched simultaneously with the Norwegian data set held by the Museum Project at Oslo University. This achievement builds on earlier work by the AQUARELLE project (Dawson 1997) and HEIPORT (Austin et al 2002).

The AQUARELLE project linked a number of research organisations, commercial companies and cultural organisations, including MDA, RCHME, and the Culture Ministries of France, Italy and Greece, using the Z39.50 protocol. It is not however, available on line any longer. HEIRPORT (http://ads.ahds.ac.uk/heirport/) linked organisations within the United Kingdom (RCAHMS, The Computing Laboratory: University of Kent, EDINA, English Heritage, Portable Antiquities Scheme, SCRAN, University of Oxford and the Archaeology Data Service at the University of York). The HEIRPORT project utilises Z39.50 in conjunction with Zava software devised at the University of Kent. The ARENA portal set up at the ADS utilises the Z39.50 protocol and Zava to submit a search to the ARENA data sets. The ADS and Museum project archives are the first two data sets to become part of the portal; four more will be added over the summer of 2004 (Iceland, Denmark, Poland and Romania). The Z39.50 protocol sits at the heart of the portal, this in turn relies on the Dublin Core (DC) metadata standard to recognise the various fields in client databases.

Z39.50 was originally developed within the libraries community to facilitate the simultaneous searching of geographically distributed library cataloguing systems. It is a protocol or set of rules that govern the discovery and retrieval of information within such a system. Z39.50 follows a client/server model and provides a number of facilities.

Z39.50 is made up of a number of structural blocks called facilities and services therein. These facilitate operations between client and server. The most important of these are Init (initialise a session), Search, and Present (results).

Users interface with the client or origin usually using a Web browser. Web based clients have variously been described as Gateways, Portals or Portlets. They allow users to formulate queries, which are then sent simultaneously to any number of target servers.

A target server receives user queries aimed at an underlying database. As there can be any number of target servers with differing database systems and data structures the user query is broadcast in a generic form, which must be decoded, by the target server. Similarly results must be presented uniformly in order to be meaningful across targets. The adoption of shared standards achieves this interoperability. The main standards take the form of attribute sets and record syntaxes.

Z39.50 uses "attribute sets"; essentially these are values or numbers that set various properties pertaining to a database query. For example, a 'use attribute' value of 2047 from the most commonly used attribute set, BIB-1 (http://www.bibliotech.com/html/z39_50_bib-1.html), indicates a query on Dublin Core subject terms. Other attribute types are used by the target include structure (search for word or phrase, etc), position (where the search should appear within a database field (beginning or anywhere for example).

"Record syntaxes" are also used to define how information is returned to a user following a successful search with the consequence that local data structures need to be mapped to shared syntaxes. Showing its origins, Z39.50 systems have tended to use library-cataloguing syntaxes such as SUTRS (Simple Unstructured Text Record Syntax) or MARC (MAchine-Readable Cataloguing). Recently there have been moves towards XML encoded Dublin Core as, for example,

specified by the Bath profile. The ARENA portal has utilised XML as its medium of choice for transporting data within the architecture. Z39.50 uses profiles to group attribute set and record syntax definitions. Profiles can become international profiles in their own right, for example, the Bath profile (http://www.ukoln.ac.uk/interop-focus/bath/) and CIMI (Consortium for the Interchange of Museum Information) profiles which specify which suite of Attributes, Record Syntaxes, and other factors to use.

An important consequence of the use of Bath Profile and CIMI is that datasets must be mapped to shared standards including the Dublin Core metadata schema and the CIMI schema for the 4 W's (Where, When, What and Who) and for spatial referencing systems which creates a framework for the semantic searching of these elements in any combination. The Bath profile also requires the use of XML for record delivery and the possibility of exporting data so tagged to other applications.

There are problems associated with Dublin Core, different communities and even individuals within communities may interpret the semantics of specific elements differently. The key task for the ARENA partners has been to use the same interpretation making interoperability sustainable. Other problems concern granularity in that information may be held at different levels, for example, record and collection level. For the ADS it has become apparent that some DC elements are more easily dealt with at the record level and others at collection level. This has been a problem for some of the ARENA partners, perhaps unsurprisingly given the diversity of the data sets. The ADS interpretation and mapping of its data to DC has been driven by a number of influences; notably *Discovering Online Resources Across the Humanities: A Practical Implementation of the Dublin Core* (Miller and Greenstein 1997), the *Dublin Core Metadata Initiative (DCMI) definitions of elements* (see DCMES, 2003) and more recently the UK *e-Government Metadata Standard* (2003).

The ARENA portal set out to investigate the possibility of making other technologies available to search the partner's data sets. To do this the ARENA partnership decided to adopt the OAI protocol into the portal architecture.

Originally developed to facilitate information sharing within the e-prints community. OAI is a protocol or set of rules that define the structure for querying remote hosts. These queries piggyback on HTTP which itself is the protocol governing the exchange of files (text, images, etc) on the World Wide Web. OAI uses HTTP requests; get and post. Specifically it defines a fixed query to recover OAI encoded metadata records to a schema, default Dublin Core or as otherwise defined.

As a component of the ARENA portal OAI is used to harvest records from the partners (data providers) who chose to use this approach. A record is an XML-encoded byte stream that is returned by a data provider in response to an OAI protocol request for metadata. OAI mandate the use of the Dublin Core metadata set. Records have unique identifiers and are date stamped.

The OAI Data Providers amongst the ARENA partners maintain repositories on a network accessible server to which OAI protocol requests, embedded in HTTP, can be submitted. The server needs to be able to return XML encoded Dublin

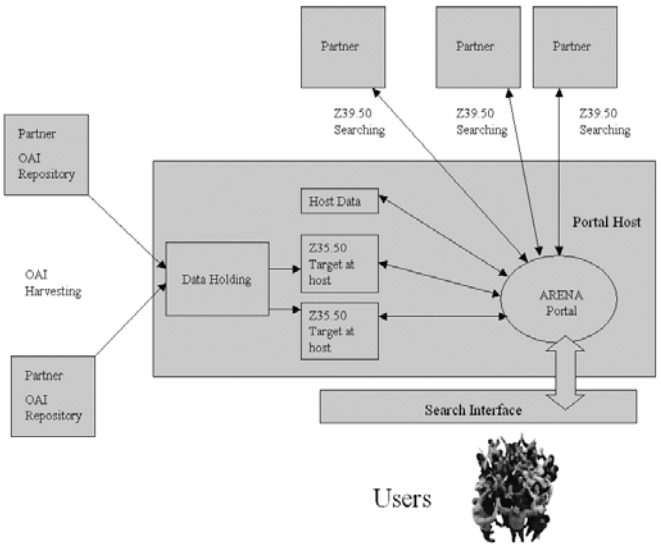

Fig. 2. The ARENA portal architecture.

Core metadata in response to the request., that either meet a default unqualified Dublin Core schema or an agreed and defined schema.

The ADS as Portal host acts as a harvester, a remote data collector that is used periodically to collect data. This creates a holding repository or set of repositories, depending on the number of partners using OAI. A record date stamp allows selective collection of new records added or records updated since the last harvest.

The ADS will harvest records from those partners who elect to use OAI and make their records available in the holding repositories. OAI does not prescribe how this is done. In the ARENA portal the holding repositories are held up as local Z39.50 targets that are then queried by the portal.

The strength of the ARENA architecture for European network building is its flexibility. Using Z39.50 or OAI to make data available for searching requires different technical skills. The skills base across European heritage organisations is varied and the choice of protocol brings flexibility.

On top of the architecture for interoperability a multilingual search interface has been added.

4. The ARENA Portal Search Interface

The portal interface structure is shown in Fig 3. The user enters the portal by clicking on a flag representing the six native languages of the ARENA partners. This then takes the user into the ARENA search interface using the selected language. Setting up the interface in this manner is time consuming but allows the initial search to be set up in a familiar language.

The user is faced with three options to refine their query of the six ARENA data sets. These are the tried and tested *When*, *What* and *Where* options first piloted by the AQUARELLE project. A search of the six data sets that is not refined in any way is likely to generate such a torrent of hits that the user will find the data unusable, simply in some cases attempting to down load the entire national sites and monuments record. The easiest way to narrow down a search is through the *Where*

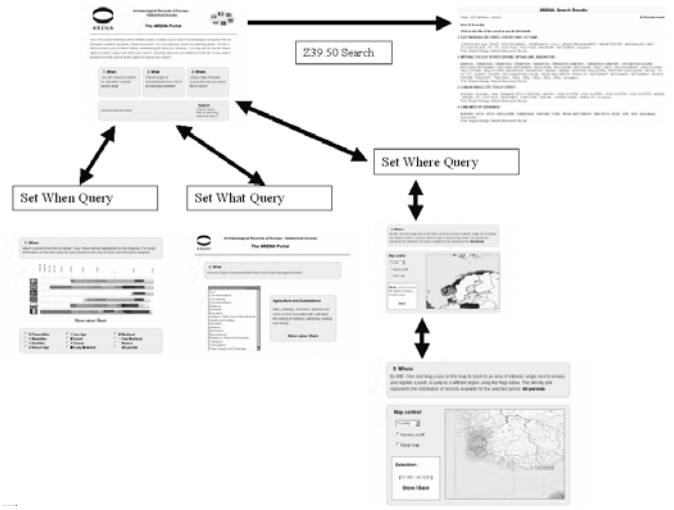

Fig. 3. The ARENA search interface structure.

option, but it is also possible to search by period and if required theme (*What*).

The *When* option is selected from a basic set of eleven top-level period terms. These basic period terms were selected to allow partners to map to their own period terms into the top level. Thus for example any site record from the Epoka Zelaza in Poland maps onto the Epoka Brazu or Bronze Age in English. Once a preferred period option is selected it is stored and the user returns to the main search page.

The *What* option utilises the definition of sites and monuments by use or type. This is achieved at a very top level again, mapping monument types into the set of themes used by the *Thesaurus of Monument Types* devised in the UK by English Heritage (1998). This is clearly a very basic mapping but in itself is the beginning of a multilingual thesaurus and ontology building process that will require considerably more time and resources than a three year project can offer. The principle is demonstrated however, that a variety of thesauri can map onto basic themes at the top level.

The *Where* option is the key to working with manageable numbers of returns from such a large combined data set. A combination of *When* and *What* will often generate tens of thousands of hits but when concentrated on a specific place the search generates a manageable set of results. Fig 3 illustrates the structure of the basic search interface through which the user selects and stores search criteria.

The ARENA partnership set out to demonstrate the potential to search across multiple data sets. Yet by searching using the *Where* option, that searches for sites within a predefined radius around the point selected, most searches will only hit one partner data set. The search interface also includes an advanced option that allows the user to select *When*, *What* and then any of the partner data sets. As a result of the large data sets involved for some partners this will return many hits (in excess of 10,000), but for specific combinations it will be possible to carry out a search across all partners. Allowing more detailed searches is really the only solution to the problem of returning too many hits. To do this an extensive multi-lingual thesaurus is required for sites and monuments across the whole of Europe, or of course map defined area searches. The ARENA project has shown that it is possible to map to a top level set of terms but the work required to take European cross searching forward is extensive. The technology is there but a great deal of "person time" is required to negotiate such a thesaurus. This has been managed for specific terminologies, especially those for heritage management terms created by the HEREIN project, but there is a great deal more yet to be done.

Even after a multi lingual thesaurus of sites and monument terms is compiled the results from any cross searching will still be returned in local languages. This is a lesson that anyone researching data from multi national resources will have to learn. Ultimately researchers will have to learn some basic terminology in the language of the country in which the data was created.

5. The Future for ARENA

The ARENA portal is timetabled for presentation at the European Association of Archaeologists conference at Lyon in September 2004. One of the strengths of the ARENA project has been the multinational partnership that has been created. Many of the partners are involved in other projects, particularly those created by the new sixth framework funding programmes. The partners have all agreed to keep their servers and targets open for searching through the portal for a fixed term. The ADS will maintain the portal for the same period. Holding annual ARENA meeting, probably at CAA, will also keep up the maintenance of the portal and other potential projects.

Lastly, the ARENA partnership and network has demonstrated the vulnerability and potential of digital archaeological archives to a European audience. It has also demonstrated that the technology is already available to share archaeological data across national boundaries. These demonstrations naturally show potential paths to make the most of the potential of the digital archaeological record of Europe. An extended ARENA project would take the work forward by extending the thesaurus, improving the map interface by expanding the use of GIS and building in an absolute date to the "When" search option. Just as important of course will be the addition of new partners to the portal.

References

Austin, T., Pinto, F., Richards, J. and Ryan, N., 2002. Joined up writing: an Internet portal for research into the Historic Environment, in G Burenhult (ed) *Archaeological Informatics: Pushing the Envelope Computer Applications and Quantitative Methods in Archaeology: CAA 2001* Oxford, Archeaopress, BAR International Series 1016, 243–252.

Dawson, D., 1997. A use for SPECTRUM: AQUARELLE. mda Information Vol 3 No 1, *Papers from the Standards in Action* Workshop Churchill College, Cambridge, 1–3 October 1997.
http://www.mda.org.uk/info31aq.htm
(downloaded 22.07.2004)

DCMES. 2003. *Dublin Core Metadata Element Schema*, version 1.1: Reference Description, DCMI. http://dublincore.org/documents/2003/02/04/dces/ (downloaded 22.07.2004)

Kenny, J., 2002. Enter the ARENA... *ADS Archaeology Data Service NEW*S issue 12 autumn 2002. http://ads.ahds.ac.uk/newsletter/issue12/arena.html (downloaded 22.07.2004).

Kenny, J., 2003. European Archaeology Archives: ARENA launches major new data sets. *ADS Archaeology Data Service NEWS* issue 13 spring / summer 2003. http://ads.ahds.ac.uk/newsletter/issue13/arena.html (downloaded 22.07.2004).

Kenny, J., 2003. Networks of Excellence and the Excellence of Networks. ADS Archaeology Data Service NEWS issue 14 autumn 2003. http://ads.ahds.ac.uk/newsletter/issue14/networks.html (downloaded 22.07.2004).

Kenny, J. and Austin, T., 2004. Data preservation: Exploring the 'rescue' role of the Archaeology Data Service. *Content Management Focus* vol 3 issue 5, 25–29.

Kenny, J. and Kilbride, W. G., 2002. Networked Access to Digital Archaeological Archives in the European Arena *CSA Newsletter* Vol. XV, No. 2 – Fall, 2002. Online at: http://csanet.org/newsletter/fall02/nlf0202.html (downloaded 22.07.2004).

Kenny, J. and Kilbride, W. G., 2003. Europe's Digital Inheritance: ARENA archives launched. *CSA Newsletter* Vol. XV1, No. 1 – Spring, 2003. http://csanet.org/newsletter/spring03/nls0302.html (downloaded 22.07.2004).

Kenny, J. and Kilbride, W. G., 2004. Europe's electronic inheritance: The ARENA project and digital preservation in European archaeology. In Ausserer, K. F., Borner, W., Goriany, M. and Karlhuber-Vockl L. (eds), *Enter the Past: The E-way into the Four Dimensions of Cultural Heritage, Computer Applications and Quantitative Methods in Archaeology 2003*. Oxford, Archaeopress BAR International Series 1227, 130–133.

Kenny, J., Kilbride, W. G., Aldred, O. and Dam, C. H., 2004. Deploying digital data: Making the most of digital archives for archaeology. *The European Archaeologist* 20, Winter 2003/2004, 16–19.

Kenny, J., Kilbride, W. G., and Richards, J. D., 2003. Enter the ARENA: preservation and access for Europe's archaeological archives. In Doerr, M. and Sarris, A. (eds), *The Digital Heritage of Archaeology Computer Applications and Quantitative Methods in Archaeology 2002*, Archive of Monuments and Publications, Hellenic Ministry of Culture. 349–353.

Miller, P. and Greenstein, D. (eds), 1997. *Discovering Online Resources Across the Humanities: A Practical Implementation of the Dublin Core*, UKOLN.

Office of the e-Envoy. 2003. *e-Government Metadata Standard* version 2 (draft). http://www.govtalk.gov.uk/documents/metadataV2.pdf (downloaded 22.07.2004)

UKOLN. *Collections Description Focus*, http://www.ukoln.ac.uk/cd-focus/ (downloaded 22.07.2004)