Cross-Border Data Sharing: A Case Study in Interoperability and Web Services

Henriette Roued Olsen[1] – Stuart Eve[2]

[1]University of Oxford, UK
[2]L – P : Archaeology, UK
[1]henriette@roued.com
[2]stuarteve@lparchaeology.com

Abstract

Releasing heritage information online is to be encouraged but it is important that any project contemplating this should also look into the issues of access to their data. There are vast amounts of heritage data available online in a format which is of little use for researchers who would like to reuse or interrogate the datasets. XML could provide a solution to this and combined with Web Services it will allow users to create their own aggregated searches across a number of online datasets. This paper discusses an attempt to develop a deep portal which searches the Swedish Sites and Monuments Record (FMIS) and the ARK system developed by L – P : Archaeology through Web Services and creates a combined output.

Keywords

heritage portal, interoperability, cross-border, data sharing, accessibility, open data, deep portals, XML, MIDAS, ARK, FMIS.

1. Introduction

This paper is based on the MSc dissertation: 'Heritage Portals and Cross-border Data Interoperability' (Olsen 2007b). The primary aim of this dissertation was to enable collective searches across heritage related datasets from different European countries through the use of a heritage portal. The dissertation produced a deep portal, which is defined as a portal bringing the contents of other web applications together through connection technologies and often facilitating cross-search functionality (Clifford 2007).

The basis of this idea is that across Europe there is a great amount of heritage data stored which could be used for research and teaching if it was more accessible. In 1992 Henrik Jarl Hansen presented a utopia suggesting a joint search of European databases of Sites and Monument Records (SMRs) (Hansen 1992, 236). He illustrated this with a map of Europe with octopi each symbolising SMRs with interconnected arms (*Fig. 1*).

The main argument for sharing heritage data is that prehistoric and historic cultures did not share borders with contemporary Europe (Dam and Hansen 2005). Therefore, any research into these cultures should cross modern borders in the same way. Furthermore, due to an increasing amount of material available through excavations brought on by modern development there is an immense amount of new data which could and should be used in modern interpretations of the history and pre-history of Europe. These data might not be accessible for researchers who are, in many cases, forced to base their interpretations on outdated secondary literature. Finally, a general openness to primary material allows new generations to reinterpret older and present hypotheses, which should eventually lead to a better understanding of the past. In conclusion, there is much to be gained from researchers gaining access to collective searches across European datasets through heritage portals.

2. The Project

The project was soon to realise that a number of heritage portals had already been produced creating perfectly good cross-searches on heritage datasets. One example is the ARENA (Archaeological Records of Europe: Networked Access) project which combines services from Poland, Romania, Denmark, Iceland, Norway and the United Kingdom as Dublin Core (DC) standard records based on the Z39.50 protocol (Kenny and Richards 2005, 1). This project allows for a cross-search based on WHERE, WHAT and WHEN parameters outputting the data as HTML. Another example is the newly developed Heritage Gateway, which uses SOAP Web Services outputting

Fig. 1. Hansen's Octopi (Hansen 1992).

XML which is then transformed into a HTML style fitting to the Heritage Gateway web site. This portal combines regional SMR data across England and other online national heritage databases (Colman and Bowen 2007).

However, most portals at this point ensure that the Web Services they combine are outputting data according to the same standard (Waller 2005, 5) and the end user is seldom presented with other output opportunities than HTML. So far we have only come across one project within humanities, which has made their data available through a multitude of methods such as Z39.50, OAI-PMH, SRW and RSS (Intute 2007). The Intute project, a shallow portal providing access to resources for education and research within humanities, is very eager to share their search functionality and hopefully more projects will follow in their footsteps.

In consequence, this meant that the aims of the current project were revised. First of all it would now attempt to prove that it is possible to successfully create a deep portal based on datasets, which do not use any related standard. Secondly, the dissertation sought to output the result of the portal search in a choice of machine readable XML or human readable HTML.

3. Issues

Allowing public and easy access to Sites and Monument Records (SMR) is becoming increasingly popular and while some agencies might feel the need to protect their data (Fernie 2003, 6) others happily share it motivated by a policy which states that *"if you don't know the precise location of a site you can't raid it, but if you don't know where it is you also cannot help protect it"* (Dam and Hansen 2005, Chapter 3.1).

At this point any institution wanting to make data available online should decide what their ideal user requirements are. If they want their users to be able to go to a certain website, search for certain data and view the output on the website in plain text then there is no need to go any further than a web site with search functionality. However, if they want their users to be able to re-use the data they have found they should make it available in a machine-readable format such as XML. Finally, if they want their users to re-use the dataset dynamically and create their own search functionalities in other applications they should consider making it available through Web Service techniques. However, with this final option comes other issues such as server overload and maintenance.

Creating a combined search of two or more SMRs is a great idea. Nevertheless, one of the biggest problems when trying to do this is that the datasets are seldom built in the same way even within the same country (Lock 2003, 198). Apart from the obvious language issues there is also a big difference in what different institutions choose to record. While most SMRs record WHAT, WHEN and WHERE for each record they do not all allow searches through these parameters in their online versions. Furthermore, there might be big differences in how they record for example the WHAT parameters. One institution might have a tradition of calling something an axe, which another institution has always called a hammer.

Although some archaeologists might dream of a world where everybody adheres to one standard for recording this is extremely unlikely, if not impossible. Most European countries have their own standards and the institutions within these might even have different standards too. Therefore, expecting everybody to adhere to one standard is unrealistic. One solution to this could be to map the concepts (i.e. a placename or a parish) to a specific element of an ontology (e.g. CIDOC Conceptual Reference Model (CRM) (Croft *et al.* 2007). But the real issue here is who should do this and whether it can be done on different levels. In the dissertation a very simple example has been used which maps one concept of a parish to another concept of a placename (*Fig. 2*). This is an ad hoc mapping which is not obvious in every context but is useful for the sake of this project.

The point is that an institution, which creates a dataset, will have a good idea of what is in this dataset and the standard they are using makes sense in the context of their dataset. If they decide to produce a Web Service to allow machine readable and searchable access to their dataset it would be perfectly fine to use their existing standard for this output. If they then wish to provide users with the same material in another standard they can transform the output on the fly to XML formatted in this standard and thus provide the user with two different outputs. Furthermore, this also allows the user to create their own remapping for their own specific use.

Fig. 2. Mapping of two parameters to one CIDOC CRM element.

Finally, if the institution has mapped their dataset to an ontology (e.g. CIDOC CRM) then in theory it should be easier to combine their dataset with another dataset mapped to the same ontology. The issue here is that the process of mapping to an ontology has to be undertaken manually and is very subjective and thus only provides a step of the way towards making two datasets comparable.

4. Methodology

A Web Service defined as '*a software system designed to support interoperable machine-to-machine interaction over a network*' (Graham *et al.* 2005, 11) was created for data presented through the ARK system.

The ARK system (Archaeological Recording Kit) developed by Stuart Eve and Guy Hunt from L – P : Archaeology is versatile in the sense that it is not built of contexts, records and descriptions but instead modules, items and fields. This means that the ARK system can be used to record anything. The project must simply decide what they would like to record (e.g. archaeological contexts, files, books) which will be modules and then for each module they can record information about each item, for instance the soil colour or the author of the publication. For the dissertation a dataset from Olsen's bachelor project 'Reflections on culture connections – Examining connections between South Scandinavia and the Sîntana de Mures/Çernjachov culture from AD 270–410 (Periods C_2 to D_1)' was used. The dataset contained information about facetted glass, bowls and glass production sites across Europe (Olsen 2007a). The dataset was set up as an ARK project with two modules (i.e. Sintana for the site points and Bibliography for the references) and several fields in each module (*Fig. 3*).

The ARK Web Service used the SOA Protocol to send messages through HTTP to a server which then returned XML based on the query received. The SOA Protocol was chosen above the newer and simpler REST (Representational State Transfer) because it allows the implementation of WSDL (Web Services Description Language) and UDDI (Universal Description, Discovery and Integration) (Freitag 2005). WSDL is basically a XML schema which allows a Web Service to be described and furthermore allows the user easy access to the service based on information like what the service does,

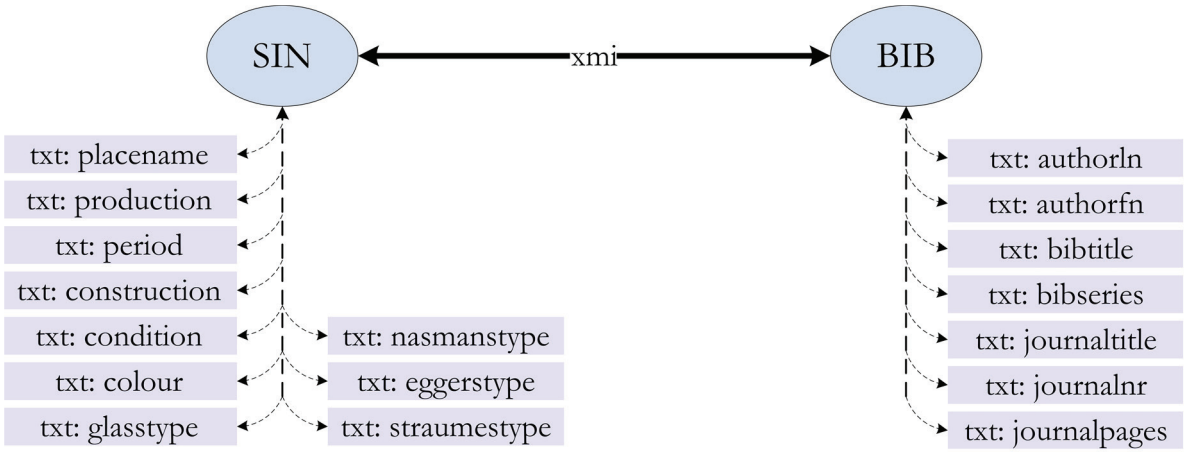

Fig. 3. Model of the ARK setup for the Sintana project.

how it is accessed and where it is located (Graham *et al.* 2005, 173). The UDDI is an initiative, which gives users access to a directory of references to Web Services (WSDL) (Graham *et al.* 2005, 311). Ideally all heritage Web Services should be collected under a heritage UDDI giving users one access point to all.

FMIS (Fornminnesinformations-systemet), the Swedish SMR is maintained by the Swedish Cultural Heritage Agency RAÄ (Riksantikvarieämbetet) who provide a simple Web Service which gives access to metadata based on a WHERE search. The project created a script, which searched for certain parishes and retrieved a dataset for each.

The last step of the project was to create the collective search of the two datasets and output a combined XML file on-the-fly. This was done by writing a script which would search through the two Web Services for the pre-determined WHERE parameters. It would then transform the two output datasets into XML formatted in another standard. This could have been any one of the two original standards or a new standard made for this project. However, it was decided to use the already well respected standard for Sites and Monument datasets MIDAS. After the two outputs were transformed they were then combined to one XML file and output as the search result (*Fig. 4*).

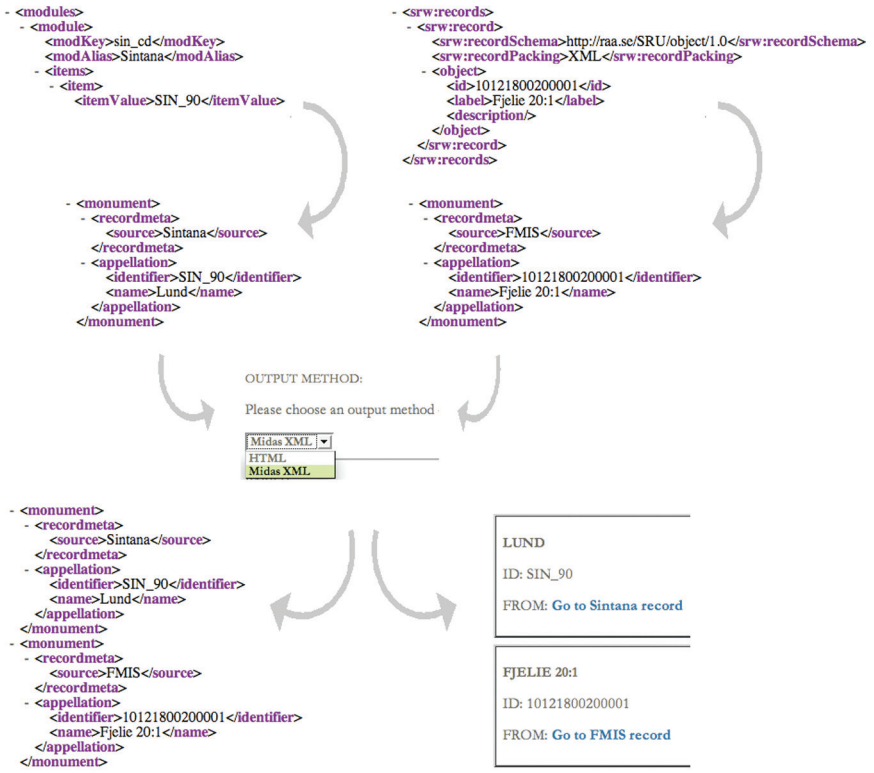

Fig. 4. Model of the process of creating a heritage portal.

5. Conclusion

The first aim of the project was to create a deep portal which cross-searched different datasets and output a combined result set. This has been created successfully as a part of the dissertation. However, during the process of the dissertation it was discovered that other larger projects had already created similar portals and therefore the project began to look into what they were missing. This evolved into two new aims, which have been discussed throughout this paper. The first was based on an observation of a tendency to disregard datasets for the portals if they could not be presented in a uniform standard. The dissertation proved that this was not necessary.

Secondly, the project observed that many search results are presented as HTML even though some of them are built through XML. If a project wishes to create interoperability between their work and others they should provide the end user with the chance of re-using the data as dynamic XML. The dissertation presented this option by simply letting the portal service specify the output as a parameter in the search URL.

In conclusion the dissertation reached its goals and the ARK system now has a Web Service, which can be used on any ARK project. In the future it will be possible to use this ARK Web Service to allow the ARK projects to interact with other projects and might also provide a solution for web publishing of datasets in ARK. The next step will be to implement these ideas in present projects together with L – P : Archaeology.

Finally, it seems that mapping datasets to an ontology like the CIDOC CRM is the future for heritage interoperability. An ontology will not solve all problems as there will still be differences in recording systems that can not be bridged by any mapping system. It is important to remember that if this mapping is done on-the-fly, the whole dataset need not be mapped, just enough fields to allow

the datasets to interact with each other to allow the discovery of relevant records which can then be explored in more detail if necessary. Therefore, the difference in recording systems might not prove a problem as long as it is possible to find some meaningful parameters, which can be cross-searched to return the datasets.

Note

The results of the dissertation can be viewed and experienced on the dissertation web site (http://www.roued.com/diss), which also contains information about Web Services. The ARK web site (http://ark.lparchaeology.com) contains information about the ARK system and it is possible to track the process of the development here. The ARK system will soon be distributed as open source code.

Bibliography

Clifford, Lisa (2007). *Portals – Frequently Asked Questions*. JISC. (http://www.jisc.ac.uk/whatwedo/programmes/programme_portals/ie_portalsfaq.aspx)

Colman, M. and R. Bowen (2007.) *Heritage Gateway. Phase 2: Technical Specification v1.0 (Draft)*. Unpublished document by English Heritage.

Croft, Nick, Martin Doerr, Tony Gill, Stephen Stead and Matthew Stiff (2007). *Defiition of the CIDOC Conceptual Reference Model*. Version 4.2.2. (http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.2.pdf)

Dam, Claus and Henrik Jarl Hansen (2005). The European Digital Resource in Archaeology: Sites and Monuments Data as a common European Web Resource. *Internet Archaeology* 18. (http://intarch.ac.uk/journal/issue18/4/toc.html)

Fernie, Kate (2003). Getting it together on-line: HEIRNET and Internet-based resource discovery tools for the Historic Environment. *Internet Archaeology* 13. (http://intarch.ac.uk/journal/issue13/4/toc.html)

Freitag, Pete (2005). *Rest vs SOAP Web Service*. Posted on 03.08.2005. (http://www.petefreitag.com/item/431.cfm)

Graham, Steve, Doug Davis, Simeon Simeonov, Glen Daniels, Peter Brittenham, Yuichi Nakamura, Paul Fremantle, Dieter König and Claudia Zentner (2005). *Building Web Services with Java. Making sense of XML, SOAP, WSDL and UDDI*. (2nd ed.) Indiana: Sams Publishing.

Hansen, Henrik Jarl (1992). European archaeological databases: problems and prospects. In: Andresen, J., Madsen, T. and Scollar, I. (eds.) *Computing the Past. Computer Applications and Quantitative Methods in Archaeology*. 229–237. Aarhus.

Intute (2007). *Integrating Intute into your site or service*. (http://www.intute.ac.uk/integration/)

Kenny, Jonathan and Julian D. Richards (2005). Pathways to a Shared European Information Infrastructure for Cultural Heritage. *Internet Archaeology* 18. (http://intarch.ac.uk/journal/issue18/6/toc.html)

Lock, Gary R. (2003). *Using Computers in Archaeology: Towards virtual pasts*. London: Routledge.

Olsen, Henriette Roued (2007a). Reflections on Cultural Connections – Examining connections between South Scandinavia and the Sintana de Mures / Çernjachov culture from AD270–410 (Period C2 to D1). *LAG 8*. Hikuin: Højbjerg.

Olsen, Henriette Roued (2007b). *Heritage Portals and cross-border data interoperability*. MSc Dissertation University of Southampton.

Waller, Stewart (2005). Future Connections: The potential of Web service and Portal technologies for the historic environment. *Internet Archaeology* 18. (http://intarch.ac.uk/journal/issue18/8/toc.html)