*Stuart Jeffrey – William Kilbride – Stewart Waller – Julian Richards*

# Thinking Outside the Search Box: The Common Information Environment and Archaeobrowser

*Abstract:* The paper describes how the Archaeology Data Service, in collaboration with Adiuri Systems, has tackled some of the problems of searching complex heritage datasets using a standard search box by creating a browser-based demonstrator. The traditional search box approach requires significant knowledge of the datasets to be interrogated and often results in low user confidence. The Archaeobrowser makes use of facetted classification technology to guide the user through over 1,000,000 records from multiple datasets. Each element of the datasets was indexed into hierarchical structures based on the concepts of 'what', 'where' and 'when'. The browser also incorporated geospatial searching and user-friendly 'trails'.

## *Introduction*

Whether we query a library catalogue or search a computerised monument inventory, the majority of archaeological research is now carried out on computer. How though do we find what we are looking for, without the aid of a librarian or information scientist looking over our shoulder? Once upon a time, someone who understood the vagaries of their catalogue or database would be there to help us tweak our enquiries to locate what we were after. However, as more and more archaeological resources can be searched online, an increasing proportion of such queries are now conducted remotely, without the benefit of expert human assistance. Invariably we are presented with an empty search box and invited to second guess the right combination of words to allow the database to do its magic. Even if the database has been constructed using a carefully controlled vocabulary, this can be problematic if the user does not understand how the vocabulary is structured, or if there are internal hierarchies of related terms. If the search combines records drawn from several resources, each with their own vocabularies, then it becomes almost impossible to be confident that all the relevant items have been found. In the context of archaeological research, that can be damaging to say the least; in a development control context, it could be disastrous.

The Archaeology Data Service (ADS) catalogue, ArchSearch, now contains over one million monument records, drawn from a variety of sources. Finding precisely those sites one is looking for demands a higher level of search skills than many users possess. The ADS has therefore worked with a number of partners to develop a "geospatial demonstrator". Rather than being faced with an empty search box, users are invited to navigate their own pathway through a million records which have been pre-indexed according to the three key variables of When, What and Where. The aim was to enable users to "Point and Click" not "Type and Hope". The Common Information Environment Demonstrator project (hereafter the CIE Demonstrator) was a collaborative project between the ADS and Adiuri Systems (http://www.adiuri.com/), building on two earlier demonstrator projects and using Adiuri's Waypoint classification software. The project was initiated in 2004 by the CIE working party through Dr Paul Miller, CIE Director. The demonstrator has been online since late December 2006 (http://ads.ahds.ac.uk/project/cie).

The vision for the project was to build a robust, engaging and technically sophisticated online demonstrator that permitted browsing and searching with cultural heritage data. The demonstrator was to combine cross-searching of discrete, remote data sets using a new generation of facetted classification tools. Facetted classification has not been used much in the heritage sector but provides a new way of interrogating data sets which can be indexed according to multiple attributes, or "facets" (Ross / Janevski / Stoyanovich 2007). The data sets in question were cognate though not explicitly linked and therefore the demonstrator enabled users to find and make intellectual links between the data sets. At the core of this linking was a move from cross-searching to cross-browsing. Geography was identified as a

readily understood way to cross-reference data, though it was also recognised that this introduced a need for software development and data cleaning, the latter coming through use of the geoXwalk gazetteer (http://www.geoxwalk.ac.uk/). In addition, the project explored issues of personalisation, ensuring that rich content could surface and also ensuring the accessibility of content by providing a series of trails (predefined sets of results on specific topics for pedagogic purposes).

## *Archaeobrowser Development*

The development of the Archaeobrowser centred on three main strands: software development; data acquisition and preparation; and classification. The project team was able to develop approaches to each of these issues which contributed to a greater understanding of facetted classification with geospatial data.

### Software Development

A range of software developments were required by the project. The web interface for Waypoint was enhanced by three important additions. A simple keyword-based search tool was made available which in turn was used to drive three elements of the demonstrator. The keyword search function was wrapped within an HTML code fragment to allow embedding on different sites – and was ultimately used internally within the demonstrator. Search terms were saved within user-designed profiles that presented a series of seeded searches within the classification. These keyword-based profiles searches were then used to complement the trails that highlighted rich or connected content within the demonstrator. Finally the simple keyword search box was added to the tools already available, and in this way made the interface more immediately approachable to users. These interface enhancements were over and above the redesign of Waypoint required by the addition of a map interface.

Based on a three-stage process of data acquisition, normalisation and transformation into XML, ADS developed a data preparation module for Waypoint that greatly enhanced the preparation of the index file produced by the facetted classification. Moreover, the data preparation module includes a series of tools for cleaning and enhancement of the data, such as lookup tables to a known fixed data point and a new client interface to the geoXwalk gazetteer. This latter function was specifically designed such that it would only be called on when local sources were insufficient.

### Data Acquisition and Preparation

A toolkit was developed in order that data could be presented to the classification engine. These java-based tools were written by ADS and took each data set through from the point of data gathering to the point that they were ready to be processed. Data was gathered in different ways from a number of sources, such as English Heritage, the British Library, The National Archives and the Council for British Archaeology. A number of different gathering techniques were also employed from manual transfer to HTTP transfer to OAI harvesting. The demonstrator also benefited from the inclusion of the ADS's ArchSearch catalogue, which was also made available as an internal disk transfer.

Data processing began by a mapping process in which the relevant fields in the data set were assigned to different points in the classification. Although most of the data sets were compliant with some form of data standard each had implemented these standards differently. Once completed, the data sets were exported out of their native formats into a common XML format that included simple Where, What and When fields, with limited descriptive text and a further flag for rich media. Once the fields had been rendered, it became easier to clean the data against a control set to ensure consistent vocabulary. This was acceptable in respect of the What fields but more successful with the When and Media Type fields.

This data preparation was time consuming but worthwhile. It was originally undertaken because early builds of the demonstrator presented an excess of 'false positives'. The classification engine had no inherent ability to distinguish an address in "Castle Street" from a "Castle" or a reference to the "Duke of Buckingham" to the place "Buckingham". By associating fields with specific types of information, it became possible to reduce (though not eliminate) these false positives, and also forced a manual quality assurance on the data.

## Classification

The first step in the development of a facetted classification is the development of a knowledge map against which the data sets are to be classified. In practice, this iterative process happened throughout the project as refinements were suggested and tried.

Early discussions focussed on what sorts of elements the classification ought to contain. Five basic units of information were considered – who, what, where, when and how. The foundation of the demonstrator was to cross-reference spatial and temporal information and so the Where and When elements were retained. Reviewing the data sets available it was relatively simple to identify two different types of What question – one relating to the nature of the heritage object being described, another relating to the media type that represented the heritage object.

Who and How were more difficult to pin down. In relation to How, it was noted that it may be possible to list the methods by which a heritage object had been assessed, and that this could be presented in a hierarchical manner. But it was also noted that this approach was likely to be arcane and unlikely to be of much interest to the target audience. In relation to Who, it was noted that the audience would very likely be interested in this topic but that it was difficult to envisage a useful hierarchical arrangement. In effect, the demonstrator would simply list all historical personages, adding little to the user experience. The same information would in any case be available through a keyword search of the data. Thus, the team decided to develop a concept map based on What, When and Where, and did not progress Who and How classifications.

### *The "What" Hierarchy*

The target audience for the demonstrator was for a group of interested amateur archaeologists who might want to draw together data from a variety of sources, but who might not know or understand the technical vocabulary of the professional archaeology sector. At the outset, the team recognised the opportunity inherent in the English Heritage Thesaurus of Monument Types and gained permission to use this within the project. The thesaurus is comprehensive and poly-hierarchical, providing scope notes and preferred terms across several thousand terms, starting with 18 high level concepts. However the thesaurus was never intended for the lay user and at times the hierarchies seem counterintuitive; it was never intended as a browsing structure. For example, the concept of "cathedral" appears at a level below the concept "piscina", even though a piscina is an architectural detail found within cathedrals and other churches. Therefore, the project team set itself the task of recasting the thesaurus for lay users, so as to provide a more accessible browsing structure. Although care was taken to ensure that the resulting browsing structure was logical, additional user-focused rules were also introduced. For example, it was decided that no hierarchy should have more than 12 elements, on the premise that the hierarchy would have to be displayed to users and any more than 12 would introduce extra scrolling. All the terms within the thesaurus were re-assigned in this way, including those terms that were deprecated, ensuring the maximum efficiency in terms of data recovery. The resulting browsing structure was then presented back to the authors of the thesaurus and the Forum for Information Standards in Heritage (FISH) for comment and review.

### *The Media Hierarchy*

One of the initial requirements of the demonstrator was that it should identify rich content and, in the presentation of results, that it should favour rich content. This was implemented in the demonstrator by allowing users to select only records that had rich media attached to them. To achieve this, an additional hierarchy was created that defined different media types – audio, video, documents and images – and therefore allowed users to use only those with rich media. The number of data sets offering rich content remained proportionally small: just over 4% of the available data points opened anything more than a database record. Nonetheless, the media hierarchy is simple and successful, allowing for some unexpected insights.

### *The "When" Hierarchy*

The temporal classification is based on the simple controlled vocabulary for period terms recommended by FISH: the MIDAS Archaeological Periods list. This controlled vocabulary is based on a combination of numerical dates and period terms, and though promoted as a national standard, it was

in fact designed as a period definition for Southern England. This produces a number of anomalies within the demonstrator: for example the "Roman period" is defined as starting in AD 43 and ending in AD 410; dates consistent with traditional chronologies for southern England, although the Roman occupation of Scotland was considerably shorter, and the occupation of Wales and South-western England quite possibly longer. The resulting anomalies were noted and where possible steps were taken to minimise them. However, the project team recognised that thorough resolution of this issue would require a significant effort beyond the resources of this project.

Although presented as a simple word list, the temporal classification is hierarchical and this hierarchy is implemented in the demonstrator. However, as an archaeological period list, the terminology for modern periods is limited. Consequently a number of additional terms were added to the catch-all 'modern' classification to increase the likelihood of success for the classifier and to improve the user experience.

### *Geographic and Cartographic Classification*

The spatial classification is in two parts: a geographic classification and a cartographic one. The first relates to place-names and areas, while the latter is concerned with the UK national grid. These two classifications were developed as separate strands, though the relationship between them is close. In outline, the geographic classification starts with England, Scotland, and Wales. Once again, the classification leaned on the FISH recommendations. The regions in England were the nine government regions, the counties were the post-1974 counties, and the districts were a mix of the district authorities that form part of the remaining counties and unitary authorities. In Scotland, a regional split between north and south was borrowed from a Scottish Executive definition, though it is recognised that this definition is unlikely to be more widely recognised. Thereafter the unitary authority areas are used. Wales is divided into the operating areas of its four archaeological trusts, and thereafter into pre-1974 counties.

Considerably more detail is invisible to the user but is essential for the geographical classification. The lowest level of the classification (English District, Scottish Authority and Welsh County) is an extensive list of associated place names. For example, parishes, non-parish areas and communities are all included in the hierarchy. This level of detail seems most attractive for usability reasons and is technically possible given extended resources. Space as a container for information is infinite and consequently every system for classifying against space can always be more detailed. A balance between detail of analysis and the ability of the computing infrastructure to sustain detailed analysis is unavoidable.

This same balance between detail and computing power is more evident in the cartographic classification. At the outset, the project team analysed the quality of cartographic information available, and sought to balance the detail on screen against the quality of the data. Although most of the data points were accurate to within 100 metres, it was evident that a significant minority were accurate to no more than 1000 metres, and often less than that. We therefore assumed a display threshold of 10 km, and divided Great Britain into 8418 10 km squares, corresponding to 690 km on the east-west axis and 1220 km on the north-south.

### **System Architecture**

At the core of the CIE demonstrator are four basic items: a data set of around 1,000,000 entries, a classification with upwards of 12,000 entries, the Waypoint classification engine and a Sun server with 12 GB of RAM. The performance and scalability of the demonstrator are functions of these four variables. It would be possible to expand and enhance any one of these elements, but only at the expense of the others: and only with extra investment would it be possible to extend each of them together. The project team believes that this demonstrator has been the largest experiment of its kind with facetted classification, using the largest data set and largest classification scheme so far brought together.

## *The Archaeobrowser Interface*

A key element of the project was bringing the benefits of the facetted classification system to the user through a sophisticated and thoroughly user tested interface. The interface consists of a number of complementary elements that guide the user through the discovery process by allowing them to select

facets with a mouse click rather than simply typing in a search term and hoping for the best. Of course, a traditional search box is still provided, with a slight enhancement whereby recognised terms are 'auto-completed' from the systems range of thesauri. At each stage in the resource discovery process, the user is allowed to switch between numerous data views that show, for example, the geographical distribution of records and the actual record set itself. Naturally, the geographical interface can be used to further narrow a search in the Where category, again with a click of the mouse.

A key element of the interface is the constantly available indication of the number of records that match the search criteria alongside the search criteria themselves. For example, with no criteria selected the record interface clearly shows over a million records are available. With just 3 mouse clicks, corresponding to when, what and where, a user can refine this to under 20 records. At that point, the user may decide to switch to the browsing panel, 20 being a more easily readable number of records than a million. The user can also reset the selected facets and start again with the original million or so records. The most innovative element of the interface is the clickable facet selection element that visually cues the user to select facets in the facet hierarchy and then automatically reveals the next level of facets in the hierarchy. This allows but does not force the user to continue to refine his or her search to the lowest level of detail allowed.

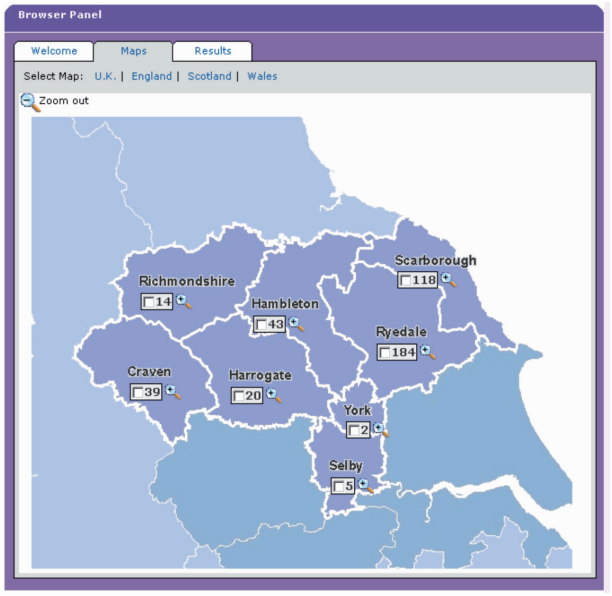


Fig. 1. The browser panel displaying the number of monuments selected within each region.

Although the project invested considerable time and effort into the development of the classification system and the associated software tools, the user is most likely to appreciate the clickable, intuitive and informative interface that was developed to take advantage of the facetted classification system.

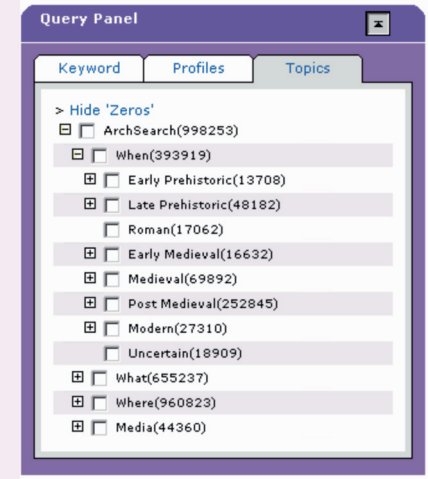


Fig. 2. The query panel with the When hierarchy expanded to the top level of period browsing.

## *Conclusion*

The Archaeobrowser demonstrator fulfilled the objectives of the project as originally conceived. It was a major achievement to provide proof of concept for a distributed common information environment using facetted classification across a range of data sets and the extensive test of data standards and protocols for information sharing across a range of public sector agencies. However, its most significant outcomes may be the large-scale test of facetted techniques to classification, with special reference to geospatial data and gazetteer services. In addition, it produced a range of XML-based normalisation tools, a simplified "knowledge map" for historic environment data, and a portable "code fragment" access point (portlet) into the facetted index.

At the core of facetted classification lies the idea of a "knowledge map" – in effect a poly-hierarchical specialist thesaurus in which the relationship between different concepts are ordered. This knowledge map becomes the principal support for users browsing the underlying data. Experience on this project shows that subject specialist thesauri are not inherently user friendly, so it is naïve to presume that a thesaurus can be deployed as knowledge

maps without considerable editorial input. However, contrary to the trend in information retrieval, it is also apparent that a well-formed knowledge map reduces our dependence on metadata. By explicitly structuring the relevant concepts, many of the problems in metadata search and retrieval were resolved. In some cases, the metadata presented to the indexer was simply superfluous to requirements. In others, the indexer was able to resolve inconsistencies by referring to the concept map. Consequently, the demonstrator presents two counter-intuitive conclusions: that where there is a well worked out knowledge map, even summary metadata schemas (e.g. Dublin Core) can be over-specified; and schemas which muddle concepts need not hinder resource discovery.

Fundamental to the success of this demonstrator have been services, such as geoXwalk, which translate between the semantics of different information communities. Geography is a uniquely powerful tool to access and unify diverse information, but this opportunity can only be exploited when reliable translation services exist. This is true of other non-spatial semantics; though equivalent middleware "cross walking" services are still in their infancy.

The most significant additional enhancement to the browser interface would be to develop further the geospatial element by adding GIS functionality to the user interface. This enhancement is eminently achievable, but will require additional work separating the interface from the underlying index. The relatively simple raster display currently used also dictates the sorts of geography adopted behind the scenes. The use of more sophisticated vector-based graphics may have the additional benefit of enabling adopt more sophisticated polygonal geometries for resource discovery and analysis

## *References*

Ross / Janevski / Stoyanovich 2007
K.A. Ross / A. Janevski / J. Stoyanovich, A Faceted Query Engine Applied to Archaeology. Internet Archaeology 21, 2007. http://intarch.ac.uk/journal/issue21/stoyanovich_index.html [2007].

*Stuart Jeffrey*

*Archaeology Data Service, Department of Archaeology*
*The King's Manor, University of York*
*Exhibition Square*
*York YO1 7EP, United Kingdom*
*sj523@york.ac.uk*

*William Kilbride*

*Glasgow Museums Resource Centre*
*200 Woodhead Road*
*South Nitshill Industrial Estate*
*Glasgow G53 7NN, United Kingdom*
*William.Kilbride@cls.glasgow.gov.uk*

*Stewart Waller*

*Archaeology Data Service, Department of Archaeology*
*The King's Manor, University of York*
*Exhibition Square*
*York YO1 7EP, United Kingdom*
*sjw143@york.ac.uk*

*Julian Richards*

*Archaeology Data Service, Department of Archaeology*
*The King's Manor, University of York*
*Exhibition Square*
*York YO1 7EP, United Kingdom*
*jdr1@york.ac.uk*