musicSpace: Integrating Musicology's Heterogeneous Data Sources

David Bretherton, Daniel Alexander Smith, mc schraefel, Mark Everist, Jeanice Brooks, Richard Polfreman and Joe Lambert.

~ ~ ~

A significant barrier to the research endeavours of musicologists (and humanities scholars more generally) is the sheer amount of potentially relevant information that has accumulated over centuries. Whereas researchers once faced the daunting prospect of physically scouring through endless primary and secondary sources in order to answer the basic whats, wheres and whens of history, these sources and the data they contain are now increasingly available online. Yet the vast increase in the online availability of data, the heterogeneity of this data, the plethora of data providers, and, moreover, the inability of current search tools to manipulate metadata in useful and intelligent ways, means that extracting large tranches of basic factual information or running multi-part search queries is still enormously and needlessly time consuming. Accordingly, the musicSpace project is exploiting Semantic Web technologies (Berners-Lee et al., 2001) to develop a search interface that integrates access to musicology's largest and most significant online resources. This will make previously intractable search queries tractable, thus allowing our users to spend their research time more efficiently and ultimately aiding the attainment of new knowledge. This brief paper gives an overview of our work.[1]

[1] musicSpace (http://www.mspace.fm/projects/musicspace) is a joint music and e-science research project based at the University of Southampton. The Principle Investigator is dr mc schraefel, Co-Investigators are Prof. Mark Everist, Prof. Jeanice Brooks and Dr Richard Polfreman, and the project is funded by the Arts and Humanities Research Council (http://www.ahrc.ac.uk), the Engineering and Physical Sciences Research Council (http://www.epsrc.ac.uk), and the Joint Information Systems Committee (http://www.jisc.ac.uk).

~ ~ ~

The digitisation of musicology's central resources has revolutionised the research process, yet dispersal of material across numerous libraries and archives has now been replaced by segregation of data into a plethora of discrete and disparate online database resources. These are typically segregated according to media type (text, image, audio, video), date of publication, subject, language, and/or copyright holder, and often limited funding inevitably results in databases of modest remit. Yet almost all musicological research cuts across these artificial divisions, meaning that musicologists are routinely forced to consult an extraordinarily heterogeneous body of online data repositories and catalogues. In short, a significant amount of valuable research time is expended in establishing basic factual information, not because the data is unavailable, but because the lack of database integration requires extensive manual collation. Not only is this an inefficient use of time, but also it means that large, complex data queries are essentially intractable (especially when the quality of a resource's metadata is poor).

This can be a major disadvantage at any stage of the research process. For example, a musicologist trying to mould an inchoate thought about Monteverdi's madrigals into a well-formed research question would need to execute the same keyword searches several times each because there are several relevant data sources, and would also need to perform numerous additional searches to account for all the synonyms that exist for the term 'madrigal' ('concerted madrigal', 'concerto', 'madrigale', 'madregal', 'madriale', 'marigalis' and 'matricale'). Similarly, because of the segregation of data into discreet and disparate databases, and the limitations of currently deployed search interfaces, real-world multi-part questions such as 'which scribes have created manuscripts of Monteverdi's works, and which other composers' works have they inscribed?' or 'which singers have recorded the operas Mozart composed during the 1780s, what other operatic roles have they taken, and where can I get hold of their recordings?' have to be broken down into their component parts, queried separately using multiple data sources, and finally collated, all of which takes hours or even days.

Oxford University Press and Alexander Street Press, two leading providers of musicological material, have recently responded to the need for database integration by providing integrated portals to their respective online repositories (see http://www.oxfordmusiconline.com and http://music.alexanderstreet.com). But, because both press's portals only provide access to the material from that press, and rely on existing search technology, the difficulties described above are fundamentally unresolved. By contrast, musicSpace is working in partnership with musicology's major repositories (the British Library Music Collections, the British Library Sound Archive, Copac, Cecilia, OUP's Grove Music Online, Naxos, RILM and RISM) to integrate access to their data sources, while developing an advanced user interface for the manipulation of metadata, so that search queries like those identified above can be answered in minutes or even seconds. This will allow musicologists to find the information they need more easily and to discover information that they did not think to look for, and will also encourage additional whimsical – but potentially fruitful – searches.

~ ~ ~

Because the data held by our data partners has been created by different organisations and for different purposes, it is marked-up using different schemas, and as such, an ontological alignment process has to be performed. Our work in integrating these data sources is not completely from scratch, however, as some of our data partners have already mapped their internal schema into MARC encoding (a system for machine-readable cataloguing created by the Library of Congress in the 1960s). In order to further align the sources, we have developed a shallow hierarchy based on information type, which provides the facets for a faceted browsing interface. For each data source, we developed a mapping from their schema (or their choice of MARC encodings) to our shallow type hierarchy. We developed software to use our mappings to map the data into an RDF representation of our type hierarchy. By using RDF for the integrated set of data, we can make use of many benefits of Semantic Web technologies, one of which is the facility to create multiple files of RDF at different times and using different tools, and assert them into a single graph of a knowledge base, and query all of the asserted files as a whole.

One of the challenges in aligning heterogeneous data sources is that of entity co-reference. It is rare that data providers share identifiers for entities (such as people and works), and as such, we have to perform co-reference mapping ourselves. For the musicological data we are aligning in musicSpace, a straightforward string matching system is appropriate to match entities across sources. To ensure greater confidence in these matches, we have developed a semi-automated system that enables musicologists to check the mappings and inform the system of any changes that need correcting. Whenever a mapping is automatically performed, our system adds the mapping to a gazetteer, using the two strings that were matched, and a small amount of contextual metadata from both records to aid understanding. The gazetteer is then ordered by confidence, so that a musicologist can check over the low-confidence mappings carefully, update the gazetteer (either to remove the mapping, alter it, or provide a replacement), and inform the co-reference software of the changes. By using this approach we can be confident that the data sources are aligned properly, and that any updates to the data sources will re-use the manually corrected gazetteers.

Exploration of the integrated data sources is performed through the mSpace faceted browser (schraefel et al., 2006), which provides a scalable web-based faceted browsing interface for large-scale data sets. Faceted browsing is an alternative complementary search paradigm to keyword searching, which is the most common form of large-scale data exploration. The faceted interface customisation used by musicSpace presents columns that list attributes from a number of facets of the data, such as ‘date’, ‘musical work title’, ‘composer’ and ‘genre’, allowing the user to make selections in these facets in order to filter down results. The interface is reactive, in that the lists of facets are updated every time a selection is made, so that subsequent choices are limited to those that would yield results (the user is never offered a choice that would yield no results).

The faceted and reactive nature of the interface enables complex questions to be addressed, such as that posed earlier concerning Monteverdi and scribes. Figure 1 shows a screenshot of the interface, in which ‘Monteverdi’ is selected in the ‘Composer’ column so that

associated scribes are returned in the 'Copyist/Scribe' column. Selecting a scribe, in this case John Immyns, and reordering the columns so that 'Copyist/Scribe' precedes 'Composer', as in Figure 2, means that the 'Composer' column now returns composers whose works have been inscribed by Immyns.

Figure 1. The column interface in the musicSpace prototype shows four facets: 'Source Collection', 'Composer', 'Copyist/Scribe' and 'Manuscript Score'. Selection of 'Monteverdi, Claudio' in 'Composer' has been made, as well as 'Immyns, John' in 'Copyist/Scribe', and the interface has filtered the results in 'Manuscript Score' to a single record that matches these selections: 'Giovinetta pianta, La'.

Figure 2. Following from the interaction in Figure 1, the user has dragged the column 'Copyist/Scribe' leftwards, so that the selection 'Immyns, John' now filters on the 'Composer' column, as well as the 'Manuscript Score' column, so that the user can see works by other composers that had John Immyns as the copyist.

~ ~ ~

mSpace is a modern interface that utilises Web2.0 technologies such as AJAX (a client-server query mechanism built on existing web architectures) to improve the response time of the service, and supports

sharing of findings over other Web2.0 services such as *del.icio.us*, *Facebook* and *StumbleUpon*, so that users can save and share their results with colleagues and the wider internet.

Over the next year a team of musicologists will use musicSpace during their everyday research. We will monitor how they use musicSpace in order to assess its efficacy as a research tool.

References

Berners-Lee, T., Hendler, J., Lassila, O.: 'The Semantic Web'. *Scientific American* (2001).

schraefel, m. c., Wilson, M. L., Russell, A. and Smith, D. A.: 'mSpace: improving information access to multimedia domains with multimodal exploratory search'. *Communications of the ACM*, 49 (4), pp. 47-49 (available at http://eprints.ecs.soton.ac.uk/12376) (2006).