musicSpace: Improving Access to Musicological Data

mc schraefel, David Bretherton, Daniel Alexander Smith, and Joe Lambert

Problem

Not unlike the sciences, musicological data is widely distributed and exists in numerous formats and in many databases. Efforts over the past decade to digitize bibliographies, artists' works lists, recordings, program guides and related ephemera should mean that data that was once physically distributed (requiring a researcher to visit various sites to consult items) is all readily accessible from the comfort of one's desktop. But this has not entirely been the case: the geographical dispersal of material in the physical world has been replaced by the dispersal of data into a plethora of discrete and disparate online databases ("data silos") in the virtual world, according to, for example, media type (text, image, audio, video), date of publication, subject, language, and/or copyright holder. Because musicological research typically cuts across such artificial divisions, musicologists thus routinely expend valuable research time consulting a bewildering number of heterogeneous data repositories.

A related problem is that the quality of the metadata often of course determines the quality of the search. In this regard, note that even some of musicology's leading data providers use legacy or ad hoc metadata specifications that are unsuited to modern demands. For instance, a source that provides lists of the works that a composer has composed has no consistent format for the production of these lists, while another is inconsistent with terminology – a composer in one record is an author in another. So, while the data has been digitized, there is no easy way to ask questions like "which scribes have created manuscripts of a composer's works, and which other composers' works have they inscribed?", and "which poets have had their poems set as songs by Schubert, and which other song composers have also set them?"

The combined effect of these shortcomings is that such real-world musicological research questions are effectively intractable, not because the data needed to answer them is not in a data source somewhere, but because there is insufficient metadata or metadata granularity, and a lack of data source integration (meaning that metadata from one source cannot be used as the basis for a query of another source).

Towards solutions

There is one seemingly obvious solution to the above query dilemmas that has been well modelled in eScience: enable integrated real-time querying over all the available metadata, and enable people to use that metadata to guide their queries. This means making existing metadata usable and surfacing data that might be more useful as metadata to guide queries. The musicSpace project, drawing on its roots in the

UK eScience Program [1], has taken a dual approach to realizing this vision: designing back-end services to integrate (and where necessary surface) available (meta)data for exploratory search; and providing a front-end interface to support rich exploratory search interaction. The musicSpace interface (Figure 1) uses the "mSpace" faceted browser [2], which provides a scalable web-based faceted browsing interface for exploring large datasets and utilizes AJAX to improve response times. As we have discussed our UI elsewhere [3, 4], we concentrate below on the data side of our work.

Integrating datasets

musicSpace's data partners include the key international providers of musicological data: the British Library and British Library Sound Archive, Cecilia, Copac, Grove Music Online, Naxos Music Library, RILM, and RISM UK and Ireland [5]. Despite the establishment of protocols for sharing metadata from the Open Archives Initiative [6], and developments in federated search [7] and Semantic Web approaches to music data [8, 9], only a very small number of our partners have implemented such systems, typically due to funding constrains or the desire to protect intellectual property. Hence data is currently provided to us manually.

Our partners supply data adhering to a number of different schemas and serializations (including MARCXML, MODS XML, custom MARC, source-specific XML, and CSV tables), and we have taken a purpose-driven approach to unifying the data sources using a multi-level metadata hierarchy with a common purpose-built high-fidelity ontology. (We found the more general Music Ontology [9] to be unsuited to our partners' music-bibliographic research data, although we do allow metadata to be exported as Music Ontology RDF.) The upper level of our metadata hierarchy includes, for example, "Person" and "Score", while the sub-level adds granularity to "Composer" and "Manuscript Score" respectively (among other possibilities). Imported data is mapped to an RDF representation of our type hierarchy. By using RDF we can make use of the 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, assert them into a single graph of a knowledge base, and query all of the asserted files as a whole.

musicSpace's combined dataset currently includes some 200,000 records and 4.5 million RDF triples. While small in relation to science databases, for humanities subjects (particularly musicology) this constitutes an integrated dataset of considerable size and demonstrates the scalability of our approach.

Enhancing metadata

In many cases we were able to directly map a record field from a partner's dataset to our combined

type hierarchy, but in other cases some light syntactic and/or semantic analysis needed to be performed. For instance, some sources (including RISM) state a person's name followed by a three-letter "relator code" [10] to indicate their role in that record, such as "Arnold, Samuel [fmo]"; in these cases we extract the name and role as two individual but related facts to allow us to associate "Arnold, Samuel" with the role of "Former Owner" in that record. Useful information can similarly be extracted from the "leader" code [11] of MARC records. This small amount of work in the pre-processing stage adds granularity that significantly enriches the data, allowing for more refined filtering and browsing of records via the UI.

Likewise, to deal with inconsistent naming and formatting conventions of works lists in Grove, our solution is to use a semi-automated approach whereby a purpose-built tool extracts data incrementally, which is then checked and edited by a domain expert before being deposited in our database.

Significantly, although in the above cases the "hidden" data we extract is present in the original records, it is neither exposed to nor exploitable by the end-user via our data providers' existing UIs. In musicSpace, however, this hidden data is surfaced so that it can be used by the musicologist for the purposes of querying the dataset, and can thus aid the process of knowledge discovery and creation.

Evaluation

The mSpace UI upon which musicSpace is built has been evaluated for exploratory search usability in a variety of contexts [12, 13], and so our main focus in testing the musicSpace browser is its impact on musicological research. Specifically, we are interested in how well it supports the kinds of queries musicologists want it to enable, and what new kinds of research questions, as yet unanticipated, it might enable. An initial phase of evaluation was completed during April-May 2009 and feedback was very encouraging – the speed of the interface, search flexibility, and level of data granularity were all praised [4]. A longitudinal evaluation is currently ongoing, but early indications are that numerous previously intractable queries have indeed been enabled, including those mentioned earlier.

Generalizability

The main takeaway from this project has been understanding how the application of approaches to data developed for eScience may enhance humanities research. From our work on the musicSpace project, we offer an effective generalizable framework for data integration and exploration that is well suited for Arts and Humanities data. Our benchmarks have been (1) to make tractable previously intractable queries, and thereby (2) to accelerate knowledge discovery towards innovation. We look forward to sharing our work and tools for your consideration and uptake.

Figures

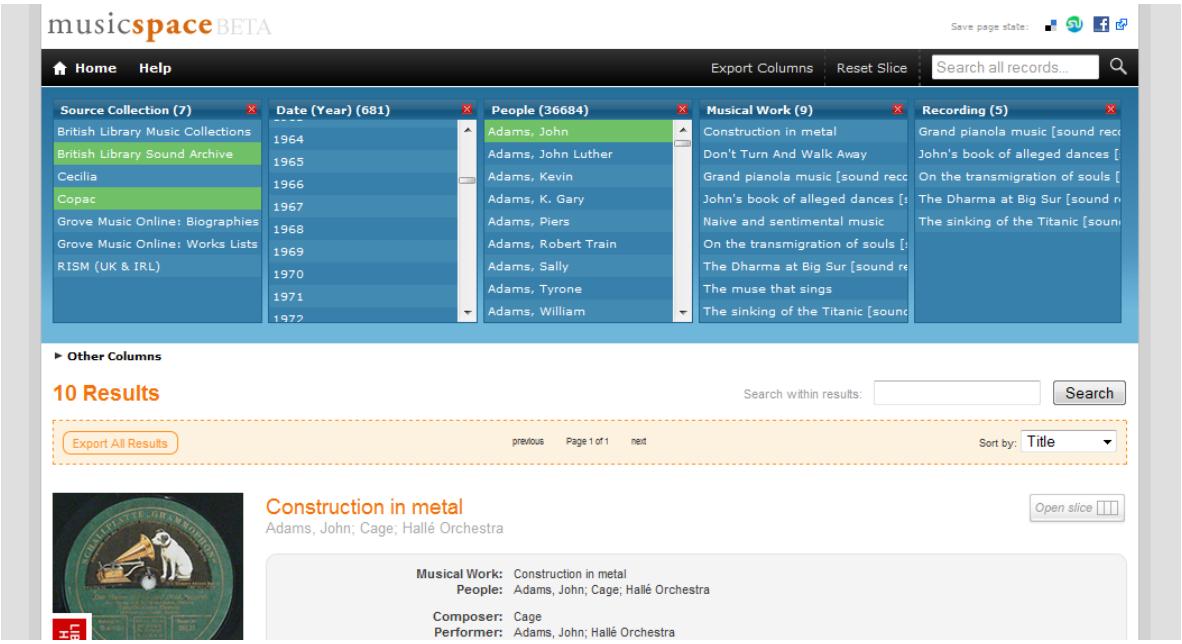

Figure 1: A screenshot of the musicSpace interface (try at http://musicspace.mspace.fm).

References

[1] See: http://smarttea.org, http://mytea.org.uk, http://www.combechem.org.

[2] See: http://mspace.fm

[3] mc schraefel, M. L. Wilson, A. Russell, and D. A. Smith, "mSpace: Improving Information Access to Multimedia Domains with Multimodal Exploratory Search", *Communications of the ACM*, 49:4, 2006, pp. 47-49.

[4] D. Bretherton, D. A. Smith, mc schraefel, R. Polfreman, M. Everist, L. J. Brooks, and J. Lambert, "Integrating Musicology's Heterogeneous Data Sources for Better Exploration", *Proceedings of the 10th International Society for Music Information Retrieval Conference*, 2009, pp. 27-32.

[5] Data partners: the British Library, http://www.bl.uk; the British Library Sound Archive, http://www.bl.uk/nsa; Cecilia, http://www.cecilia-uk.org; Copac, http://copac.ac.uk; Grove Music Online (OUP), http://www.oxfordmusiconline.com; Naxos Music Library, http://www.naxosmusiclibrary.com; RILM, http://www.rilm.org; RISM UK and Ireland, http://www.rism.org.uk.

[6] C. Lagoze, and H. Van de Sompel, "The Open Archives Initiative: Building a Low-Barrier Interoperability Framework", *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, 2001, pp. 54-62.

[7] C. N. Cox, ed., *Federated Search: Solution or Setback for Online Library Services*, Haworth Information Press, Binghamton NY, 2007.

[8] C. Lai, I. Fujinaga, D. Descheneau, M. Frishkopf, J. Riley, J. Hafner, and B. McMillan, "Metadata Infrastructure for Sound Recordings", *Proceedings of the 8th International Conference on Music Information Retrieval*, 2007, pp. 157-158.

[9] Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson, "The Music Ontology", *Proceedings of the 8th International Conference on Music Information Retrieval*, 2007, pp. 417-422.

[10] See: http://www.loc.gov/marc/relators/index.html

[11] See: http://www.loc.gov/marc/bibliographic/bdleader.html

[12] M. L. Wilson, and mc schraefel, "A Longitudinal Study of Exploratory and Keyword Search", *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2008, pp. 52-56.

[13] M. L. Wilson, mc schraefel, and R. W. White, "Evaluating Advanced Search Interfaces Using Established Information-Seeking Models", *Journal of the American Society for Information Science and Technology*, 60:7, 2009, pp. 1407-1422.