

Orchestrating Musical (Meta)data to Better Address the Real-World Search Queries of Musicologists

David Bretherton (D.Bretherton@soton.ac.uk), Daniel Alexander Smith, mc schraefel (PI), Richard Polfreman, Mark Everist, Jeanice Brooks and Joe Lambert

1. musicSpace's Objectives

- To integrate access to musicology's heterogeneous data sources so that they can be explored effectively and efficiently via one interface service.
- To deliver an optimally interactive approach to support this exploration.
- To develop a better understanding of how musicologists use the musicSpace interface, so that it can be optimised to support the process of discovery and aid the attainment of new knowledge.

2. Motivation

Musicologists consult a wide variety of heterogeneous data sources in their research, many of which are now available online. Yet digitisation has been accompanied by the segregation of data into numerous discrete repositories, and this, along with inadequate metadata, insufficient data/search granularity, and poor search UIs, prevents the intelligent manipulation of metadata. This means that questions requiring advanced cross-source filtering on metadata fields or the running of complex multipart search queries have to date been effectively intractable.

Sample intractable questions provided by our musicologists include:

'Which scribes have created manuscripts of a composer's works, and which other composers' works have they inscribed?'

'Which poets have had their poems set as songs by Schubert, which other song composers have also set them, and where can I get recordings of these settings?'

To address such questions using existing online database resources, one would have to break them down into their component queries, perform those queries on several different sources, and manually collate the results. Given the volume of results returned at each stage, and that the granularity of the data held is often poor, this process can require significant human intervention. Herein lies their effective intractability.

3. Experimental Solutions

There is at least one seemingly obvious solution to the above query dilemmas: enable integrated real-time querying over all the available metadata, and enable people to use that metadata to guide their queries. The associated issues for this solution also imply that all data that could be construed as useful is extracted in some way, and that there is an interaction approach that will enable this metadata to be explored effectively to formulate rich compound queries. To this end, we have taken a dual approach to addressing this exploration problem: 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.

3.1 Multi-Source Integration

Our data partners include the British Library Music Collections catalogue, the British Library Sound Archive catalogue, Cecilia, Copac, Grove Music Online, Naxos Music Library, RILM, and RISM (UK and Ireland). Their metadata is supplied adhering to a number of different schemas and serialisations (MARCXML, MODS XML, custom MARC, source-specific XML, and relational database CSV tables). On import data is mapped onto a representation of our custom multi-level metadata type hierarchy in RDF, which allows us to take advantage of the many benefits of Semantic Web technologies.

For some datasets, such as RISM, it was necessary to enhance or better expose detail already buried in the records. In others, such as Grove, data in full-text articles needed to be given semantic meaning. After trying in vain many methods for providing this meaning, we found the most reliable solution was to employ human intelligence, specifically that of domain experts. To this end, we designed a web-based tool that shows a domain expert small amounts of data at a time, so that he/she, taking into account the context, can enter the data into a form (see the screenshot above) for eventual import into musicSpace's triplestore.

3.2 User Interface

Data sources integrated into musicSpace are explored via a customised version of the 'mSpace' faceted browser, which provides a scalable web-based faceted browsing interface for large-scale datasets and utilises the AJAX client-server query mechanism to improve response times. Faceted browsing is an alternative complementary search paradigm to keyword searching, the latter currently being the most commonly deployed 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,' 'Composer' and 'Genre', allowing the user to make selections in these facets in order to filter down results (see the screenshot at the top right of this poster). 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.

4. Queries Made Tractable

Pre-processing of metadata and data source integration, combined with mSpace's optimally interactive search interface, enable musicSpace to be used to address complex multi-part research questions. Consider the discussion below of the two sample questions given earlier in Section 2.

4.1 'Which scribes have created manuscripts of a composer's works, and which other composers' works have they inscribed?'

To use musicSpace to address the above question, the user would: (i) arrange the three facets 'Composer', 'Copyist/Scribe' and 'Manuscript Score' in that order; (ii) select a composer in the 'Composer' column (e.g. 'Monteverdi, Claudio') so that associated scribes are displayed in the 'Copyist/Scribe' column, (iii) select a scribe (e.g. 'Immyms, John') so that results associated with both that scribe and the selected composer are returned in the 'Manuscript Score' column (see screenshot to the right).

Following from this interaction, (iv) by leaving the scribe ('Immyms, John') selected, dragging the 'Copyist/Scribe' column to the left of the 'Composer' column, and deselecting the selection made in the 'Composer' column, the user can discover other works by other composers that were inscribed by the selected scribe (see screenshot to the left).

4.2 'Which poets have had their poems set as songs by Schubert, which other song composers have also set them, and where can I get recordings of these settings?'

The basic search procedure for this question is similar to that just described in Section 4.1, but, given its greater complexity, more facets are required: (i) the columns 'Genre', 'Composer', 'Author/Lyricist/Librettist', 'Musical Work' and 'Recording' are arranged in that order, then (ii) by selecting 'Songs' in the 'Genre' column and 'Schubert, Franz (Peter)' in the 'Composer' column, poets who have had their words set by Schubert are returned in the 'Author/Lyricist/Librettist' column. (iii) By selecting 'Goldoni, Carlo' (for example) in the 'Author/Lyricist/Librettist' column, associated musical works and musical recordings are returned in the 'Musical Work' and 'Recording' columns (as the screenshot below illustrates).

Next, (iv) to identify other song composers that have set texts by the same poet, the selection 'Goldoni, Carlo' (for example) is retained in the 'Author/Lyricist/Librettist' column, the user drags that column leftwards so that it sits between the 'Genre' and 'Composer' columns, and deselects 'Schubert, Franz (Peter)', with the effect that other composers who have set texts by the same poet are now returned in the 'Composer' column, and associated musical works and recordings are returned in the 'Musical Work' and 'Recording' columns (see the screenshot below).

5. Evaluation

Our musicologists' sample questions have had a profound effect on the musicSpace project, providing us with a user-focused means for continually assessing and judging the development of our interface. The benefits of focusing on real-world user queries in this way was born out by an initial period of internal testing and evaluation of musicSpace that took place in May 2009, during which testers reported speed gains and the improved tractability of complex queries in comparison to existing interfaces. These very positive early results suggest that musicSpace has the potential to become a powerful tool for helping musicologists to deal intelligently and productively with large and heterogeneous datasets.

During the summer of 2009 we refined our methods for mapping, enhancing and generating metadata, and polished the user interface. A broader public trial has just begun, which will enable us to assess and evaluate musicSpace's efficacy as a research tool. We anticipate that the results of this large-scale evaluation will demonstrate that musicSpace allows musicologists to find the information they need more easily and to discover information that they did not think to look for, thus enabling the discovery of new knowledge.

We would welcome additional testers working in the following areas of musicology: Monteverdi recordings, Schubert song, 19th-century Italian opera buffa, and 20th-century electroacoustic music.

Our Funders:



Arts & Humanities
Research Council

EPSRC

Engineering and Physical Sciences
Research Council

JISC