INTEGRATING MUSICOLOGY'S HETEROGENEOUS DATA SOURCES FOR BETTER EXPLORATION

David Bretherton, Daniel Alexander Smith, mc schraefel,
Richard Polfreman, Mark Everist, Jeanice Brooks, and Joe Lambert
University of Southampton, Southampton, UK, SO17 1BJ
D.Bretherton@soton.ac.uk; {ds, mc}@ecs.soton.ac.uk;
{R.Polfreman, M.Everist, L.J.Brooks}@soton.ac.uk; jl2@ecs.soton.ac.uk

ABSTRACT

Musicologists have to consult an extraordinarily heterogeneous body of primary and secondary sources during all stages of their research. Many of these sources are now available online, but the historical dispersal of material across libraries and archives has now been replaced by segregation of data and metadata into a plethora of online repositories. This segregation hinders the intelligent manipulation of metadata, and means that extracting large tranches of basic factual information or running multi-part search queries is still enormously and needlessly time consuming. To counter this barrier to research, the "musicSpace" project is experimenting with integrating access to many of musicology's leading data sources via a modern faceted browsing interface that utilises Semantic Web and Web2.0 technologies such as RDF and AJAX. This will make previously intractable search queries tractable, enable musicologists to use their time more efficiently, and aid the discovery of potentially significant information that users did not think to look for. This paper outlines our work to date.

1. INTRODUCTION

A significant barrier to the research endeavours of musicologists is the sheer volume of potentially relevant information that has accumulated over centuries. Researchers once faced the daunting prospect of manually scouring through seemingly endless primary and secondary sources in order to answer the basic whats, wheres and whens of musicology, particularly when making lists of people or repertoire according to specific criteria. Many of the sources needed to address these queries are becoming available online. Yet the dramatic increase in the online availability of data, the variety of data subjects, the growing number of data providers, and, moreover, the inability of current mainstream search tools to manipulate the associated metadata in useful ways, means that extracting large tranches of basic factual information (e.g. manuscripts once owned by "a," opera roles performed by "b") or running multi-part search queries (e.g. composers from place "c" that were active during decade "d") is still enormously and needlessly time consuming.

Accordingly, the "musicSpace" project <http://www.mspace.fm/projects/musicspace> is exploiting Semantic Web [1] and Web2.0 technologies to develop an experimental innovative search interface that integrates access to some of musicology's largest and most significant online data and metadata repositories, including 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. We anticipate that integrating heterogeneous metadata sources into one exploratory search user interface will allow our users to spend their research time more efficiently, make previously intractable search queries tractable, and ultimately open up new avenues for musicological study.

musicSpace is exploring and developing numerous methods for enhancing and generating additional metadata from our data partners' particularly heterogeneous data sets, and a primary focus is the development of web-based UIs and the longitudinal analysis of their effects on musicological scholarship and human-computer interaction. This distinguishes our work from that of previous notable projects concerned with music data source integration, such as Variations2 <http://variations2.indiana.edu> and EASAIER <http://www.easaier.org> [2, 3]. The "mSpace" framework and interaction layer of musicSpace has been designed and evaluated [4, 5] specifically to support multiple browsing and exploratory search tactics that go beyond common keyword search. Our user interface gives the provenance of all records, and is designed not only to help musicologists discover relevant resources, but also to enable them to go from musicSpace to those resources in their original context in a single click. Beyond these core features, there are numerous support services based on related usability research to assist with collecting,

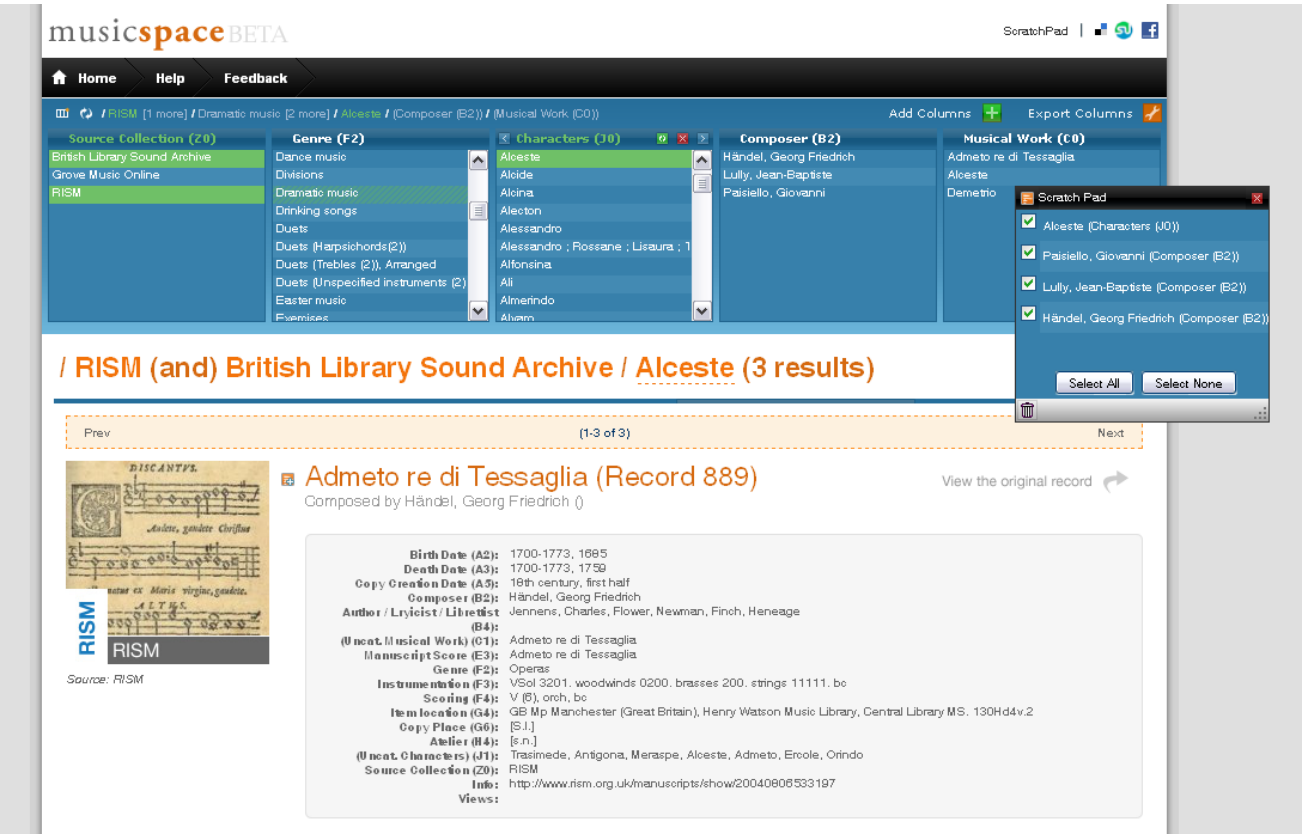

Figure 1. The musicSpace interface in use.

organising, exporting, and sharing information relevant to a particular query. It should also be noted that as musicSpace is a Web2.0 application, a web browser is all that is required to access the interface, a screenshot of which is given in Figure 1.

In this paper we give an overview of our work so far and outline the findings of our initial trial of the musicSpace browser interface. To begin, we review the motivation for our approach to supporting musicological knowledge building.

2. MOTIVATION: BARRIERS TO EFFICIENCY

2.1 Database Heterogeneity

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 usually segregated according to media type (text, image, audio, video), date of publication, subject, language, and/or copyright holder. Yet typical musicological research cuts across these artificial divisions, meaning that musicologists are routinely forced to consult an extraordinarily heterogeneous body of online data repositories. In short, a significant amount of valuable research time is expended in establishing basic factual information, not because the data is unavailable, but because a lack of database integration requires extensive manual collation of discovered data. This problem of heterogeneity is exacerbated by the fact that search interfaces to data providers' content remain almost universally rooted in the now somewhat dated 'textbox-based' search paradigm. Not only does the current situation mean that users' research time is used inefficiently, but it also means that large, complex data queries are essentially intractable.

These barriers 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. Similarly, because of the segregation of data into disparate, discreet databases and the limitations of currently deployed search interfaces, real-world multi-part queries 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 that 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 can take hours or even days.

Recently, a number of academic publishers, including Oxford University Press (with Oxford Music Online

<http://www.oxfordmusiconline.com>) and Alexander Street Press (with Alexander Street Press Music Online <http://muco.alexanderstreet.com>), have recognised the benefits of integrating their musicological data sources [6, 7]. However, because their portals only provide access to their own data repositories, and because their interfaces rely on existing textbox-based search technology, their work only takes us partway towards overcoming the barriers to research highlighted above; there remains a pressing need for further integration of data sources and better interaction support for more diverse search paradigms.

2.2 "Intractable" Queries

The musicSpace team includes musicologists who specialise in four pilot research areas: Monteverdi recordings, Schubert's songs, nineteenth-century opera buffa, and twentieth-century electroacoustic music. At the start of the project we asked our musicologists for examples of queries that they considered intractable (or, more specifically, not readily tractable) using the current search interfaces of our data providers, such that they had largely given up on a particular line of enquiry, and which they hoped that musicSpace would be able to facilitate. The list of queries suggested included:

A. Which scribes have created manuscripts of a composer's works, and which other composers' works have they inscribed?
B. Which performers have recorded Monteverdi's madrigals, and what else did they record in the same years?
C. 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?
D. Which singers have sung the role of Malatesta in *Don Pasquale*, and what else have they sung?
E. Which comic operas were composed in the nineteenth century and premiered in the twentieth?
F. Which electroacoustic works were published within five years of their premier?

It will be noted that all the above queries have multiple parts, and, therefore, if one were to use current search interfaces, one would have to break them down into their component queries and manually collate the results. There are several further obstacles to tractability. Queries B, C, D and F call (in particular) for several data sources to be consulted (for Queries B and D, for example, one would want to consult both the Naxos Music Library and the British Library Sound Archive catalogue), and so data source integration would clearly be beneficial in these cases. In addition, increased metadata granularity is a necessary prerequisite for the tractability of Queries A, C, D and F (for example, in Query A one would rely on metadata in RISM, yet although it is possible to use RISM's interface to search by "Person," it is not possible to further restrict this to "Composer" or "Scribe"). Finally, in addressing Queries C, E and F one would necessarily wish to consult the works lists in Grove Music Online. However, because these works lists are not marked up semantically, a system to generate relevant metadata from the raw data is needed (this particular issue is currently being addressed by musicSpace, and will be reported on at a later date).

3. EXPERIMENTAL SOLUTIONS: APPARENT INTEGRATION

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, even if buried in the database records, is extracted in some way, and that, similarly, there is an interaction approach that will enable this metadata to be explored effectively to formulate the kinds of rich compound queries described above.

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. We discuss these components below.

3.1 Multi-Source Integration

Despite advances in the development of protocols for shareable metadata in the form of the Open Archives Initiative <http://www.openarchives.org> [8], federated search [9], and, more recently, the application of Semantic Web technologies to the domain of music [10, 11], only a very small number of musicSpace's data partners offer such systems for the harvesting of metadata. This is typically either because funds are presently unavailable to meet the costs of implementing such systems, or, in the case of some data providers, because metadata is considered to be as much of an intellectual property asset as data content itself. Hence our data partners' data sets are currently provided to us manually.

We have thus taken a purpose-driven approach to unifying the metadata from our data partners, which is supplied adhering to a number of different schemas and serialisations (MARCXML, MODS XML, custom MARC, and source-specific XML). In order to unify these sources for the purposes of cross-source exploration, we have created static mappings from the schemas used by each data provider to a two-level hierarchy based on metadata type. The upper level of the hierarchy includes, for example, "Person" and "Score," while the sub-level respectively adds granularity to "Composer" and "Manuscript Score" (among other possibilities). In some cases we were able to directly map a record field to our type hierarchy, while in other cases some light syntactic and/or semantic analysis was performed on the source data. For example, some sources denote a person with

their name, followed by their role in that record, e.g. "J. S. Bach (composer)." In this case we extract the name and role as two individual related facts to allow us to associate "J. S. Bach" as "Composer" in the record, rather than simply "J. S. Bach (composer)" as "Person." This pre-processing of the metadata adds granularity to the source data and allows richer filtering and exploration through the browsing interface. We developed a tool to map the imported data to an RDF representation of our type hierarchy. By using RDF for the integrated set of data, 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.

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, 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; we use Alignment API [12], which uses Wordnet. 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, documenting the two strings that were matched along with a small amount of contextual metadata from both records to aid understanding. The gazetteer is then ordered by confidence, so that a musicologist – with reference to the Library of Congress Authorities website <http://authorities.loc.gov> – 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 sure that the data sources are aligned properly, and that any updates from our data partners will re-use the manually corrected gazetteers.

Because of the legacy issues that many of our data partners have to contend with, there are inevitably shortcomings and inconsistencies in their database structures, schemas, and records. But by using gazetteers in the string matching process, adding contextual metadata, and increasing granularity as records are imported, we are able to negate any such data quality issues. In addition, our approach means that we do not have to maintain copies of our data partners' databases for ourselves; rather, we provide a user interface service that provides a single point of entry to our data partners' repositories.

3.2 User Interface

Data sources integrated into musicSpace are explored via a customised version of the "mSpace" faceted browser [4, 5], which provides a scalable web-based faceted browsing interface for large-scale data sets 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. 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.

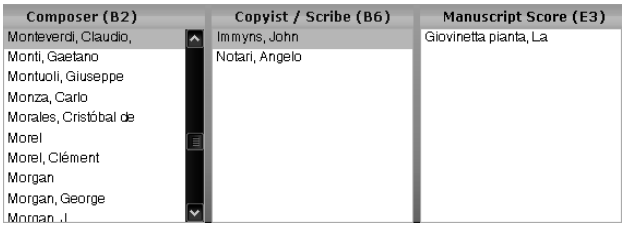

Figure 2. Scribes associated with the composer "Monteverdi, Claudio."

Figure 3. Composers associated with the scribe "Immyns, John."

The faceted and reactive nature of the interface enables complex queries to be addressed. Let us consider the query "which scribes have created manuscripts of Monteverdi's works, and which other composers' works have they inscribed?" In Figure 2, the musicSpace interface is showing three facets: "Composer," "Copyist/Scribe," and "Manuscript Score." The selection "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." Following from this interaction, in Figure 3 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 scribe.

3.3 Saving, Exporting, and Sharing Findings

Each interaction with the musicSpace interface generates a specific URL that, when re-entered into a web browser at a later stage, will return users to exactly that same point in the data exploration process. Thus users can pause and resume their research at any time by using the bookmarking feature common to all web browsers, and, moreover, they can save, share, and disseminate their findings with colleagues, students, and the wider internet by using Web2.0 services such as del.icio.us, Facebook and StumbleUpon, all of which can be accessed by clicking the appropriate icon in the musicSpace interface. Exporting of findings via email is also supported. In addition, musicSpace has the facility to allow users to access and export metadata as RDF (using the Music Ontology <http://musicontology.com> [11] as a data model), but licensing restrictions with our data partners currently prevent us from doing so for all data sets.

4. EVALUATION

Since the mSpace UI has been evaluated for exploratory search usability in a variety of contexts, our main focus in testing the musicSpace application is its impact on research: how well is it supporting the kinds of queries musicologists want it to enable? And, likewise, what new kinds of research questions, as yet unanticipated, may it enable? Towards answering these questions, we have recently completed an early pilot study. We describe our findings below. While these are early stage tests, our intention in outlining our findings here is to have knowledge of our approach and preliminary results available within the Music IR community in order to enhance engagement with the project.

4.1 First Phase

A version of the musicSpace interface was released internally to a team of six musicologists for an initial period of testing and evaluation on 29 April 2009, and their feedback was very encouraging. Although this initial release did not integrate our full spread of data sources, testers nevertheless reported significant improvements with search speed and ease:

- "All the information showed up very quickly, and it was easy to find material. It was really good to have different kinds of material in the same place."
- "[musicSpace offers] a speedier way to research crossed search pathways."
- "Excellent interface – very simple to understand."

Testers were also impressed with the way that musicSpace's faceted interface allowed for browsing around a subject and for instantaneous paradigmatic shifts in search focus:

- "I would recommend musicSpace for its ability to manipulate queries in order to get results that you wouldn't otherwise be able to get [without starting over]."
- "I liked the ability to explore around a topic once I'd identified something of interest."
- "The ability to switch columns around and add new columns was most useful."

Aside from these early hoped-for indications that musicSpace will provide a quicker and more flexible way to explore a variety of musicological data sources, testers also reported that increased search data granularity (as compared to that of our data partners' search interfaces) was a substantial benefit. For example, a number of testers were pleased by musicSpace's facility to browse by opera character:

- "[Without using musicSpace] it would not be at all easy to do a character search. You would have to use printed reference books like *Pipers Enzyklopädie des Musiktheaters* [13], but even this does not have an index of characters, so you'd have to look at the entry for each opera and draw up character lists by hand. You would also have to know what you were looking for before you started out!"
- "I used musicSpace to explore how many operas have a character named Alceste. This information simply isn't get-at-able using other search interfaces – you'd have to sort through the information on your own."

There was similar enthusiasm for musicSpace's ability to browse by scribe and the former owner of manuscripts.

4.2 Future Phases

Over the coming months there will be incremental releases of musicSpace, each expanding the data set, refining our data mappings, and polishing the UI. This process will culminate in a broader public release towards the end of 2009, which will enable us to assess its real-world efficacy as a research tool.

5. CONCLUSION

Early results from our testing of musicSpace's ability to enable rapid and effective exploratory search across heterogeneous musicological sources are promising. Our testers clearly appreciated the speed gains of integrating data sources; in fact the only recurring negative comments from testers during our initial period of evaluation concerned their desire to see still more data repositories integrated into musicSpace. In addition to data source integration, both increased data granularity and the flexibility of faceted browsing were found to be very beneficial. These three features enabled testers to explore data in a way that had not previously been possible, and a number of intractable queries were indeed made tractable.

In his keynote address to this conference in 2005, Nicholas Cook predicted that "working with larger data sets will open up new areas of musicology" [14]. But if

Cook's prediction is to be realised, then increasing the size and number of data sets that musicologists work with both demands and allows for better systems to integrate those data sets, and also for far more sophisticated systems for manipulating data. To this end, our research demonstrates a potentially powerful approach for helping musicologists to deal intelligently and productively with large and heterogeneous data sets. We believe that musicSpace will allow musicologists to find the information they need more easily, and to discover information that they did not think to look for. In so doing, it may also encourage additional speculative – but potentially fruitful – searches, thus enabling the discovery of new knowledge.

6. ACKNOWLEDGEMENTS

The musicSpace 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>. Our data partners include 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>, and RISM UK and Ireland <http://www.rism.org.uk>.

7. REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila: "The Semantic Web," *Scientific American*, Vol. 284, No. 5, pp. 34–43, 2001.

[2] J. W. Dunn, D. Byrd, M. Notess, J. Riley, and R. Scherle: "Variations2: Retrieving and Using Music in an Academic Setting," *Communications of the ACM*, Vol. 49, No. 8, pp. 55–58, 2006.

[3] C. Landone, J. Harrop, and J. Reiss: "Enabling Access to Sound Archives through Integration, Enrichment and Retrieval: the EASAIER Project," *Proceedings of the Eighth International Conference on Music Information Retrieval*, pp. 159–160, 2007.

[4] mc schraefel, D. A. Smith, A. Owens, A. Russell, C. Harris, and M. L. Wilson: "The Evolving mSpace Platform: Leveraging the Semantic Web on the Trail of the Memex," presented at *Hypertext*, Salzburg, 6–9 September 2005.

[5] 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*, Vol. 49, No. 4, pp. 47–49, 2006.

[6] L. Macy: "Letter from the Editor," *Oxford Music Online*, March 2008 <http://www.oxfordmusiconline.com/public/page/letter_08>.

[7] A. Hall: "Alexander Street Press: New Developments," presented at the *Academic Music Librarians' Seminar*, Birmingham Conservatoire, 21 May 2009.

[8] 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.

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

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

[11] 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.

[12] J. Euzenat: "An API for Ontology Alignment," *Proceedings of 3rd International Semantic Web Conference*, pp 698–712, 2004.

[13] C. Dahlhaus, and S. Döhring, eds: *Pipers Enzyklopädie des Musiktheaters: Oper, Operette, Musical, Ballet*, 7 Vols, Piper, Munich, 1986–1997.

[14] N. Cook: "Towards the Complete Musicologist," *Proceedings of the Sixth International Conference on Music Information Retrieval*, 2005.