The MusicNet Composer URI Project

Daniel Alexander Smith, David Bretherton, Joe Lambert, and mc schraefel
University of Southampton

Motivation

In any domain, a key activity of researchers is to search for and synthesize data from multiple sources in order to create new knowledge. In many cases this process is laborious, to the point of making certain questions effectively intractable because the cost of the searches outstrip the time available to complete the research. As more resources are published as Linked Data, and with the development of appropriate tools, data from multiple heterogeneous sources should be more rapidly discoverable and automatically integrable, enabling previously intractable queries to be explored, and standard queries to be significantly accelerated for more rapid knowledge discovery. But Linked Data is not of itself a complete solution. One of the key challenges of Linked Data is that its strength is also a weakness: anyone can publish anything. So in classical music, for instance, 17 sources may publish data about ‘Schubert’, but there is no *de facto* way to know that any of these Schuberts are one and the same, because the sources are not aligned. Without alignment, much of the benefit of Linked Data is diminished: resources can effectively be stranded rather than discovered, or tangled nets of only guessed at associations can cost more time than they are worth to determine whether a particular dataset is relevant or not.

Our previous work on the musicSpace project [1] attests to importance of data alignment, and was in fact the main impetus for the MusicNet project. In musicSpace we integrated access to musicology’s key online resources [2], using the ‘mSpace’ [3] faceted browser, to demonstrate how commercial- and research-developed heterogeneous data resources could be integrated for rapid exploration and knowledge building [4] (a longitudinal evaluation of this work is currently ongoing). One of the most tenacious challenges to aligning heterogeneous data sources for musicSpace proved to be that of entity co-reference, specifically the fact that our data partners rarely used the same identifiers for entities such as composers. Although the Library of Congress, for example, provides an authority service for names and items that can be used in the creation of library metadata, this operates as a commercially run subscription service, and so there is a price barrier to smaller organisations and individual creators of datasets. Many data providers have also voiced objections to us regarding the ‘authorized names’ promoted by the Library of Congress, on the basis that they rarely represent the fullest or truest version of a name (middle names are usually omitted, for example). Furthermore, even where data providers have subscribed to the Library of Congress’s authority service, the size of their data legacies makes it impractical for them to retrospectively amend the name-forms used in pre-existing records.

Our experiences working with stakeholders and with their datasets suggests that a service to support the alignment of identifiers across data sources is a *sine qua non* necessity for the creation of new Linked Data musicology resources, and for the translation of existing resources into Linked Data, if such resources to be optimally useful and usable, as well as sustainable.

Solution

The MusicNet project is set to address the challenge just outlined by “minting” URIs for key musicology assets, to provide a framework for the effective exploration of Linked Data about classical music. The Linked Data that MusicNet produces (and which it will shortly begin to publish) is derived from the metadata of the musicSpace project’s data partners, with unique URIs being minted for each composer that exists within their current datasets. The data will be exposed using existing Linked Data technologies (RDF and the Music Ontology [5, 6]) and will form the basis of an online source of canonical data about – and, in time, comprehensive index of – the names and vital details of classical music composers. As well as exposing basic metadata about each composer (including full name, place and date of birth and death, and nationality), we will also expose URLs that reference back into the online web catalogues of our data partners, so as to allow musicologists immediate access to relevant data from each partner collection, and to enable partner collections a means to link to each other via our URIs. The URIs will be maintained to secure future archival integrity.

In addition to URI minting, the datasets from each data partner will need to be aligned to ensure that composers from one dataset will match up with composers from another if they represent the same person. This matching should be capable of handling different formatting of names (composer disambiguation) as well as input errors occurring when the data partners digitised their catalogues. A subset of this co-reference alignment has been performed in the musicSpace project, and we propose for this project that the existing alignments are exposed as Linked Data, and that the alignment work be expanded to all composers within the data sets, by using an expanded version of our prototype alignment tool created for musicSpace [4].

Planned Demonstrator

In order to directly engage the benefits of the Linked Data to end-users, a web portal will be created over the data that will allow musicologists to search all of the Linked Data to find items of interest, and to get links to all references to those items in the partner collections. For example, a user interested in the works of Johann Sebastian Bach searches the portal, and the search will find all Linked Data that references Johann Sebastian Bach (but not other Bachs) and will offer relevant links to all of our partners' collections, as well as Linked Data publishers such as DBPedia [7], MusicBrainz [8] and the BBC [9] (who publish Radio 3 playlists as Linked Data).

Impact

While the MusicNet project is in its early stages (it runs for a year until May 2011), our intention in outlining the project here is to have knowledge of our work available to the research community in order to enhance engagement with the project.

The outputs of MusicNet will greatly benefit researchers by improving the workflow of musicological research, proving a trusted codex of links to scholarly and commercial data sources where information on specific composers can be located. This will be equally valuable in music teaching environments. By aligning data sources and allowing open linking to other academic and commercial data sources, the visibility of these sources to academic researchers and students will be increased. Additionally, publishers of Linked Data recognise the usefulness of authoritative identifiers, in order for their data to be useful outside of its context. For example, although DBPedia provides URIs for some composers, the coverage is limited, while MusicBrainz URIs confuse performers and composers, leading to ambiguity when applied to classical music. Yves Raimond (BBC) has reported to us that these problems have impacted the usefulness of the BBC's classical music Linked Data output; a problem that will be solved by MusicNet's approach.

Conclusion

MusicNet, by minting URIs for composers, will remove the limitations on how data partners represent composers' names in their data sources by enabling them to link to other sources that represent names differently. In addition, by including data about name variants in different sources, MusicNet will also address the issue of compatibility with legacy data.

Composer data is fundamental to the work of musicologists and music educators; the establishment of authoritative URIs for composers, and moreover the disambiguation of composers in online data sources that will flow from this, is an essential first step in the provision of Linked Data services for classical music and musicology. Our work provides a model that can usefully be applied to other humanities disciplines, and beyond.

Acknowledgements

MusicNet is funded by JISC [10] and runs from June 2010 to May 2011.

References

[1] http://musicspace.mspace.fm/

[2] Data partners: http://www.bl.uk, http://www.bl.uk/nsa, http://www.cecilia-uk.org, http://copac.ac.uk, http://www.oxfordmusiconline.com, http://www.naxosmusiclibrary.com, http://www.rilm.org, http://www.rism.org.uk.

[3] http://mspace.fm/

[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] http://musicontology.com/

[6] 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.

[7] http://dbpedia.org/

[8] http://musicbrainz.org/

[9] http://www.bbc.co.uk/

[10] http://www.jisc.ac.uk/