Prospects for a Big Data History of Music

Stephen Rose
Music Department,
Royal Holloway, University of London
Egham TW20 0EX
44 1784 443806
stephen.rose@rhul.ac.uk

Sandra Tuppen
The British Library
96 Euston Road
London NW1 2DB
44 207 412 7500
sandra.tuppen@bl.uk

ABSTRACT
This position paper sets out the possibility of a musicology based on the analysis of musical-bibliographical metadata as Big Data. It outlines the work underway, as part of the AHRC-funded project A Big Data History of Music, to align seven major datasets of musical-bibliographical metadata. After discussing some of the technical challenges of data alignment, it suggests how analysis and visualization of this data might transform musicological understandings of cultural transmission and canon formation.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Online Information Services – Standards

General Terms
Digital humanities, data alignment.

Keywords
Musicology, canon, music publishing, metadata, bibliography.

1. INTRODUCTION
Studies of the history of Western music are frequently dominated by notions of musical greatness; scholars focus on a handful of composers whose work has been canonized for its aesthetic qualities. This bias towards studying 'Great Composers' is still evident in the documentational infrastructure of musicology: online reference sources such as Grove Music Online (www.grovemusic.com) are primarily structured via articles on individual composers, whereas the rich topographies of Western musical culture could equally well be interrogated via an analysis of locations, publishers, performers, genres, social practices or migratory patterns.

The AHRC-funded project A Big Data History of Music (running from January 2014 to March 2015) seeks to develop alternative ways to explore the history of Western music, principally through analyzing and visualizing the bibliographical data collected by research libraries. Through such quantitative analyses, musicologists will be able to develop research questions which probe the development of musical taste and compositional genres, and which examine trends in the transmission of music.

DLfM '14, September 12 2014, London, United Kingdom
 ACM 978-1-4503-3002-2/14/09…$15.00.
http://dx.doi.org/10.1145/2660168.2660177

2. RESEARCH CONTEXT
Since 2000 literary studies have been revolutionized by the opportunities offered by big data. Franco Moretti has pioneered the concept of 'distant reading', whereby quantitative analyses offer new perspectives on literary production and taste (as opposed to 'close reading' that focuses on a single text). By analyzing the production of novels during the 18th and 19th centuries, he has shown how the genre waxed and waned in response to external events and changing taste [1]. By analyzing the titles of seven thousand novels of the late 18th and early 19th centuries, he has offered a way to enter the 'archive of the "Great Unread"' as opposed to the 'world of the canon' [2]. Moretti has shown that the metadata (catalogue records) generated by libraries, containing such information as titles, author names and publication dates, can be valuable material in constructing alternative histories of literature and culture.

Musicologists, by contrast, have rarely used quantitative analyses in their work. Some pilot studies have been done by historians of music publishing, notably Rudolf Rasch's analyses of the retail costs of 18th-century printed music [3], and Tim Carter's investigation of the fluctuating output of printed music in Italy around 1600 [4]. Yet no systematic attempt has been made to analyze the existing datasets held by research libraries, comprising the metadata relating to their holdings of sheet music and other musical artefacts.

3. CORPORA
There currently exist numerous datasets of metadata relating to music publications, manuscript scores, and also the ephemeral materials that surround classical concert-giving (e.g. programme notes, handbills). Several of these datasets are the product of cumulative bibliographical research over many decades. Some are maintained by research libraries as part of their online catalogues, whereas others are curated by RISM (Répertoire International des Sources Musicales).

The AHRC-funded project A Big Data History of Music aims to combine seven of these datasets:

1. The British Library catalogue of printed music (http://explore.bl.uk), approx. 1 million bibliographic records
2. The British Library catalogue of music manuscripts (http://searcharchives.bl.uk), approx. 16,000 bibliographical records
3. Augustus Hughes-Hughes's *Catalogue of manuscript music in the British Museum*, 3 vols. (London, 1906–09) (not currently available online), containing descriptions of approx. 31,000 musical works (or parts of works)

4. RISM's bibliography of European music printed between 1500 and 1800 (RISM A/I), containing approx. 100,000 bibliographical records, currently available only in hard copy (*Répertoire International des Sources Musicales A/I: Einzeldrucke vor 1800*, 14 volumes, Kassel, 1971–2003) or as a CD-Rom (Kassel, 2011).
5. RISM's international union catalogue of music manuscripts dating from 1600 to c.1850 held in libraries worldwide (RISM A/II, http://opac.rism.info), approx. 900,000 records.
6. The RISM UK Music Manuscripts Database (www.rism.org.uk), approx. 55,000 records
7. The Concert Programmes Project database, describing collections of concert programmes held in UK libraries, and hence a record of historical musical performances ((http://www.concertprogrammes.org.uk), approx. 6,000 records.

Most of these datasets are currently held in proprietary library systems and are encoded using one of three international standards: MARC21, MARC XML or ISAD(G)/EAD.

After aligning these constituent datasets, our project will test various data analysis and visualization tools on the combined dataset. An open dataset will be made available in RDF-XML format and in delimited text files via the British Library website, and we will seek user evaluations at a symposium to be held in March 2015. For licensing reasons, the fourth and seventh datasets listed above will not be included in the open dataset.

4. CHALLENGES OF DATA ALIGNMENT

Before the datasets can be analyzed in any meaningful way, a certain amount of data alignment is required. The constituent datasets have been created over a long period of time; the descriptive and encoding standards used vary between datasets, and sometimes within an individual dataset. Different rules and standards exist for the cataloguing of printed and manuscript sources; British and American institutions have traditionally used different rules from libraries in mainland Europe; and data published originally in printed catalogues has sometimes been corrupted or misaligned during the process of conversion to electronic form.

Indeed, musicologists have previously been deterred from analyzing bibliographical datasets by concerns about the quality and reliability of the constituent data. As bibliographer David Hunter states: 'Accurate quantification of the output of printed music during our period [the long eighteenth century] is impossible at present owing to inadequate bibliographical control'. Hunter identifies 'inconsistency of coverage' and 'problems of definition' as the biggest pitfalls in existing bibliographical data [5].

The following paragraphs summarize the principal challenges of data alignment, as well as some of the solutions employed by our project.

4.1 Unreliable data

Bibliographical data may be unreliable for several reasons. Much data was created over a century ago and, depending on the subsequent development of musicological research, may contain outdated or erroneous information. Here manual updating of catalogue records may be the only way forward, despite being time-consuming and labor intensive.

Other data is conjectural, such as the dates that bibliographers assign to printed music of the late 18th and 19th centuries. The original items are usually undated, and bibliographers may assign dates that are rounded to the nearest decade. Such conjectural dates would distort any analysis of year-by-year production of printed music. In this respect it is necessary for any Big Data study to acknowledge the limitations of the original data. When we release the open dataset, we also aim to release documentation alerting analysts to the cataloguing conventions under which it was created.

Data may also be unreliable because coding problems occurred when data was migrated between catalogue formats, leaving data in the wrong field. Our project uses the tool MarcEdit to locate data that has been incorporated into the wrong field or corrupted during the migration process. Within the British Library Catalogue of Printed Music, for instance, our use of MarcEdit has detected data erroneously placed in the field for place of publication (MARC field 260$a), such as information about volume numberings and dates of publication.

4.2 Heterogeneous data

Another challenge is the heterogeneous nature of the data. The British Library catalogue records for printed music show a high degree of variability in their level of detail. The disparity is most extreme for bibliographical records concerning anthologies of 16th-century printed music (anthologies are those books containing pieces by more than one composer). The old catalogue records, created in the 19th century, generally record only the title of each book and the place / date of publication, with no information about the names of composers or the titles of compositions in these volumes. By contrast, the catalogue records upgraded in 2011 as part of the Early Music Online (EMO) digitization project (www.earlymusiconline.org) contain detailed information for every composition and composer in these books, transcriptions of title-pages, plus the names of printers, publishers, dedicatees, former owners, etc. The Big Data project is employing teams of research assistants and cataloguers to enhance all the catalogue records for the British Library's 16th-century anthologies to the level of detail used in EMO.

Further challenges of heterogeneity arise with the various forms in which names of composers, places and compositions have been recorded. While names of composers have usually been subject to standardization within a particular dataset, there is no international consensus on the 'preferred' form of a composer's name, with different countries having selected different authorized forms. For example, among the various 'preferred' forms of Handel's name are the following:

- Händel, Georg Friedrich
- Händel, George Frideric
- Haendel, Georg Friedrich
- Handel, George Frideric

In many instances it is possible to align personal names using recognized numerical identifiers for individual persons, from the Virtual International Authority File (VIAF) or International Standard Name Identifier (ISNI).

Place names and work titles are also problematic. While places can be aligned by geo-referencing, there is currently no internationally recognized standard for referencing a particular musical work. We are experimenting, however, with the creation of standardized titles for Latin-texted vocal music.

5. PROSPECTS FOR MUSICOLOGY

By analyzing a large corpus of bibliographical data, The Big Data History of Music project aims to offer many new perspectives on musicological debates, such as how music travelled across geographic boundaries, and how the formation of a canon occurred. Because the project's data analysis phase is yet to start, this paper will give a foretaste of its likely scope with examples of how the metadata created in 2011 for the Early Music Online project can be analyzed. EMO catalogued 324 anthologies of 16th-century printed music, less than 25% of the extant anthologies from this century. Hence the following comments are not definitive statements, but rather indications of how a Big Data approach might allow new ways of interrogating music history.

Many textbooks on 16th-century music emphasize such figures as Josquin des Prez and Giovanni Pierluigi da Palestrina [6]. Even remaining within a paradigm that focuses on individual composers, the EMO data suggests an alternative canon. The composers whose works appear most frequently in EMO are the chanson composer Thomas Crecquillon (149 works), followed by the prolific and versatile Orlande de Lassus (111 works), while Josquin has only 55 works and Palestrina only 7 works. To be sure, the low figures for Josquin and Palestrina reflect the different dissemination patterns of their works: Josquin favored manuscript dissemination and Palestrina favored single-composer editions. Still, the printed anthologies in EMO give an idea of which composers were most popular among the publishers assembling anthologies for commercial reasons.

The EMO data can also be interrogated in ways other than by composer name – for instance by place of publication, name of dedicatee, or by title of work. Many of these searches are facilitated by the interface built by the SLICKMEM project (Semantic Linking of Information, Content and Metadata for Early Music), which incorporates the EMO data [7] [8]. Searching by title can show interrelationships between pieces based on the same text or musical theme; this shows how many 16th-century compositions were based on popular tunes, and can be the basis for studies of *imitatio* (the practice whereby composers emulated each other's works).

Even the EMO records for the wording of title-pages can be profitably analyzed. Although the first music printers such as Ottavio Petrucci had sparse title-pages with limited information, subsequent firms such as Pierre Attaignant included more detailed title-pages to attract buyers, often noting that the music was 'new' or had never before been printed. The rise of this practice can be documented by an analysis of books in EMO that include the word 'new' (*nova, nouvelle, neue* etc.) on their title-pages. This practice began in the 1530s and increased until the 1550s, when about 40 per cent of the EMO anthologies claim to be 'new'. Figure 1 shows that these claims of novelty were primarily made on secular publications, which had to appeal to a market of amateur musicians; books of sacred music, by contrast, were more likely to emphasize that they conformed to liturgical requirements or to the demands of religious authority. Novelty and change were desirable factors in secular music but not in sacred music.

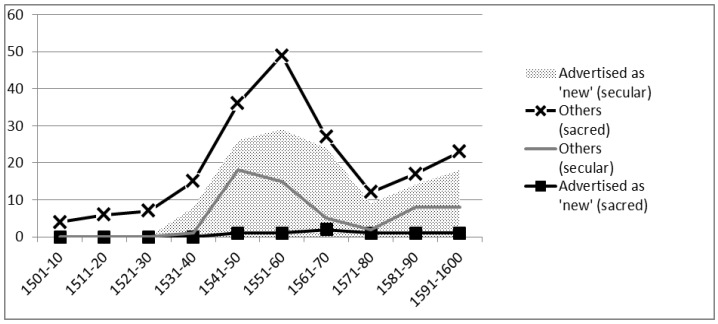

Figure 1: Printed anthologies advertised as 'new' on title-page

6. CONCLUSION

This paper has outlined some of the ways in which a Big Data approach to musical-bibliographical data might change understandings of music history. Many other avenues of exploration remain, for instance an analysis of the development of genres as indicated by the titles of published musical compositions (in which decades are the titles *Étude* or *Poem* used?) or a comparative analysis of where a composer's music was published and where it was performed. Although such studies can only be as accurate as the data within their constituent datasets, they open investigations into 'the Great Unheard' (to adapt Moretti's formulation) and show the varied ways in which musicologists can interrogate digital libraries.

7. ACKNOWLEDGMENTS

A Big Data History of Music is funded by the AHRC. This paper also analyses metadata created by Early Music Online, which was funded by JISC's Rapid Digitization Fund 2011.

8. REFERENCES

[1] Moretti, F. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London, Verso.

[2] Moretti, F. 2010. *Distant Reading*. London, Verso.

[3] Rasch, R. 2013. 'Wie viel bezahlte man für Musik im 18. Jahrhundert?' Paper at conference 'Das Leipziger Musikverlagswesen im internationalen Kontext', Sächsische Staatsarchiv, Leipzig, June 21, 2013.

[4] Carter, T. 1986/1987. 'Music Publishing in Italy, c.1580–c.1625.' *Royal Musical Association Research Chronicle* 20, 19-37.

[5] Hunter, D. 2009. 'Music'. In *The Cambridge History of the Book in Britain, vol. 5: 1695–1830*, ed. M. Suarez & M. Turner. Cambridge, Cambridge University Press. 750–761.

[6] Atlas, A. W. 1998. *Renaissance Music: Music in Western Europe, 1400–1600*. New York, Norton.

[7] Crawford, T., Fields, B., Lewis, D., Page, K. 2014. 'Explorations in Linked Data practice for early music corpora'. In *Digital Libraries 2014*. London, 8–12 Sept 2014.

[8] http://slickmem.data.t-mus.org/snorql