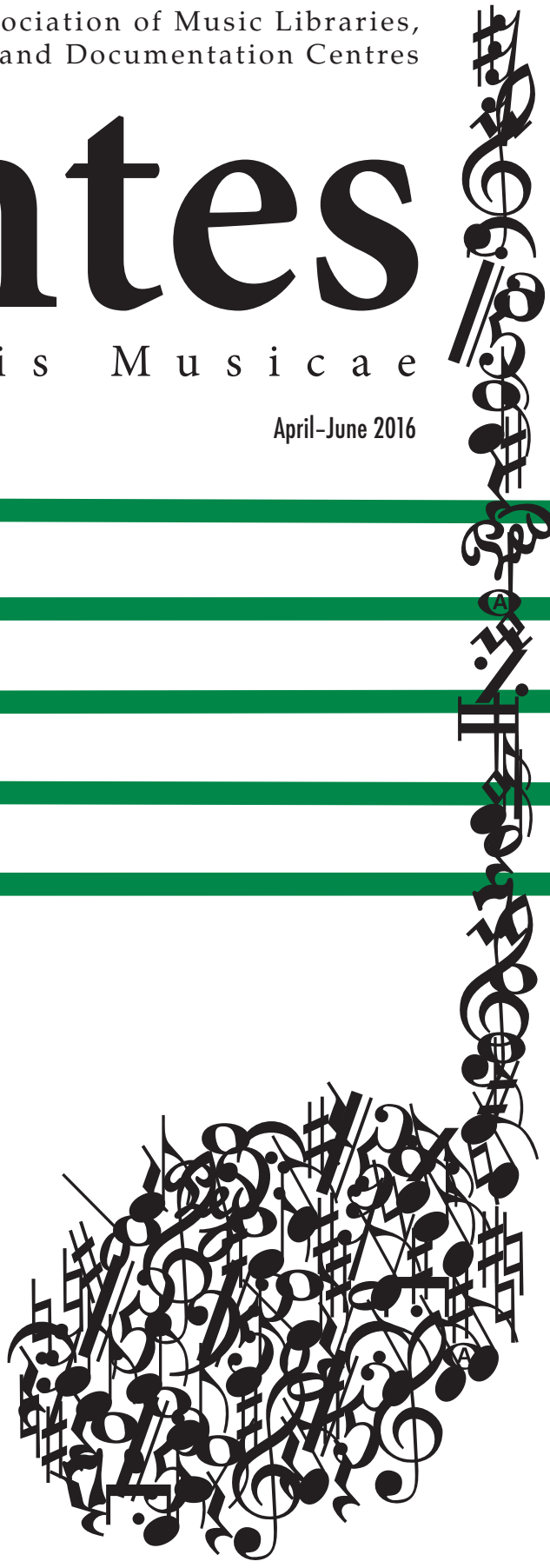
Journal of the International Association of Music Libraries,
Archives and Documentation Centres

Fontes

Artis Musicae

April–June 2016



ISSN 0015-6191 (print)
ISSN 2471-156X (online)

63/2

Journal of the International Association of Music Libraries, Archives and Documentation Centres (IAML) / Journal de l'Association Internationale des Bibliothèques, Archives et Centres de Documentation Musicaux (AIBM) / Zeitschrift der Internationalen Vereinigung der Musikbibliotheken, Musikarchive und Musikdocumentationzentren (IVMB)

FONTES is not available for sale: the journal is supplied only to members of the Association, with the price subscription included in membership dues. Membership application should be made to the IAML Secretary General or to the secretariat of the applicant's national branch. Business correspondence relating to mailing list rental, change of address, order for back issues, claims, and other matters should be sent to the Treasurer. See the inside back cover for addresses.

FONTES ist nicht im Handel erhältlich. Die Zeitschrift wird ausschließlich an Mitglieder der IAML abgegeben; der Bezugspreis ist im Mitgliedsbeitrag enthalten. Anträge auf Mitgliedschaft richten Sie bitte an den IAML-Generalsekretär oder an das Sekretariat Ihrer nationalen Gruppe. Es wird gebeten, die geschäftliche Korrespondenz bezügl. Adressänderung, Bestellung älterer Ausgaben, kostenpflichtiger Nutzung des Mitglieder verzeichnisses, Forderungen und sonstige Anfragen ausschließlich direkt an den Schatzmeister zu senden. Die Adressen finden Sie auf der hinteren Innenseite des Umschlages.

FONTES n'est pas disponible à la vente. La revue n'est adressée qu'aux membres de l'association, le prix de l'abonnement étant compris dans celui de l'adhésion. Les demandes d'adhésion doivent être faites auprès du Secrétaire général de l'AIBM ou du secrétariat de la branche nationale du demandeur. Toute correspondance concernant la location du fichier d'adresses, les changements d'adresses, la commande d'anciens numéros, les réclamations et autres sujets doit être adressée au Trésorier. Les coordonnées se trouvent en troisième de couverture.

FONTES is printed quarterly by A-R Editions, 1600 Aspen Commons, Suite 100, Middleton, WI 53562 USA

FONTES ARTIS MUSICAE

VOLUME 63/2, APRIL–JUNE 2016

CONTENTS

LIBRARY CATALOGUE RECORDS AS A RESEARCH RESOURCE: INTRODUCING 'A BIG DATA HISTORY OF MUSIC'

Sandra Tuppen, Stephen Rose, and Loukia Drosopoulou

Librarians and archivists are curating increasingly large quantities of digital data, not merely the data contained in catalogues and databases, but also digitised documents and data that originated in digital form. The dramatic growth in digital data that occurred between the 1980s and 2000s is graphically illustrated by the archives of two U.S. presidents. The library of George Bush senior (served as president 1989–1993) contains over forty million pages of textual documents and two million still photographs; the library of his son George W. Bush (served as president 2001–2009) preserves more than seventy million pages of documents, nearly four million digital photographs and over 200 million email messages.[1] Such an explosion of data poses a challenge to the information profession, which has responded by developing new roles such as (in the academic sector) the research data librarian.

Similar trends can be observed in scientific disciplines and in business, where the collection and analysis of vast quantities of data have become fundamental activities. The work of astronomers has been transformed by the Sloan Digital Sky Survey, which has released over 116 terabytes of open data collected by a wide-angle optical telescope at the Apache Point Observatory in New Mexico.[2] Experiments in particle physics at the Large Hadron Collider produce over thirty petabytes of data annually, for analysis by a grid consisting of more than 140 centres spread over twenty-five countries.[3] In the world of business, supermarkets and other retailers amass immense quantities of data on purchasing patterns, analysis of which then shapes their marketing and sales strategies. Although humanities researchers typically work with far smaller quantities of data than those in the scientific and business fields, they are also increasingly drawing on large datasets in their research. Scholars in linguistics and literary history now routinely use corpora of digital texts to analyse the use of language in particular settings: for instance, the Early English Books Online Text Creation Partnership has manually keyed in the full text of over 40,000 pre-1700 English-language books and made the data freely available for analysis.[4]

The phrase 'big data research' is now often used in association with such research activities. Big data has been defined as data which is high volume (in other words, it exists

---

Sandra Tuppen is Lead Curator, Modern Archives & Manuscripts, 1601–1850 at the British Library; Stephen Rose is Reader in Music at Royal Holloway, University of London, and a specialist in German and English music between 1550 and 1750; Loukia Drosopoulou was Postdoctoral Research Assistant on the project 'A Big Data History of Music'. Research for this article was funded by the Arts & Humanities Research Council (grant AH/L010046/1). Project datasets can be downloaded from www.bl.uk/bibliographic/download.html, accessed 11 March 2016.

1. http://bush41library.tamu.edu; https://www.georgewbushlibrary.smu.edu/en/Research/Presidential_Records.aspx, accessed 11 March 2016.

2. http://www.sdss.org/dr12/, accessed 11 March 2016.

3. http://home.web.cern.ch/about/computing, accessed 11 March 2016.

4. See http://quod.lib.umich.edu/e/eebogroup/, an interface which allows complex searches to be performed.

in large quantities), high velocity (more data is being added rapidly), or high variety (the data is heterogeneous); the phrase is also sometimes used to describe data which is too large or complex to be processed using traditional information technology (IT) systems.[5] A more nuanced view of big data has also emerged, that rather than the actual size of the dataset, it is its comparative size and complexity within the field in question that is important. Christof Schöch argues that big data is 'a relative term and a moving target, depending on context and available technologies', and that the distinctive mark of big data in the humanities appears to be a methodological shift rather than a primarily technological one.[6] Yuanjen Chen advocates defining big data as 'being able to use as much data as you need to derive actionable insights'.[7]

Within the various disciplines that study music, there are no datasets of comparable size to those used by scientists and social scientists. Yet there is considerable interest in applying 'big data' approaches to datasets about music bibliography and also data generated from sound recordings or machine-readable scores. In 2014–2015 the U.K.'s Arts and Humanities Research Council funded three projects exploring the application of 'big data' approaches to aspects of musicology. This article introduces one of these projects: 'A Big Data History of Music', a collaboration between Royal Holloway, University of London and the British Library.[8] This project aimed to explore ways of unlocking library catalogue data about printed and manuscript music and to pilot its use in new ways in the study of music history. Here we outline some of the challenges faced, describe some of the project's research findings, and highlight how others can now access the 'big data' music datasets that have been released for all to explore.

Contexts

At the heart of 'A Big Data History of Music' is the bibliographic data about printed and manuscript music held in library catalogues. While the prime function of this data is to enable library users to search for, identify, and access individual items of interest to them, it is also a rich source of information about the works and genres that were disseminated in a particular period and locality, and about the people involved in the music's creation, production, and transmission.

The potential value of analysing bibliographical data has already been demonstrated by the literary historian Franco Moretti. A specialist in the eighteenth- and nineteenth-century novel, he was frustrated with existing scholarship that tended to focus on the 'close reading' of a handful of canonised novels, ignoring the thousands of 'Great Unread'

5. The information technology consultancy Gartner has defined big data as 'high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing' (http://www.gartner.com/it-glossary/big-data, accessed 11 March 2016).

6. Christof Schöch, 'Big? Smart? Clean? Messy? Data in the Humanities', *Journal of Digital Humanities* 2, no. 3 (2013). Available at http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities, accessed 11 March 2016.

7. http://www.datasciencecentral.com/profiles/blogs/big-data-analytics-in-a-snap, accessed 11 March 2016. Chen is referring to big data in a business context but the definition is also applicable in an academic setting.

8. https://www.royalholloway.ac.uk/music/research/abigdatahistoryofmusic/home.aspx. The other projects funded were the Digital Music Lab (led by City University, London), exploring the analysis of large-scale audio collections (http://dml.city.ac.uk/) and Optical Music Recognition from Multiple Sources (led by Lancaster University), http://insight.lancaster.ac.uk/?p=262; all accessed 11 March 2016.

books from the period. As a solution, Moretti proposed the concept of 'distant reading'—in other words, the quantitative analysis of bibliographical data, which could thereby offer an overview of the production of novels in the eighteenth and nineteenth centuries. In his manifesto *Graphs, Maps and Trees* he demonstrated how political or military conflict often led to a collapse and then belated rise of novelistic production (as in France after 1789, or in Milan after the revolutionary movements and wars of the late 1840s). In an overview of the genres of English novels, he showed how each genre (for instance, the sentimental or the Gothic novel) was in favour for about twenty-five to fifty years, before fashion moved on to another genre.[9] Moretti has also analysed the titles of about seven thousand British novels from the late-eighteenth into the early-nineteenth centuries, noting the move towards shorter titles and then one-word titles after 1800, including titles that consist of the names of female protagonists, for example, *Emma*, *Lucy*, or *Caroline*.[10]

Moretti's work is controversial and has attracted accusations of positivism.[11] His quantitative analyses can give an aura of objectivity, when the data on which they are based may be far from comprehensive or accurate. Yet 'distant reading' can radically change the study of literature, showing a vast sweep of thousands of titles, and helping us judge whether the canonised novels (that are the usual subjects of study) are representative or not. Arguably Moretti's approach works best when practised in conjunction with the 'close reading' still favoured by most literary critics.

The notion of 'distant reading' is a radical challenge to the current scholarly infrastructure for researching music history. Most musicological reference works (such as *Grove Music Online*) are primarily set up to be searched by composer name. That is helpful for a researcher who wants to find out more about canonised figures such as Bach, Haydn, Mozart, or Beethoven, but less useful for a researcher who wishes to ask questions about the development of particular genres, or the texts favoured by vocal composers. In 'A Big Data History of Music', we wanted to experiment with reconfiguring musicological data so it could be interrogated in other ways, enabling a richer understanding of the topographies of music history. We also wanted to examine the usefulness of bibliographic data as a source of contextual information. As with literary scholarship, there can be a tendency in music history to focus on the canonised composers, at the expense of the more peripheral.

The starting-point for our project was a previous collaboration between Royal Holloway and the British Library, *Early Music Online* (*EMO*, www.earlymusiconline.org [accessed 11 March 2016]). Funded by a grant from JISC (the U.K.'s funding council for digital innovation in higher education), *EMO* digitised over 320 anthologies of sixteenth-century printed music. *EMO* was not just about digitisation; equally important was to open access to these books by thoroughly cataloguing their contents. The old catalogue records for these books, created in the nineteenth century, generally recorded only the title of each book and the place and date of publication, with no information about the names of composers or the titles of compositions in these volumes. As part of *EMO*, the catalogue

9. Franco Moretti, *Graphs, Maps and Trees: Abstract Models for a Literary History* (London: Verso, 2005), 11, 18–19.

10. Franco Moretti, 'Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850)', *Critical Inquiry* 36 (2009): 134–58.

11. *Reading 'Graphs, Maps and Trees': Responses to Franco Moretti*, ed. J. Goodwin & J. Halbo (Anderson, SC: Parlor Press, 2011), 4.

records were upgraded to contain detailed information for every composition and composer in each of these books, plus the names of printers, publishers, dedicatees, and former owners. This metadata is already a valuable scholarly resource that can be analysed for studies of music publishing, in addition to its more obvious function of enabling users to locate specific pieces of music in anthologies.

Project Datasets and Their Challenges

Inspired by the experience of creating the *EMO* metadata, our project sought to work with the largest and most significant musical-bibliographical datasets. Foremost among these are the datasets created by RISM (Répertoire International des Sources Musicales). RISM has been collecting data about printed and manuscript music held in libraries across the world since the 1950s, so it now holds the most comprehensive body of information on musical sources between ca.1500 and 1800. For data about printed music, we drew principally on RISM series A/I (publications containing works by a single composer before 1800), and RISM B/I (anthologies between 1500 and 1700). For manuscripts, we used RISM A/II, the vast database detailing manuscripts copied principally between 1600 and 1850 (http://opac.rism.info [accessed 11 March 2016]). Between them, these datasets contain over a million bibliographic records. While there are inevitably many omissions in RISM, the A/I and B/I inventories are the closest we have to a comprehensive and representative listing of music printed in Europe before 1800.

Although RISM has a very wide geographical scope, its chronological coverage is strongest for the sixteenth to eighteenth centuries. We also wanted to look at music history over a broader time frame, and for this purpose we drew on the catalogues of the British Library (BL). The BL's main online catalogue, 'Explore the British Library' (http://explore.bl.uk [accessed 11 March 2016]), contains over a million catalogue records describing printed music published between 1500 and the present day. Legal deposit legislation means that the BL is entitled to receive one copy of every music publication issued in Britain and Ireland. Through this the BL has amassed the biggest and most representative collection of British publications, which is supplemented by a vast collection of material published overseas. The BL's other online catalogue 'Explore Archives and Manuscripts' (http://searcharchives.bl.uk [accessed 11 March 2016]) contains some 16,000 descriptions of music manuscripts and music-related archival material, many of them brief descriptions of volumes of material that do not itemise the contents. More detailed information about the BL's pre-1900 music manuscripts is available in the printed catalogue edited by Augustus Hughes-Hughes, hereafter referred to as Hughes-Hughes.[12] This had previously been digitised and Optical Character Recognition (OCR) technology used to extract the text, and these text files were made available to our research team.

Once our project team began to study these various datasets, it became apparent that it would not be feasible to combine the disparate data into a single dataset. The individual datasets would instead need to be analysed separately, and where relevant the results could then be compared. This was partly because there are overlaps between RISM and the BL catalogues (much material listed in RISM is held in the BL) but also because of the heterogeneity of the data. This heterogeneity was one of the greatest challenges faced by the project team. Not only was there great variety in the type and quantity of data captured

12. Augustus Hughes-Hughes, *Catalogue of Manuscript Music in the British Museum* (London, 1906–1909).

in the datasets, but the data models underpinning them also varied, meaning that the 'unit of description' was not the same in each.[13]

In the RISM A/II dataset (fig. 1a), there is a separate catalogue record for each musical work within a manuscript. These are linked to a 'parent' record describing the manuscript as a whole. (Where a manuscript contains just one work, there is a single catalogue record which conveys information about the work, composer and the manuscript source.)

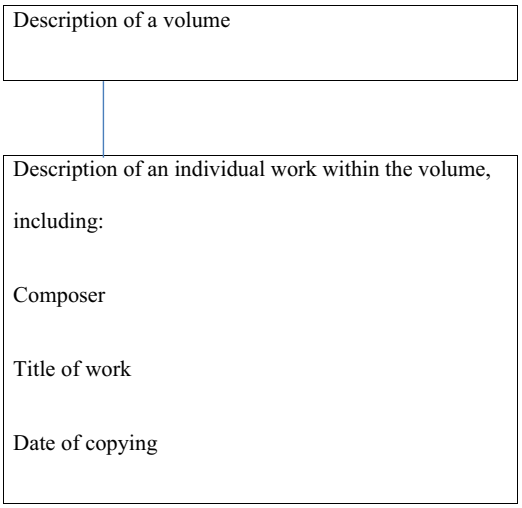


Fig. 1a. RISM A/II Record Structure

In contrast, each catalogue record in the RISM A/I dataset (fig. 1b) describes a volume of printed music, whether or not it contains more than one work. RISM A/I covers single-composer publications, and the composer is named, where known. However, where the publication contains multiple works, these are not usually itemised within the catalogue

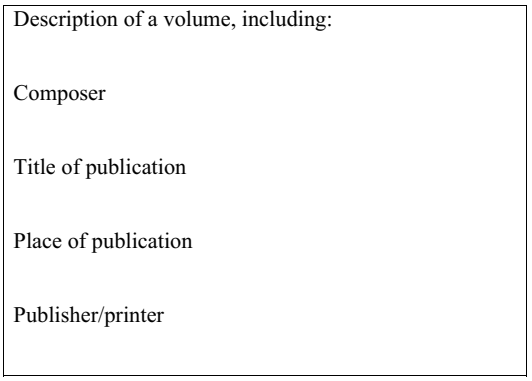


Fig. 1b. RISM A/I Record Structure

13. Schöch believes that heterogeneity is arguably the biggest challenge of data in the humanities, and also notes the impossibility of integrating heterogeneous datasets into one unified dataset. 'Big? Smart? Clean? Messy? Data in the Humanities', Ibid.

record. (The keys of individual works within a set of instrumental works such as sonatas might be given, but nothing more.)

Similarly, in RISM B/I (fig. 1c) the unit of description is the publication, in this case the anthology. Descriptions do not list the works within the volume, although there is an index of composers' names.

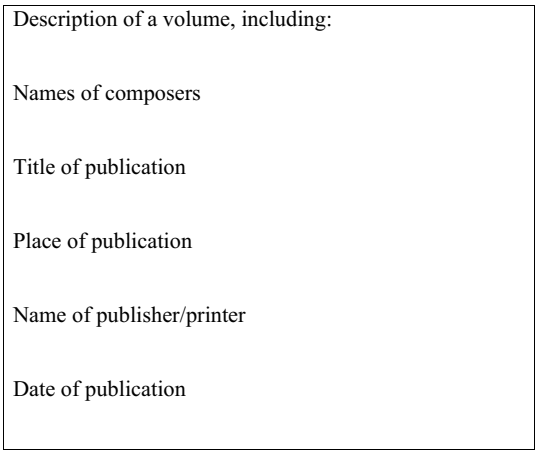


Fig. 1c. RISM B/I Record Structure

The British Library's printed music dataset (fig. 1d) also treats the individual publication as the unit of description and does not provide separate catalogue records for individual works in a volume of pieces. The records for single-work publications and collections of pieces by a single composer include the composer's name, where known. However, in single-composer collections the titles of individual works are rarely given.

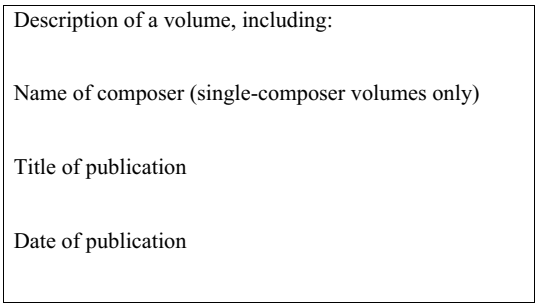


Fig. 1d. Record Structure for Printed Music in the British Library

Anthologies are more problematic. There is a catalogue record for each anthology, which gives the title of the anthology. However, until recently the contents of anthologies have not been enumerated, and the composers' names were not included in the catalogue record.

The *Early Music Online* project (fig. 1e) was the first attempt to provide a detailed inventory of the contents of anthology volumes.

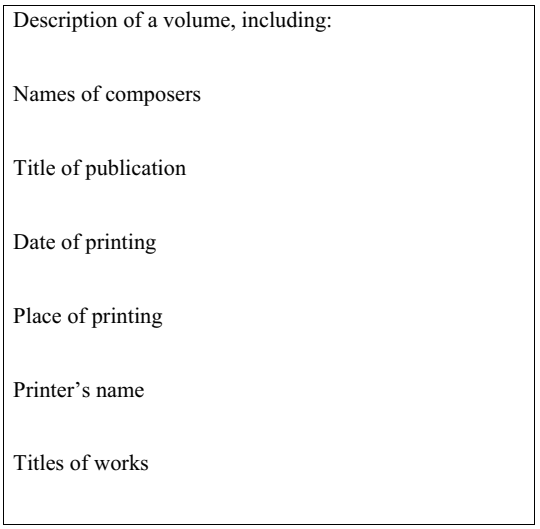


Fig. 1e. *Early Music Online* Record Structure

The BL's online catalogue of manuscripts and archives (fig. 1f) follows archival conventions and includes hierarchical descriptions of archives. Although there is usually a record for each volume of material, the contents are enumerated within this, rather than having their own records.

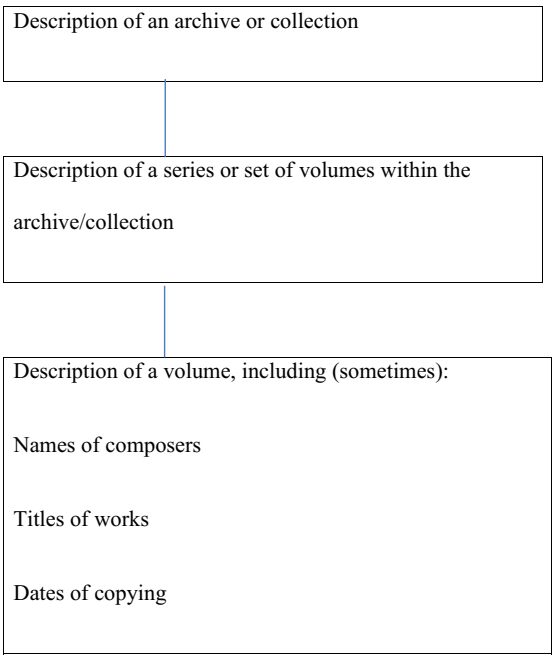


Fig. 1f. Record Structure for the British Library's Catalogue of Archives and Manuscripts

Hughes-Hughes (fig. 1g) includes detailed descriptions of each manuscript, in which the individual pieces are enumerated, and very rich indexes of people, and titles and first lines of works.

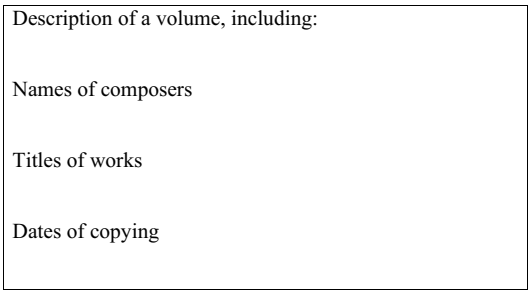


Fig. 1g. Information Contained Within Hughes-Hughes

The project team was faced not only with data heterogeneity, but also the fact that not all of the datasets contained the data elements that were of interest to the project team. We were particularly interested in the following categories of information, and in examining relationships between them:

PEOPLE e.g., composers, printers, former owners
PLACES e.g., countries, cities, holding libraries
DATES e.g., dates of publication or manuscript copying, composer life dates
WORKS e.g., work titles, genres

Table 1 shows the distribution of certain of these data elements across the datasets.

Material type	**Catalogue**	**Composer**	**Work title**	**Country of origin**	**City of origin**	**Creation date**	**Subject/ genre**	**Printer/ publisher**
Printed music	**RISM A/I**	Y	partial coverage	N	Y	partial coverage	Y	Y
Printed music	**RISM B/I**	Y	N	N	Y	Y	N	Y
Printed music	**Explore the BL (except *EMO* records)**	partial coverage	partial coverage	Y	Y	Y (many are conjectural and approximate)	partial coverage	partial coverage
Printed music	***EMO* records**	Y	Y	Y	Y	Y	Y	Y
MS music	**RISM A/II**	Y	Y	N	N	Y (mostly conjectural date ranges)	Y	N/A
MS music	**Explore Archives & MSS**	partial coverage	partial coverage	N	N	Y (mostly conjectural date ranges)	N	N/A
MS music	**Hughes-Hughes**	Y	Y	N	N	Y	Y	N/A

Table 1. Categories of Data Captured in the Music Datasets

One of the objectives of the project was to track the transmission of certain composers' music geographically and temporally. As Table 1 shows, some of the datasets were more suitable for this than others. RISM A/I, B/I, A/II, and Hughes-Hughes were deemed to be good sources for composers' names, but the BL online catalogues less so, because of the number of composers whose music is 'hidden' in printed and manuscript anthologies.

Critically, none of the manuscript catalogues captures information on the place of copying. While it is often not possible to pinpoint the city of origin of a manuscript, the country of origin can often be identified, so the absence of this information from all the manuscript datasets was a problem. While the RISM A/II records include the current location of the manuscript, this cannot generally be taken as a proxy for the place of origin of the manuscript, because much material has been carried overseas. Including country of origin in manuscript descriptions is perhaps something that could be considered by cataloguing organisations.

With the catalogues of printed music, geographically-focused research was more feasible: RISM A/I, B/I, and the BL catalogue all generally include city of publication where this information is provided on the publication or can be ascertained. The BL catalogue also includes country of origin, but this was found to be less useful given the shifting nature of national borders.

For an analysis of the transmission of composers' works over time, it was necessary to have access to creation dates for the material, either publication or printing dates for the published material or copying dates for the manuscript sources. While all the catalogues include creation dates of sorts, these had to be treated cautiously, as many were conjectural. Manuscript sources cannot often be dated precisely, so a conjectural date range is generally provided in the RISM A/II database, for example '17th century'.

From the eighteenth century onwards, printers rarely included the date of publication on the item, and so for material bearing no date, a conjectural date of publication has been included in the BL catalogue. Such dates are usually rounded, for example up to the nearest decade. An unwary researcher could be misled by an apparent spike in the number of publications issued at the beginning of each decade. Figure 2 shows how data from the British Library catalogue could give the impression that there were sudden spikes in the publication of Handel's music in the eighteenth century. Today, some of the conjectural dates may be challenged by musicologists and bibliographers who have access to information that was not available to the nineteenth-century cataloguer. In contrast, RISM A/I cataloguers generally included dates of publication only when these were printed on the edition. Thus a large proportion of the post-1700 catalogue records in RISM A/I include no date of publication.

Even where the same data elements are included in several datasets, there are differences in the way the information has been recorded. While names are authority-controlled in the BL catalogue and in the RISM datasets, different authority files have been used. (The BL uses the NACO/Library of Congress authority data, while RISM has its own set of authority files.)

As well as heterogeneity between the datasets, there is also heterogeneity within the dataset in the case of the BL. Because of the long history of the BL and the changes to library cataloguing rules and standards that have occurred over the years, the catalogue records for printed music vary in the level of detail they contain. Most records were created at the time that the particular scores were acquired, using the cataloguing standards of the time and under the constraints of the day. The Library of the British Museum (the

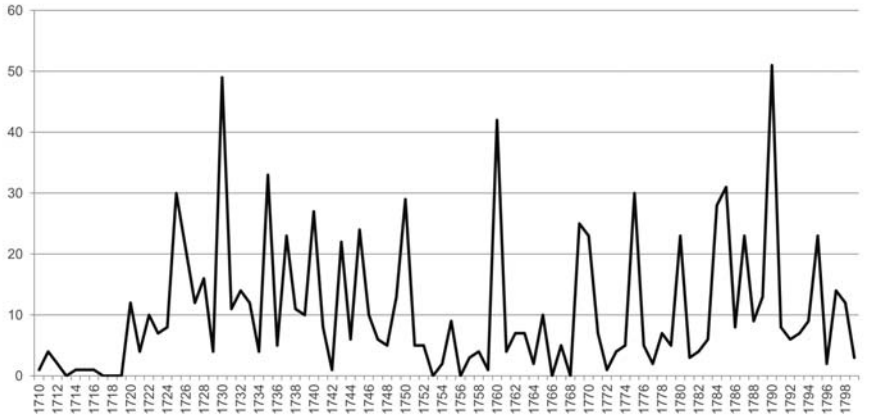


Fig. 2. Graph Showing Distribution of Publication Dates for Handel's Music in the British Library Catalogue, 1710–1800

predecessor of the British Library) was founded in the eighteenth century, and began acquiring music immediately. Its first music cataloguer, Thomas Oliphant, was not appointed until 1841, but by 1849 he had apparently catalogued all the printed music by then acquired by the Library. Legal deposit laws required British and Irish music publishers to send one copy of every publication to the British Museum Library, and this, given the burgeoning of the music publishing industry in the nineteenth century, left Oliphant's successors struggling to keep up with the ever-increasing amount of printed music being acquired. They did manage to catalogue all the musical repertory they deemed to be 'serious', but much of the lighter, popular music was not catalogued at all.[14]

Popular pieces aside, then, most of the music in the British Library that was published in the nineteenth century (and a considerable amount of the earlier material) was acquired and catalogued in the nineteenth century. As well as bearing in mind that much of the bibliographic data was created a long time ago, it was important to recall the purposes for which it was gathered, namely to enable library users to discover whether a particular piece of music was held by the Library, and to see what else by that composer was also held. Only information deemed essential for these purposes was included, and it was necessarily brief, to enable it to be written onto small catalogue slips. Many of the British Library's catalogue records created in the nineteenth century do not therefore include the name of the publisher, and sometimes the place of publication was omitted too, presumably to save time.

By contrast, records created since the 1980s conform to modern cataloguing standards. They are usually very detailed (including information, for instance, about the language of the text of vocal works), and draw terms from controlled vocabularies, such as the Library of Congress Subject Headings and NACO name authority file.

With a dataset as long-lived and large as the BL's, there are inevitably errors in the way some of the data has been captured. With the advent of machine-readable cataloguing,

14. Brief records for the popular vocal music were created in the late-twentieth century. Much of the popular instrumental music is yet to be catalogued (see the article by Christopher Scobie, 'Ephemeral Music? – The "Secondary Music" Collection at the British Library', *Fontes Artis Musicae* 63, no. 1 [January–March 2016]: 21–32).

the data has had to be marked up to show to which field it belongs. The BL uses MARC21 for this. Encoding and migration errors can affect the way the data has been assigned to fields. While to a user of the online catalogue, these encoding errors might not always be problematic, they can have an effect on the accuracy of data analysis.

Data Enhancement and Preparation

With the data being so heterogeneous, it was recognised that a substantial amount of data preparation, alignment, and cleaning would be needed before any data analysis could take place. For the BL printed music data, the project team decided to undertake some basic improvements to the whole dataset, and in parallel to enrich selected records, focussing on music printed in the sixteenth century, in a continuation of the work begun during the *EMO* project. As with *EMO*, it was decided to concentrate on enriching records for sixteenth-century anthologies, rather than single-composer publications, adding to the catalogue record the names of all the composers who contributed to a volume of music, and the titles of all of the pieces. The British Library utilises the Resource Description and Access (RDA) standard, and MARC21, so all data enhancement work had to take place within this framework, in a database in which a single record represents the whole publication. Lists of the pieces in each volume were added to the contents note field, where the composers' names and work titles were recorded as spelled on the item in hand. For *EMO*, authority controlled names had also been included; for 'A Big Data History of Music' we went a step further and added authority controlled work titles as well. Where the work appeared in the NACO/Library of Congress Name-Title authority file, this was used, but in practice it was necessary to construct a local list of titles to supplement this. Also recorded in controlled form were names of dedicatees and former owners.

In the sixteenth century, music printing began to enable music to be reproduced in greater quantities than before, and individual composers' works could be disseminated far more widely. Printers' names and places of printing are therefore of interest to researchers of printing history. However, they can be difficult to track in library catalogues. Traditionally, library cataloguing rules have required the publication details to be recorded exactly as on the item. However, on the title page or colophon, these details may be given in a variety of ways, and often in a Latin form. The date of publication may be in Roman numerals, or in words. To enable the publications of a particular printer or locality to be tracked, we went beyond what is required by traditional cataloguing rules and also created authority controlled access points for printers' names and places of printing.

A team of research assistants (mainly Ph.D. students with an interest in this repertoire) created lists of contents of each volume, checked composer attributions, and described the physical characteristics of the volumes. Rather than teaching the students RDA and MARC21, we asked them to enter their data in a template in Microsoft Word. Experienced cataloguers then checked the information and added it to the BL's library management system (Ex Libris' Aleph), contributing the authority-controlled data, and adding the required encoded information in MARC21. By the end of this data enhancement phase, all the BL's sixteenth-century anthologies of printed music had been re-catalogued to modern standards.

The research assistants undertook some data cleaning and reformatting work on the electronic version of the Hughes-Hughes catalogue data. OCR errors were corrected and the largely unstructured data added to spreadsheets for analysis. Further data cleaning work on the printed music dataset was undertaken by the British Library's Metadata

Services team, who have developed routines for making global changes to the data. They corrected historic MARC encoding errors and added MARC country codes to most of the records, deriving these automatically from the city of publication. The cleaned data was returned to the master catalogue, so all catalogue users will benefit from the data improvements. Metadata Services then provided the project team with a download of the complete printed music dataset in MARC21. An export of the 16,000 music-related records from the catalogue of archives and manuscripts was also obtained.

At the time of our research, the complete RISM A/II dataset could be freely downloaded from the RISM Web site in MARCXML format, but the RISM A/I data was publicly available only on CD-ROM or in hard copy. The RISM Central Office kindly provided us with an electronic copy of the complete RISM A/I dataset in MARCXML format in advance of its release on the RISM Web site in 2015.[15] To analyse the data in RISM B/I, which is not available electronically in its entirety, we produced our own spreadsheet of names, places, and dates drawn from the printed volume.

We wished to transfer the data supplied in MARC21 and MARCXML into a form in which it could be analysed more easily, and chose to convert it to tab delimited text files which could then be opened in Excel. For the data conversion we used the utility MarcEdit, created by Terry Reese (http://marcedit.reeset.net/ [accessed 11 March 2016]). MarcEdit allows the user to select particular fields and subfields for export, so we chose to migrate only those fields deemed most relevant to our research. Figures 3, 4, and 5 provide screenshots of the process of data migration.

Fig. 3. Screenshot of MarcEdit Showing the MARC Field Selection Screen

With each of our chosen datasets now available in spreadsheets, we then reviewed the data and planned our research strategy. It was already clear, from our analysis of the kinds of data stored in each dataset, that certain types of analysis, for example, of the geographical movement of music, would be possible only with certain datasets. We also wanted to minimize our handling of unwieldy datasets of a million or more records. While maintaining copies of the full datasets for releasing to other researchers, we produced 'slices' or subsets of the data for use in our research. From the British Library data, we chose to focus on pre-twentieth-century publications; from RISM A/I and B/I we concentrated on pre-1700 editions, as these were dated and provided us with the potential for examining publication trends over time.

15. We are grateful to the RISM Central Office for providing us with this data and for giving us permission to utilise the data in our research.

FMT	MU
LDR	cm a22001933 4500
001	004763607
008	850515s1783 ne ||| ||||||||| lin |
040	|a Uk |c Uk
1001	|a Clementi, Muzio, |d 1752-1832.
24010	|a Sonatas, |m piano, |n op.9
24500	|a Trois sonates pour le forte piano ou le clavecin / |c composées par Muzio Clementi. Oeuvre IX.
260	|a À Amsterdam : |b [s.n.], |c [1783?]
300	|a 38p. ; |c fol.
336	|a notated music |2 rdacontent
337	|a unmediated |2 rdamedia
338	|a volume |2 rdacarrier
85241	|a British Library |b MUSIC |j h.319.o.(2.)
SYS	004763607

Fig. 4. Example of a MARC Record Prior to Migration

206	Birch, W. H. (William Henry), 1826-1888.				Grand triumphal march [for pianoforte], etc.
207	Birch, W. H. (William Henry), 1826-1888.				W. H. Birch's Vocal Miscellany, a selection of choruses, glees, quartets, trios, etc.
208	Hurlebusch, Conrad Friedrich.				De 150 Psalmen Davids, met de zelver Lofgezangen, gemaakt voor het clavier en orgel, na hunne gegro
209	Hurlebusch, Conrad Friedrich.				De 150 Psalmen Davids, met der zelver Lofgezangen, gemaakt voor het clavier en orgel, na hunne gegro
210					[Sji'r segala Mazmûr-Da-ûd, dan pûdji-an jang lajin, etc. (Ta'limu-'ldini-'lmesêhiji, ija itu, pang' adjaran '
211	Haym, Nicola Francesco, 1678-1729.	Dodeci sonate à trè, cioè due violini,			<Violino primo. - Violino secondo.> [Parts.]
212	Mozart, Wolfgang Amadeus, 1756-1791.	Zauberflöte.			<Zweyter Theil.>.
213	Mozart, Wolfgang Amadeus, 1756-1791.	Zauberflöte.			Die Zauberflöte. <Zweyter Theil>.
214	Mozart, Wolfgang Amadeus, 1756-1791.	Zauberflöte.			Ouverture fürs Clavier oder Piano forte.
215	Locatelli, Pietro Antonio, 1695-1764.	Concertos,		op.1.	[XII concerti grossi a quattro, e a cinque ... opera prima ...].
216	Clementi, Muzio, 1752-1832.	Sonatas,	piano,	op.9	Trois sonates pour le forte piano ou le clavecin / composées par Muzio Clementi. Oeuvre IX.
217	Kywtok, B. I. W.				God save the King in VIII easy variations for the piano f composed by B.I.W. Kywtok.
218	Scaccia, Angelo Maria.				Concerti con Violino Obligato, due Violini, Alto Viola e Basso Continuo. Opera Prima. [Parts.]
219	Valentini, Giuseppe.				x concerti a violino primo, secondo e basso concertino, e violini primo e secondo concerto grosso, viole
220	Mossi, Giovanni.				Concerto xiio a Quatro Violini e Violoncello Obligati. [Parts.]
221	Jozzi, Giuseppe, approximately 1710-approximately 1770.				VIII sonate per cembalo, ppera prima / da Giuseppe Jozzi.
222	Simono, D.				Recueil Nouveau d'Airs, Menuets, Contredanses, Gavottes & Gigues, de Differens Autheurs, Italiens, Fr
223	Schmid, J. B.				vi. Sonates de Clavecin, avec Accompagnement ad libitum de deux Violons et Basse Continue, etc. [Par
224	Collasse, Pascal, 1649-1709.	Thetis et Pelée			Les Airs de la Tragedie de Thetis et Pelée.
225	Hol, Rijk, 1825-1904.				Ontwaking. Lied voor eene Zangstem met Klavierbegeleiding. Woorden van J. A. Bakker. Op. 78.
226	Schuurman, F. H.				Aan het Vaderland. Gedicht van L. Uhland, vertaald door B. van Meurs. Lied voor Tenor met Piano. Op. 2
227					De cl Psalmen Des Propheten Davids ... overgeset Door Petrum Dathenum: En ... op eenen Sleutel geste

Sheet1

Ready

Fig. 5. The Same Data After Migration to Tab Delimited Form

Once we had created our subsets of data, we then undertook some further data cleaning and enhancement work on these spreadsheets. For example, some dates of publication were qualified with a question mark or were given in square brackets to indicate that the date had been taken from a source other than the publication itself. Such annotations meant that the dates could not be sorted properly, so we created a new column for 'normalised' date. This involved assigning a simple four-digit date to the record. It could be argued that this introduces a level of certainty not present in the original data, but without a date assigned to the record it could not be included in the analysis. To aid a broad-brush analysis of publication trends we also assigned each record to a decade.

In a further column we created a normalised version of the city of publication, to bring together all publications issued in a single location. We then added a column for geographic co-ordinates to aid in the presentation of the data on a map. These codes for longitude and latitude were taken from the CERL thesaurus (http://thesaurus.cerl.org

/cgi-bin/search.pl [accessed 11 March 2016]), which proved a particularly useful resource for distinguishing between different printers of the same name, as much work has been done by CERL contributors to disambiguate individuals.

Research Findings

Before creating our data subsets we performed some simple analyses on the full datasets, the results from which highlight the large numbers of now-forgotten composers active in all centuries since the advent of music printing, and the vast growth in music publishing between then and the present. More than nine-thousand composers feature in the RISM A/I dataset (predominantly active before 1800). Even more composers' music was circulated in manuscript in that period: over 28,000 composers feature in the RISM A/II data. The British Library's holdings of twentieth-century publications include works of more than 100,000 composers. Strikingly, more than half of these composers have just one publication to their name; the majority could be classed, following Moretti, as the 'Great Unheard'.

Table 2 lists the composers whose names appear most frequently in RISM A/I. Here it must be cautioned that counting totals of publications is a crude measure that does not distinguish between large collected volumes and single-sheet songs; inevitably Table 2 is dominated by eighteenth-century musicians whose compositions were generally published as single items of sheet music. Despite this caveat, Table 2 is useful for showing that eighteenth-century popularity measured in quantitative terms did not secure a composer a place in subsequent canons of music history. Leaving aside the anonymous publications, some perhaps surprising names appear near the top of the table, most notably those of Daniel Gottlieb Steibelt, and of Ignace Pleyel, for whom RISM lists even more editions than for Joseph Haydn. Some composers who are considered today to be part of the eighteenth-century canon are notable by their absence from the top fifty, for example Johann Sebastian Bach, at number ninety-five on the list with 141 entries, and Domenico Scarlatti, at number 440 with thirty-one entries.

Restricting the dataset to publications issued in Britain and Ireland gives a somewhat different top twenty (see Table 3). Handel is now predominant (unsurprising, given his role in English musical culture), and Mozart less prominent, while the prolific Charles Dibdin and James Hook have overtaken even Haydn. A seventeenth-century composer, Henry Purcell, now appears in the ranks of the most frequently published, hinting at the English taste for certain types of 'ancient music' in the late-eighteenth century.[16] Although this listing does not include the sheet music that was imported from overseas, it still gives a sense of the relative importance of native versus continental composers in the British Isles in the eighteenth century.

A similar listing of the most prolific composers in RISM A/I was offered by Donald Krummel in his 1992 overview of music bibliography, in a short vignette on 'the magnitude of the output of printed music since Gutenberg's day'.[17] Krummel's listing was made by manually counting entries in the initial series of RISM A/I volumes, before the volumes of addenda were published, and hence his totals are somewhat smaller than ours. For him,

16. Sandra. Tuppen, 'Purcell in the 18th Century: Music for the "Quality, Gentry, and Others" ', *Early Music* 43, no. 2 (2015): 233–45.

17. Donald W. Krummel, *The Literature of Music Bibliography: An Account of the Writings on the History of Music Printing and Publishing* (Berkeley: Fallen Leaf Press, 1992), 67–70.

Mozart, Wolfgang Amadeus	3,988
Anonymous	3,705
Pleyel, Ignace	2,534
Haydn, Joseph	2,476
Händel, Georg Friedrich	1,711
Steibelt, Daniel Gottlieb	1,173
Hook, James	1,165
Dibdin, Charles	847
Grétry, André-Ernest-Modeste	822
Dalayrac, Nicolas-Marie	789
Vaňhal, Jan Křtitel	772
Dusík, Jan Ladislav	758
Jelínek, Josef	723
Arne, Thomas Augustine	704
Paisiello, Giovanni	639
Shield, William	608
Clementi, Muzio	561
Méhul, Etienne Nicolas	512
Hoffmeister, Franz Anton	512

Table 2. Most Frequently Occurring Composers in RISM A/I

Händel, Georg Friedrich	1,416
Anonymous	951
Hook, James	871
Dibdin, Charles	769
Pleyel, Ignace	624
Haydn, Joseph	616
Arne, Thomas Augustine	492
Mozart, Wolfgang Amadeus	469
Shield, William	442
Arnold, Samuel	309
Storace, Stephen	237
Giordani, Tommaso	229
Reeve, William	222
Mazzinghi, Joseph	184
Dusík, Jan Ladislav	183
Bach, Johann Christian	170
Callcott, John Wall	165
Purcell, Henry	164
Paisiello, Giovanni	156
Koželuh, Leopold	153

Table 3. Most Frequently Occurring Composers in British Publications in RISM A/I

the dominance of eighteenth-century names on this listing illustrated the reorientation of music publishing around 1700, away from substantial volumes and towards sheet music containing a single composition or a few pieces. But Krummel also noted some of the methodological difficulties that such a quantitative approach faced, including the question of what should be counted (anthologies or individual compositions within them?), the difficulty of estimating lost publications, and the problem posed by the fact that most music publications after 1700 are undated.[18]

We address the same methodological challenges in an article in the November 2015 issue of *Early Music*, which presents more substantial examples of research done by the project.[19] Given the difficulties of dating printed music after 1700, many of the analyses in this article focus on music of the sixteenth and seventeenth centuries. Using the RISM

18. Ibid., 67–68.

19. Stephen Rose, Sandra Tuppen and Loukia Drosopoulou, 'Writing a Big Data History of Music', *Early Music* 43, no. 4 (2015): 649–60.

A/I and B/I datasets for printed music between 1500 and 1700, this article analyses the rise and fall of music printing across Europe in the sixteenth and seventeenth centuries. The analysis shows how a plateau or decline in the European output of printed music might be caused by economic or political disruption in a specific city—for instance, plague in Venice in 1576–1577 and again in 1631 had a devastating impact on the continent's production of printed music. This case-study also quantifies the decline of European music-printing from the 1620s onwards, and its geographical reconfiguration (with the dominance of Venice being replaced by the rise of northern printing centres such as London and Paris). Further case-studies in the *Early Music* article examine Palestrina and Purcell, showing how bibliographical data can confirm hypotheses and explore new research questions about the dissemination of a composer's music in print or manuscript. Finally, the article shows how bibliographical data can be used to show trends in the publication of music that advertises itself as having an ethnic or national flavour—in this case, the rise of 'Scottish' music in the British Library catalogue between 1700 and 1900.

Data Visualisation as an Aid to Research

Much of the research for 'A Big Data History of Music' was undertaken using the data manipulation, filtering, and counting tools available in Microsoft Excel. With the very large datasets we also found it beneficial to produce visualisations of some of the data. Data visualisation can both illustrate research findings (for example in graphs and pie charts) and aid the research process itself, by enabling the researcher to see patterns and trends not always visible in the raw data. Having identified potential points of interest at the macroscopic level from the visualisation, the researcher can then drill down to examine the data at the microscopic level and seek explanations for a particular trend. Many data visualisation tools, both free and commercial, are available, and it is beyond the scope of this article to offer a full evaluation of these. Three free tools which can be used by researchers without programming skills are Google Fusion Tables (http://tables.googlelabs.com), OpenHeatMap (www.openheatmap.com), and Palladio (http://palladio.designhumanities.org [all accessed 11 March 2016]).

Google Fusion Tables allow data from a spreadsheet to be imported and then a range of visualisations to be created quickly and simply. Network graphs, in which relationships between data can be displayed, are a useful feature, and it is also possible to visualise data with a geographic element on a map. (An advantage of this tool is that the geographic coordinates can be generated automatically.) Figure 6 connects the most frequent genre descriptors in RISM A/I with related composers; this network graph shows the numerical dominance of vocal genres such as psalms, sacred songs, and occasional music (for weddings and funerals) in the surviving corpus of printed music before 1800.

OpenHeatMap allows maps to be produced showing the relative levels of density of a specific category of data. Animations can also be produced, showing changes over time. As with Google Fusion Tables, the data can be imported from a simple spreadsheet. Figure 7 is a sample of how such heat-maps can chart the relative intensity of music printing across Europe in the 1660s; this map demonstrates the decentralised nature of music printing in German-speaking lands, with relatively many firms each producing small amounts of music, whereas in France and Italy the industry was focused on Paris and the three Italian centres of Bologna, Rome, and Venice respectively. A dynamic heat-map of European music-printing before 1730, using RISM A/I and B/I data within the Dariah geo-

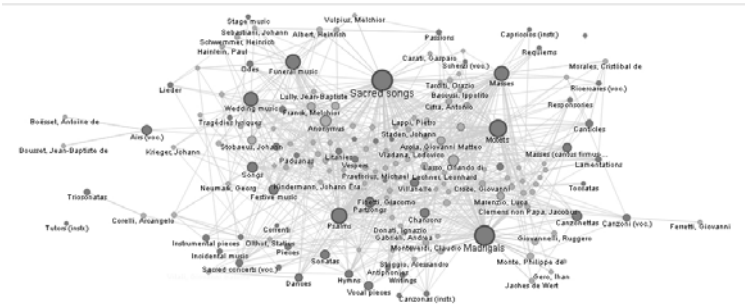


Fig. 6. Most Common Genre Terms in RISM A/I and Related Composers, in a Network Graph Created Using Google Fusion Tables

Fig. 7. The Distribution of Music Printing in Europe in the 1660s, from RISM A/I and B/I, in OpenHeatMap

browser (http://geobrowser.de.dariah.eu/ [accessed 11 March 2016]), has been created by Laurent Pugin of RISM Switzerland.[20]

The visualisation tool Palladio requires slightly more effort in terms of data preparation, including the addition of geographic co-ordinates in order to view map visualisations, but includes more sophisticated filters and timelines which can illuminate research on specific facets of the data. Figure 8, showing the genres and places of Palestrina's published music, is useful for clarifying publishing patterns: for instance, only a limited number of Palestrina's sacred genres (such as Masses, hymns, and sacred songs) were published in the counter-Reformation centres of Antwerp and Milan, whereas a greater cross-section of genres (including his madrigals) appeared in Rome and Venice.

20. http://docs.rism-ch.org/visualisation/, accessed 11 March 2016.

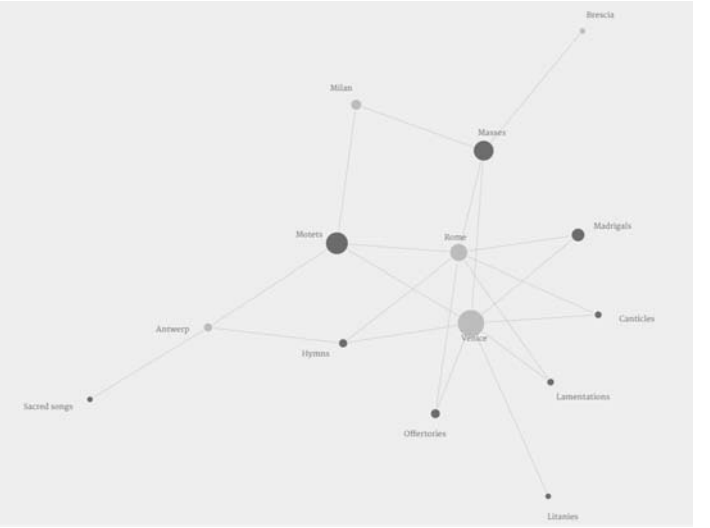


Fig. 8. Palestrina: Genres and Places from RISM A/I Data, in Palladio

Scholarly Prospects

Bibliographical data from research libraries is of immense value for researchers in digital musicology, and can also help libraries reach new scholarly audiences, as the following examples show. During the project 'A Big Data History of Music', a strong public interest in data analysis was demonstrated by a Data Exploration Day held at the British Library on 10 March 2015, attended by around twenty delegates (including academics in a range of disciplines and non-academics such as curators, publishers and librarians). The delegates experimented with analysing and manipulating the datasets made available by the Big Data History of Music project; attendees less experienced in data analysis were given individual guidance plus a tutorial on the data-cleaning tool OpenRefine.

From the visualisations produced at the day, we publish here an example by Andrew Gustar, a specialist in the application of statistics to musicology (fig. 9).[21] His two graphs seek to investigate patterns in the publication history of items in the BL catalogue, specifically whether works containing 'piano' in the title had a shorter publication history than works without the word 'piano'. (The 'piano' works hence include piano concertos and chamber works with piano, as well as solo piano repertory.) He began the analysis in Excel, extracting catalogue records that included a publication date and a composer's birth date, and ignoring composers born before 1700 or after 1930. This gave a total of 363,914 records (93,880 records with the word 'piano' and 269,266 other records), which he loaded into the statistical software 'R'.[22] The charts were created using the R package 'ggplot2',[23] and are scatterplots of each of the piano and non-piano subsets by the composer's date of birth and year of publication. The contour lines and overall depth of shad-

21. Andrew Gustar, 'Statistics in Historical Musicology' (Ph.D. thesis, Open University, 2014).

22. R: A Language and Environment for Statistical Computing. https://www.R-project.org/, accessed 11 March 2016.

23. Hadley Wickham, *ggplot2. Elegant Graphics for Data Analysis* (New York: Springer, 2009).

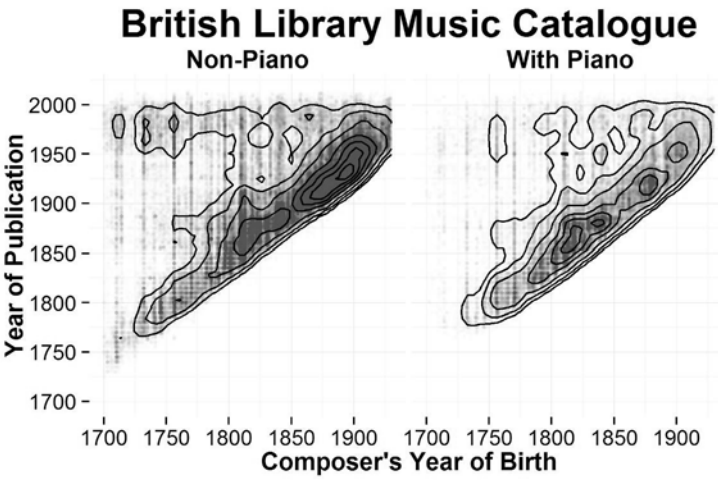


Fig. 9. Andrew Gustar's Visualisation of Publications With and Without Piano in the British Library

ing represent the density of points. The vertical lines represent the composers whose music continued to be published over the decades or even centuries after their lifetime: Haydn (b. 1732), Mozart (b. 1756), Beethoven (b. 1770), and Mendelssohn (b. 1809) can all be discerned. The graphs show that most of the BL's printed music comprises editions published during the composer's lifetime (that is, near the diagonal line at the base of the shaded area), although there was a resurgence of publications of earlier works in the second half of the twentieth century, especially pre-1800 works not for piano, which may reflect the increased interest since 1950 in early music and in scholarly editions. The 'piano' publications peak in the third quarter of the nineteenth century, presumably reflecting the popularity of piano arrangements and the growth of the domestic piano repertory in this era. Gustar stresses that this visualisation was produced during the Data Exploration Day, and with more time, a better data selection process could be used to identity works for solo piano (as opposed to chamber and orchestral works that include piano within their scoring). Despite these cautions, his visualisation shows the interest that library bibliographical data holds for researchers and also for practitioners of citizen science.

Scholars may choose to link library metadata with other datasets, or use it as the basis of a new database. An example of this approach is shown by the Semantic Linking of Information, Content and Metadata for Early Music project (SLICKMEM), carried out as part of the Transforming Musicology research project led by Tim Crawford of Goldsmiths' College, London. As a pilot exercise in the aligning and linking of data, SLICKMEM combined *Early Music Online*'s metadata with data from another digital library, the Electronic Corpus of Lute Music (www.ecolm.org [accessed 11 March 2016], covering a very similar sixteenth-century repertory). Links were created to external data sources such as the Virtual International Authority File (http://viaf.org), MusicBrainz (https://musicbrainz .org/), and DBPedia (http://wiki.dbpedia.org [all accessed 11 March 2016]). To explore the methodological ramifications of making such links, this work was done partly via automated processes, partly via human intervention; place names proved

much easier to align than personal names.[24] The resultant database can be queried via http://slickmem.data.t-mus.org/snorql (accessed 11 March 2016), where the *Early Music Online* metadata can be searched in a variety of ways, for instance to determine the most frequently occurring titles of compositions in the corpus. After discounting genre names such as 'Fantaisie' and liturgical titles such as the Benedictus of the Mass, the most frequently occurring titles in SLICKMEM are two vocal compositions by Jacques Arcadelt ('Il ciel che rado virtu' and 'Occhi miei lassi') that often appeared in instrumental tablatures. By identifying such links within the metadata, SLICKMEM can guide musicologists wishing to examine the notated music for thematic similarities or compositional borrowings, for instance scholars seeking to detect the process of *imitatio* whereby sixteenth-century composers modelled their works on authoritative musical models. SLICKMEM shows how musicologists with computer programming skills can use library metadata to develop new methods for research.

Accessing the Open Datasets

Most of the data utilized in 'A Big Data History of Music' is available as open data and may be freely downloaded for re-use. After combining RISM A/I, A/II, and part of B/I into a single dataset, the RISM Central Office released the dataset as open data in MARCXML and RDF/XML in 2015.[25] As part of the 'Big Data History of Music' project, the BL released its catalogue of printed music as open data, together with data from the Hughes-Hughes catalogue.[26] Both BL datasets are available as simple comma-separated text files (here described as 'Researcher Format'), and the printed music data is additionally available in RDF/XML. The Researcher Format provides several different 'views' of the printed music dataset, enabling users to examine the data from different perspectives, including by name and title. The Hughes-Hughes data is a set of more than 35,000 titles and first lines of individual compositions within manuscripts, with details of genre and composer where known.

Conclusion

The project 'A Big Data History of Music' has shown how the role of the academic music librarian can extend into new digital domains. Metadata that was originally created as a finding and bibliographical tool can now be harnessed as material for research in its own right. Libraries may need to invest time and money in data cleaning before they can make their metadata publicly available, but this cleaning can benefit all catalogue users, not just data analysts. However, the methodologies of big data research are making scholars aware of the need to clean and align heterogeneous data that was originally assembled for other purposes. Libraries should also educate potential users about the histories and idiosyncrasies of their catalogues, so that scholars are aware of the limitations and provenance of the data.

24. Tim Crawford, Ben Fields, David Lewis and Kevin Page. 'Explorations in Linked Data Practice for Early Music Corpora', *Proceedings of 2014 IEEE/ACM Joint Conference on Digital Libraries*, 309–12, DOI: 10.1109/JCDL.2014.6970184.

25. https://opac.rism.info/index.php?id=8&L=1, accessed 11 March 2016. The Resource Description Framework (RDF, http://www.w3.org/RDF/ [accessed 11 March 2016]) is a standard model for data interchange on the Web.

26. http://www.bl.uk/bibliographic/download.html, accessed 11 March 2016.

Scholars interested in using the data are likely to have widely varying levels of IT competence, from basic computer literacy to advanced programming skills. If a release of open bibliographical data is to have its maximum impact on the community of researchers, it may be advisable to make it available in simple delimited text files that can be opened in proprietary programmes such as Microsoft Excel, as well as in RDF/XML. However, if these obstacles can be overcome, valuable resources for research can be created. The cataloguing work undertaken over decades by music librarians can find new uses, because the data thus accumulated can allow musicologists to test hypotheses and discern previously unnoticed trends. Library catalogue data can potentially enable musicologists to transform the writing of music history.

English Abstract
Librarians and archivists are increasingly collecting and working with large quantities of digital data. In science, business, and now the humanities, the production and analysis of vast amounts of data (so-called 'big data research') have become fundamental activities. This article introduces the project 'A Big Data History of Music', a collaboration between Royal Holloway, University of London and the British Library. The project has made the British Library's catalogue records for printed and manuscript music available as open data, and has explored how the analysis and visualisation of huge numbers of bibliographic records can open new perspectives for researchers into music history. In addition to the British Library data (over a million records), the project drew on a further million bibliographic descriptions from RISM, which have also recently been released as open data. To show the challenges posed by the heterogeneous nature of the data, the article outlines the different structures of the various catalogue records used in the project, and summarises how the British Library data was cleaned and enhanced prior to its public release. Examples are given of how music-bibliographical data can be analysed and visualised, and how scholars and citizen scientists can engage with this data through hackathons, large-scale data analyses, and database construction. It is hoped this article will encourage other research libraries to consider making their catalogue records available as open data.

French Abstract
Les bibliothécaires et les archivistes sont de plus en plus amenés à recueillir et à manipuler de grandes quantités de données numériques. Dans le domaine des sciences, des affaires et maintenant des sciences humaines, la production et l'analyse de grandes quantités de données sont devenues des activités fondamentales. Cet article présente le projet « A Big Data History of Music », une collaboration entre Royal Holloway, University of London et la British Library. Le projet a consisté à rendre disponibles sous forme de données ouvertes les notices catalographiques de la British Library pour la musique imprimée et manuscrite, et à explorer comment l'analyse et la visualisation d'un très grand nombre de notices bibliographiques peuvent ouvrir de nouvelles perspectives pour les chercheurs en histoire de la musique. En plus des données de la British Library (plus d'un million de notices), le projet a aussi tiré parti d'un autre million de descriptions bibliographiques venant du RISM, qui ont récemment été diffusées sous forme de données ouvertes. Pour montrer les défis posés par la nature hétérogène des données, l'article décrit les différentes structures des notices catalographiques utilisées dans le projet et résume la façon dont les données de la British Library ont été nettoyées et améliorées avant leur diffusion. Il comporte aussi des exemples de la façon dont les données bibliographiques relatives à la musique peuvent être analysées et visualisées, et comment les chercheurs et les scientifiques citoyens peuvent traiter ces données au moyen de marathons de programmation, d'analyses de données à grande échelle et de la construction de bases de données. Il est à espérer que cet article encouragera d'autres bibliothèques de recherche à envisager de rendre disponibles leurs notices catalographiques sous forme de données ouvertes.

German Abstract
Bibliothekare und Archivare sammeln immer größere Mengen digitaler Daten, mit denen sie arbeiten. In den Naturwissenschaften, in der Wirtschaft und nun auch in den Geisteswissenschaften ist die Produktion und Analyse großer Datenmengen ("Big Data") zu einer grundlegenden Tätigkeit geworden. In diesem Aufsatz wird das Gemeinschaftsprojekt "*A Big Data History of Music*" von *Royal Holloway, University of London* und der *British Library* vorgestellt. Im Rahmen des Projektes wurden die Katalogdaten für handschriftliche und gedruckte Noten der *British Library* als *open data* verfügbar gemacht und erforscht, wie die Analyse und optische Aufbereitung einer immensen Menge bibliografischer Daten den Forschern neue Perspektiven auf die Musikgeschichte eröffnen. Zusätzlich zu den mehr als eine Million Datensätzen der *British Library* benutzte das Projekt eine weitere Million bibliografischer Beschreibungen aus *RISM*, die kürzlich als *open data* zur Verfügung gestellt wurden. Im Artikel werden die unterschiedlichen Strukturen der Datensätze aus den verschiedenen Katalogen beschrieben, um die daraus resultierenden Herausforderungen aufzuzeigen. Außerdem wird zusammengefasst dargestellt, wie die Daten der *British Library* vor der Veröffentlichung bereinigt und angereichert wurden. Beispiele illustrieren, wie musikbibliografische Daten analysiert und visualisiert werden können und wie Studenten und Wissenschaftler durch die Arbeit in Hackathons, mit großformatiger Datenanalyse sowie mit Datenbankentwicklung aus diesen Daten Nutzen ziehen können. Vorbildhaft soll dieser Aufsatz andere wissenschaftliche Bibliotheken ermuntern, ihre Katalogdaten als *open data* zur Verfügung zu stellen.