Are we ready for a big data history of music?

Athens, 22 June 2023
Frans Wiering, Utrecht University
f.wiering@uu.nl
Are we ready for a big data history of music?
Big data history of music

• expression coined by Stephen Rose and Sandra Tuppen (2014)
  • study of bibliographic data, mainly

• approach in key phrases
  • analyzing and visualizing bibliographic data
  • quantitative analysis
  • examine developments and trends
  • use visualizations to test and develop hypotheses

ABSTRACT

This position paper sets out the possibility of a metric based on the analysis of musical bibliographic 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 bibliographic metadata. After discussing some of the technical challenges of data alignment, it suggests how analysis and visualisation of this data might transform musicological understandings of cultural transmission and canonic 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 spheres; scholars focus on a handful of composers whose works are considered to be the best or most important. This bias is evident in the online realm as well: websites such as www.greatcomposers.com are primarily orientated to individual composers, whereas the rich bibliographies of Western musical culture could equally well be interrogated via an analysis of locations, publishers, performers, genres, social practices or migratory patterns.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires specific permission and/or a fee. Request permissions from Permissions@acm.org

2. RESEARCH CONTEXT

Since 2000 literary studies have been revolutionised by the opportunities offered by Big Data. Franco Moretti has pioneered the concept of 'distant reading', where quantitative analyses offer new perspectives on literary production and taste (as opposed to 'close reading' that focuses on a single text). By analysing 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 analysing 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.

Musicians, by contrast, have rarely used quantitative analyses in their work. Some pilot studies have been done by historians of music publishing, notably Rudolf Fleisch's analysis 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 bibliographical records.

There currently exist numerous datasets of metadata relating to music publications, manuscript scores, and also the ephemeral materials that surround classical concert-going (e.g., programme notes, handbills). Several of these datasets are the product of cumulative bibliographic research over many decades. Some are maintained by research libraries as part of their online catalogues.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires specific permission and/or a fee. Request permissions from Permissions@acm.org

musical culture could equally well be interrogated via an analysis of locations, publishers, performers, genres, social practices or migratory patterns
Example result from Rose et al. (2015)

...suggesting that from the 1550s composers took more initiative in publishing their works for financial gain or as symbols of prestige and skill.

1 European output of printed music by decade, 1500–1699. Red shading denotes anthologies (data from RISM B/I) and blue shading indicates single-composer editions (data from RISM A/I)
My own big data history of music

CANTOSTREAM project (November 2021)

• polyphony dating from 1500—1700
• musical objects as big data

• focus on ‘tonal structures in early music’ (Judd 1998), for example:
  • what was the role of the authentic-plagal distinction?
  • how did modality transition into tonality?
  • how did the Mixolydian modes disappear?

Mode Æolien and Sous-Æolien (on G) from La rhétorique des Dieux (1652), a 12-mode cycle consisting of 56 lute pieces.
The question, once again

• are we ready for a big data history of music?
• what would, ideally, a ‘yes’ entail?

big data analysis is able to deliver new facts and insights that enrich the study of music history

• by confirming, questioning or providing alternatives to current views
• by shedding light on open issues in the discipline
• by suggesting new directions for research
Five questions about readiness

1. community: who are ‘we’?
2. data: what data are available?
3. processing: what methods and tools are available?
4. study: what research can be done?
5. persuasion: how convincing are the outcomes?

obviously, these are highly interconnected
Q1. community: who are ‘we’?
Computational musicologists?

• computational musicologists have been eager since at least 1967 (Logemann)
  • ‘heroic period’ of CM
  • focus on musical objects (as in musicology of the day)

• then musicology developed in a different direction
  • contextualisation rather than ‘the music itself’

• are computational musicologists ready to reconnect?

The Princeton Josquin Project was first example of a big data approach in musicology (Mendel 1969)
Technology acceptance in musicology

- What Do Musicologists Do All Day 2015 outcomes (Inskip & Wiering 2015)
  - musicologists use technology widely and creatively
  - widely-shared feeling that this has changed musicology
  - access and discovery are very important
  - only weak interest in software for processing musical data (music notation programs excepted)

benefits of technology responses: terms related to processing are infrequent (n=625)
WDDMDAD the pandemic edition (2022)

Q18: Which of the [technology-related] changes you have been forced to make are you going to keep after the pandemic is over?

<table>
<thead>
<tr>
<th>amount of change</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>(almost) none</td>
<td>73</td>
</tr>
<tr>
<td>specific change(s)</td>
<td>210</td>
</tr>
<tr>
<td>most or all</td>
<td>19</td>
</tr>
<tr>
<td>don't know</td>
<td>9</td>
</tr>
<tr>
<td>unclear or unusable</td>
<td>6</td>
</tr>
<tr>
<td>total</td>
<td>317</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>area of change</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>human interaction: teaching, conferences, collaboration, ...</td>
<td>130</td>
</tr>
<tr>
<td>digital and physical materials</td>
<td>76</td>
</tr>
<tr>
<td>remote work, travel</td>
<td>50</td>
</tr>
<tr>
<td>hardware, software, services</td>
<td>24</td>
</tr>
<tr>
<td>other changes: health, skills and attitudes, ...</td>
<td>21</td>
</tr>
<tr>
<td>total</td>
<td>301</td>
</tr>
</tbody>
</table>

strongest impacts on social structure of the discipline and on research materials
One possible storyline in the responses

- Tech skills were essential for survival.
- Many discovered potential of technology for research.
- Specifically, online access to materials (matched by concerns about study of physical resources).
- Maybe a more collaborative attitude.
- New research topics and methodological reflections.
- Promising for reconnection.

Perhaps none of them [my changes] huge or transformative on their own, but in sum the changes in habit and standard operating procedure have been huge. [176]

Our tools shape the way we think - the increased speed of digitisation allows for more analysis of big data and more research questions based on questions related to quantifications in general.

I am concerned about how reliance on digital resources can shape research questions and trends - a topic which to my knowledge no one has looked into.
Q2. data: what data are available?
### Musical data come in many kinds

<table>
<thead>
<tr>
<th>type</th>
<th>examples</th>
<th>number of musical items (wild estimate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>metadata</td>
<td>RISM, library catalogues</td>
<td>&gt; 10,000,000</td>
</tr>
<tr>
<td>digitised sources</td>
<td>EMO, Gallica, DIAMM</td>
<td>100,000-500,000</td>
</tr>
<tr>
<td>digital editions (mostly PDF)</td>
<td>CPDL</td>
<td>50,000-200,000</td>
</tr>
<tr>
<td>MIDI</td>
<td>Classical Archives</td>
<td>100,000-500,000</td>
</tr>
<tr>
<td>encodings of notation</td>
<td>MEI, MusicXML, humdrum</td>
<td>20,000-100,000</td>
</tr>
<tr>
<td>audio recordings</td>
<td>CD, MP3, streaming</td>
<td>&gt; 100,000,000</td>
</tr>
<tr>
<td>feature sets</td>
<td>Million Song Dataset, Essentia</td>
<td>2,000,000-5,000,000</td>
</tr>
<tr>
<td>born-digital materials</td>
<td>generative music; tags, tweets</td>
<td>&gt; 10,000,000</td>
</tr>
</tbody>
</table>
Musical data come in many kinds

<table>
<thead>
<tr>
<th>type</th>
<th>examples</th>
<th>number of musical items (wild estimate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>audio recordings</td>
<td>CD, MP3, streaming</td>
<td>&gt; 100,000,000</td>
</tr>
<tr>
<td>born-digital materials</td>
<td>generative music; tags, tweets</td>
<td>&gt; 10,000,000</td>
</tr>
<tr>
<td>metadata</td>
<td>RISM, library catalogues</td>
<td>&gt; 10,000,000</td>
</tr>
<tr>
<td>feature sets</td>
<td>Million Song Dataset, Essentia</td>
<td>5,000,000-20,000,000</td>
</tr>
<tr>
<td>digitised sources</td>
<td>EMO, Gallica, DIAMM</td>
<td>100,000-500,000</td>
</tr>
<tr>
<td>MIDI</td>
<td>Classical Archives</td>
<td>100,000-500,000</td>
</tr>
<tr>
<td>digital editions (mostly PDF)</td>
<td>CPDL</td>
<td>50,000-200,000</td>
</tr>
<tr>
<td>encodings of notation</td>
<td>MEI, MusicXML, humdrum</td>
<td>20,000-100,000</td>
</tr>
</tbody>
</table>

what we need most is also the rarest
Situation of early music even more dramatic

• Harasim et al. (2021)
  • historical development of tonality

• MIDI representations
  • 12635 from Classical Archives
  • 777 from Du Chemin and ELVIS

• scarcity is bad news for CANTOSTREAM project

Fig. 1 Histogram of composition dates of 13,402 musical pieces in the dataset.
Early music datasets in recent research

<table>
<thead>
<tr>
<th>publication</th>
<th>size</th>
<th>comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geelen et al. (2021). Clustering analysis of renaissance polyphony</td>
<td>1248 works</td>
<td>humdrum; Josquin Research Project</td>
</tr>
<tr>
<td>Harasim et al. (2021). Exploring the foundations of tonality</td>
<td>777+ works</td>
<td>midi; Lost Voices, CRIM, ELVIS</td>
</tr>
<tr>
<td>Long (2020). Hearing homophony</td>
<td>nearly 400 works</td>
<td>not a computational study; canzonettas by Gastoldi and others</td>
</tr>
<tr>
<td>Upham, Cumming (2020). Auditory streaming complexity</td>
<td>216 works, 1042 movements</td>
<td>midi; Josquin Research Project, RenComp7</td>
</tr>
<tr>
<td>Arthur (2021). Vicentino versus Palestrina</td>
<td>707 movements</td>
<td>humdrum; complete Palestrina Masses</td>
</tr>
</tbody>
</table>

small datasets, some great composers
## Treasure hunt (selection)

<table>
<thead>
<tr>
<th>Dataset name</th>
<th>content</th>
<th>items before 1700</th>
<th>encoding system</th>
<th>url</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choral Public Domain Library</td>
<td>choral music of any age</td>
<td>unknown</td>
<td>various</td>
<td><a href="https://www.cpdl.org/wiki/">https://www.cpdl.org/wiki/</a></td>
</tr>
<tr>
<td>ECOLM</td>
<td>16-18th century lute music</td>
<td>1619</td>
<td>tabcode</td>
<td>ecolm.org</td>
</tr>
<tr>
<td>ELVIS database</td>
<td>heterogeneous, aggregated</td>
<td>1000 (c)</td>
<td>various</td>
<td><a href="https://database.elvisproject.ca/">https://database.elvisproject.ca/</a></td>
</tr>
<tr>
<td>Josquin Research project</td>
<td>works by Josquin and contemporaries</td>
<td>902</td>
<td>MusicXML, humdrum</td>
<td><a href="https://josquin.stanford.edu/">https://josquin.stanford.edu/</a></td>
</tr>
<tr>
<td>Tasso in Music</td>
<td>musical settings of Tasso's poems</td>
<td>778</td>
<td>humdrum, MEI, MusicXML</td>
<td><a href="https://www.tassomusic.org/">https://www.tassomusic.org/</a></td>
</tr>
<tr>
<td>Neuma</td>
<td>multiple subcollections</td>
<td>600 (c)</td>
<td>MusicXML, MEI</td>
<td><a href="http://neuma.huma-num.fr/">http://neuma.huma-num.fr/</a></td>
</tr>
<tr>
<td>Lost Voices</td>
<td>16th c. French chansons</td>
<td>380 (c)</td>
<td>MEI</td>
<td><a href="http://digitalduchemin.org/">http://digitalduchemin.org/</a></td>
</tr>
<tr>
<td>CRIM</td>
<td>16th c. imitation masses and their models</td>
<td>250 (c)</td>
<td>MEI</td>
<td><a href="https://crimproject.org/">https://crimproject.org/</a></td>
</tr>
<tr>
<td>Gesualdo online</td>
<td>complete works of Gesualdo</td>
<td>222</td>
<td>MEI</td>
<td><a href="https://ricercar.gesualdo-online.cesr.univ-tours.fr/">https://ricercar.gesualdo-online.cesr.univ-tours.fr/</a></td>
</tr>
<tr>
<td>miami Publication server of the University of Münster</td>
<td>3 digital editions of early 17th c. sacred music</td>
<td>230 (c)</td>
<td>Lilypond, MIDI</td>
<td><a href="https://miami.uni-muenster.de/">https://miami.uni-muenster.de/</a></td>
</tr>
<tr>
<td>Measuring polyphony</td>
<td>14th century motets</td>
<td>61</td>
<td>MEI</td>
<td><a href="https://measuringpolyphony.org/">https://measuringpolyphony.org/</a></td>
</tr>
<tr>
<td>Computerized Mensural Music Editing</td>
<td>15-16th c. polyphony</td>
<td>59</td>
<td>CMME</td>
<td><a href="http://www.cmme.org/">http://www.cmme.org/</a></td>
</tr>
<tr>
<td>Furnace and fugue</td>
<td>digital edition of Atalanta fugiens (1618)</td>
<td>50</td>
<td>MEI</td>
<td><a href="https://furnaceandfugue.org/">https://furnaceandfugue.org/</a></td>
</tr>
</tbody>
</table>
How to move forward?

• coordinated community initiative

• Optical Music Recognition
  • even the best OMR is far from perfect
  • imperfect data is usable in some scenarios (→ David Lewis)

• use audio recordings
  • much more data (including Great Composers)
  • much software development
  • intellectual property is a barrier

related items retrieved by searching OMR output (Crawford et al. 2018)

part of the audio corpus analysed by Weiß et al. (2019)
Q3. processing: what methods and tools are available?
Processing encodings

• quite a few toolboxes
  • humdrum, music21, jSymbolic, IDyOM, retrieval methods...
• weaker on detectors for high-level features
• example: cadence detection in renaissance polyphony
  • cadence is a marker for a boundary
  • and a manipulable musical object
  • studied in CRIM (Alex Morgan) and Polifonia (Christophe Guillotet-Nothmann)
• current status: cadence detection is an interactive process
  • another data requirement: expert annotations
Processing audio

• situation is more complex
• chroma features are marvellous
  • collect energy per pitch class
  • analysis of harmony
• polyphonic transcription is still an open problem
  • recent work on vocal polyphony (Cuesta et al. 2020, Rosenzweig 2022)
• most computational methods developed for popular music
  • e.g. boundary detectors perform rather weakly on polyphony
  • retraining needs yet another kind of data in quantity

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 easy pieces</td>
<td>0.90</td>
<td>0.78</td>
<td>0.83</td>
</tr>
<tr>
<td>8 hard pieces</td>
<td>0.53</td>
<td>0.37</td>
<td>0.43</td>
</tr>
</tbody>
</table>

chords extracted by chordify.net

performance of OLDA segmentation algorithm (McFee & Ellis 2014) on renaissance vocal music
Q4. study: what research can be done?
Focussed studies

- Attribution of anonymous works from *Leuven Chansonnier* (Geelen et al. 2021)
- Palestrina’s counterpoint practice vs. Vicentino’s theory (Arthur 2021)
- Establishment of mode in renaissance duets (Arthur et al. 2022)
  - Mode ‘families’ easy to distinguish
  - Authentic-plagal distinction not so clear
Studying large-scale developments

• statistical cognitive modelling of mode (Harasim et al., 2021)
  • 500 years; c. 13,000 compositions
  • modal clarity at its highest in classical period

• stylistic information in pitch class distributions (Yust 2019)
  • period of nearly 400 years

• style evolution based on harmonic intervals extracted from audio (Weiß et al. 2019)
  • intervals, chord progressions, complexity
  • overall, outcomes align with Harasim’s findings

interval category features distributed over the years
Observations

• focus on tonal aspects, less on rhythm, melody, genre
• mid-level studies seem to be rare
  • Long’s (2020) study of the canzonetta is a ‘manual’ big data study
• these require much richer contextual data, such as
  • location
  • date
  • genre
  • performance information

in short, interoperability with other (meta) data sets
Q5. persuasion: how convincing are the outcomes?
Some obstacles to persuasion

• questions that are too big
  • insights such as decreasing tonal clarity in the 19th century isn’t really new musicological knowledge

• pushing things too far
  • perplexing classification outcomes are frequent
  • these are not musicologists’ problems

• using whatever data is at hand
  • choice and quality of data matters hugely in musicological research

• reductionism
  • e.g. a mode is just another scale --> look at pitch class only

• conceptual shopping
  • not paying attention to the context from which a concept or idea originates

The assignment of Mussorgsky and Faure to this cluster is rather surprising since most of the late romantic composers (Mahler, Strauss) as well as the impressionists (Debussy, Ravel) are assigned to the Romantic cluster. This kind of unexpected observations could serve as an inspiration for musicological research (Weiß et al. 2019)
What makes computational results convincing

- choose problems at the right level of granularity
- make use of the richness of main concepts
  - mode, for example
- choose the right data
  - quality more important than size
  - appropriate contextual information
- maximise transparency and explainability
  - if necessary at the cost of performance
- apply data and tool criticism to investigate biases
Conclusion
Summary answers to the questions

- community: divided, but tech acceptance is growing
- data: problematic
- methods and tools: decent, but hoping for advances in audio analysis tools
- research: promising, creative
- persuasion and insight: rather weak, especially for studies of large-scale developments
Are we ready for a big data history of music?
We aren’t ready yet
Five easy steps towards maturity

1. take interdisciplinarity *very seriously*
   • connect to existing musicological expertise

2. sustainable integration of resources
   • projects like Polifonia are great but tend not to outlive their funding
   • lightweight: money will only become scarcer

3. practice data and tool criticism continuously

4. MEI-data button for all music notation programs

5. <...*insert your own favourite step here*...>
Thank you!


References, continued


