frontiers in Digital Humanities

SPECIALTY GRAND CHALLENGE
published: 19 August 2015
doi: 10.3389/fdigh.2015.00004

The challenge of data in digital musicology

Laurent Pugin[*]

Swiss RISM Office, Bern, Switzerland

Keywords: digital musicology, content-based music resources, music analysis, music encoding

Most of our work in the humanities is increasingly driven by digital technology. Musicology is no exception and the field is undergoing the same revolution as all disciplines in the humanities. There are at least two key areas in which digital technology is transforming research: access and scale. Technology, and the internet in particular, has radically changed how we can access data, but also how we can make research results accessible to others. Correlatively, the scope of projects can be broadened to a completely new extent.

What does this mean for musicology? Scholars in musicology base their work on a wide range of materials. Since most of the music that forms our heritage in Western culture has been preserved in a text-based form, this is by far, the most widely used type of material for musicological studies. Handwritten and printed sources constitute the core data, but historical studies also rely on various types of textual and archival material, be they letter writings, libretti, or inventories of diverse kind. These are essential for understanding the socio-economic context in which the music sources were written or produced and for better understanding of specific aspects, such as performance practice of the time. Performance practice study itself may also be based on sound recordings when focusing on relatively recent history, as it is often the case for studies in ethno-musicology or in folk-songs (Cook, 2010). Obtaining access to the sources has always been a struggle for musicologists. Only a few years ago, studying a particular source meant first locating the relevant sources using printed bibliographies, writing to the holding library, and then waiting for a microfilm to be prepared and sent out. The process could take months and be unpredictably expensive, with no guarantee of success. Such an obstacle seriously reduced the breadth of research musicologists could reasonably envisage, with a consequent inclination toward close-reading approaches on a restricted set of sources.

With the coming of the digital world, the situation changed. Many resources are now available online, including the bibliographic finding aids, which makes locating sources significantly easier. Collections are being digitized and made accessible online, which greatly facilitates access to them for musicologists. This is also the case for secondary sources. Some projects are composer-specific, such as the Digital Archive of the Beethoven-Haus, others are repertoire-oriented, such as the digital image archive for medieval manuscripts (DIAMM) or based on a particular library collection, such as the Julliard Manuscript Collection, to cite only three examples. In the archives, digital cameras are often allowed and can be used to capture sources quickly. It is now straightforward for scholars to store thousands of images on their personal computer, in the cloud, or even share them on community websites, although this in its turn raises new copyright concerns. What other issues need to be addressed?

Digital access in musicology is still overwhelmingly linked to images. Several important digital musicology research projects, such as the OCVE and the Edirom projects, focusing mostly on philological issues have been very successful in relying extensively on digital image resources (Bradley and Vetch, 2007; Bohl et al., 2011). However, digital musicology projects that address a wide range of other issues, such as music analysis or music searching, require access to the music itself in digital form, are referred to as content-based resources. Musicology has never been behind other disciplines for experimenting with computational approaches in these domains, quite on the

Edited and reviewed by:
Eleanor Selfridge-Field,
CCARH, USA; The Packard
Humanities Institute, USA

***Correspondence:**
Laurent Pugin
laurent.pugin@rism-ch.org

Specialty section:
This article was submitted to Digital Musicology, a section of the journal Frontiers in Digital Humanities

Received: *28 November 2014*
Accepted: *04 August 2015*
Published: *19 August 2015*

Citation:
Pugin L (2015) The challenge of data in digital musicology.
Front. Digit. Humanit. 2:4.
doi: 10.3389/fdigh.2015.00004

contrary. However, obtaining or accessing high quality datasets remains a serious hurdle, especially on a large scale, in a similar way to accessing sources a couple of decades ago. It is a major barrier that needs to be removed if digital musicology research is to be taken to the next level.

Several initiatives have laid down the basis for large-scale content-based resources. First and foremost, the CCARH with its KernScores repository[1], which represents years of careful data creation and curation is made available for research and is an invaluable contribution. The Josquin Research Project (JRP[2]) at Stanford is a groundbreaking project that is currently building a considerable dataset of pieces of Josquin des Prez and of other composers of the time (1400–1500). Another is the Electronic Locator of Vertical Interval Successions project at McGill University (ELVIS[3]). These two projects pursue similar goals and follow more or less comparable strategies: respectively creating or collecting a large collection of data and making it accessible and analyzable by integrating state-of-the-art analysis tools Humdrum and Music21. Their output in terms of counterpoint analysis is a breakthrough and opens new perspectives for style analysis and composition attribution. The use of the harmonic and melodic intervals in ELVIS illuminates areas in which innovative research might be needed to address the question of how to represent music appropriately for such corpus-based analysis undertakings.

These are undoubtedly models to follow, but they also illustrate how much still needs to be done. They hold a few thousand pieces[4] while nearly one million music sources are inventoried by the RISM for just the period around 1600–1800. This figure includes many copies of identical or similar pieces, but it certainly provides us with an indication of what still remains to be accomplished. The JRP, and ELVIS to some degree, is also particular in focusing on early music, which has often been the basis of leading digital research projects. Josquin was at the heart of one of the very first computational musicology projects in the 60s. There is currently no project truly comparable to the JRP for later repertoires, although some projects have admittedly begun to fill this huge gap: the aforementioned ELVIS project and also the Transforming Musicology project in the UK with its part focusing on Wagner Leitmotivs[5].

A considerable portion of all these projects must be devoted to data creation and collecting, mainly because optical music recognition (OMR) remains a challenge. Much of the time, primary (or secondary) sources must be (re-)transcribed by hand. And there it sometimes seems we are stuck in a loop. Is it not frustrating to know that the large majority of editions published over the last decade was prepared using digital tools, but that eventually only paper or PDF versions will survive? Very often, once transcribed and edited, the music is published but the digital content is not made available. For publishers, there are certainly commercial reasons for this, but this is not the only factor. There is a lack of awareness of the issue, and musicologists should make their case for preserving their work in digital format. It is a tremendous waste of resources and, as such, should be addressed before any other issue. Perhaps this is not different from what is experienced in other fields, but there is an important difference to take into account, namely that there is currently no OMR technology that performs as good as OCR, and that transcribing or correcting music scores by hand is overly time-consuming. There is a need for better tools for this task, and the single interface for music score searching and analysis (SIMSSA[6]) based at McGill is a front runner. The creation of OMR frameworks will make it possible to create and gather data in completely new ways, including by enabling crowd-sourcing correction.

Several challenges will need to be faced when creating large datasets. First of all, ensuring data quality will be crucial and should not be overlooked.[7] We may expect online tools and infrastructures for distributed data correction to make such tasks less tedious and to help in constantly improving the data quality. Ensuring interoperability will be another challenge. The development of music computer codes has shown us how different centers of interest and different focuses can lead to countless barely compatible initiatives (Selfrige-Field, 1997). Similar situations should be avoided in the future and we need to make sure that data resources will be interoperable as much as possible, despite the differences in focus or in repertoire. The music encoding initiative (MEI[8]) is definitely well placed to play a unifying role. It covers a wide range of music notations, but most importantly, it includes features for enriched encodings and is extensible. These features should be utilized as the basis for more advanced computational models. Creating possibilities for enriching the datasets while maximizing interoperability will open new research areas and bridge research fields. This should also avoid locking in analytical data in the datasets themselves, as it has been the case sometimes in the early projects (e.g., the use of pre-tonal terms in the Josquin project of the 1970s). For example, storing detailed metadata, structural or harmonic analysis hypothesis of encoded scores, or any other type of annotations, in the datasets themselves will be the best path toward making them directly usable in or by other applications. This will be particularly useful for integrating music information retrieval (MIR) tools in digital musicology applications, and it will not be limited to notation encoding processing. Quite on the contrary, it will be an excellent way of integrating higher-level musicological knowledge into audio or performance analysis. Finally, with the emergence of linked open data (LOD) resources, having reliable identifiers is a pre-requisite. There is much to be done to make musical works identifiable in the digital domain. However, it is a complex question, not to say a minefield. Where does one draw the limits on what constitutes a distinct work? To what degree of variants? Is an aria of an opera a work, or only the entire opera? These are typical and recurrent music bibliography issues that become critical in large-scale digital musicology projects collating heterogeneous types of musical works.

[1] http://kern.ccarh.org/

[2] http://jrp.stanford.edu/

[3] http://elvisproject.ca/

[4] At the date of writing (October 2014), the KernScores holds about 3,700 pieces, JRP 1,000 of pieces, and the ELVIS 6,000.

[5] http://www.transforming-musicology.org/

[6] http://simssa.ca/

[7] The JRP project includes in its data creation workflow several time-consuming verification steps. This will remain necessary.

[8] http://www.music-encoding.org/

Many projects have attempted to solve the issue, including following the abstract FRBR mode (Riley, 2011). However, none of them has succeeded so far in scaling up.

In the emerging field of digital humanities, huge gaps exist in our knowledge and capabilities, and we can see digital musicology projects as an opportunity to widen and bridge research fields. The tools we require are in the process of being invented. The foundations for our work are in place; what is required now is the opportunity to leverage these existing resources to attain a new level of knowledge and insight. This will result in a transformation of the way we access music and the extent to which we make use of it.

References

Bohl, B., Kepper, J., and Röwenstrunk, D. (2011). Perspektiven digitaler Musikeditionen aus der Sicht des Edirom-Projekts. *Die Tonkunst* 5: 270–6.

Bradley, J., and Vetch, P. (2007). Supporting annotation as a scholarly tool – experiences from the online Chopin Variorum edition. *Lit. Linguist. Comput.* 22: 225–41. doi:10.1093/llc/fqm001

Cook, N. (2010). The ghost in the machine: towards a musicology recordings. *Musicae Scientiae* 14: 3–21. doi:10.1177/102986491001400201

Riley, J. (2011). Leveraging the FRBR model for music discovery and data sharing. *OCLC Syst. Serv.* 27: 175–89. doi:10.1108/10650751111164551

Selfrige-Field, Eleanor., ed. (1997). *Beyond MIDI: The Handbook of Musical Codes*. Cambridge MA: The MIT Press.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.