Big Data for Musicology

Tillman Weyde
Stephen Cottrell
Jason Dykes
Emmanouil Benetos
Daniel Wolff, Dan Tidhar
{t.e.weyde,
stephen.cottrell.1,j.dykes,
emmanouil.benetos.1,
daniel.wolff.1,
dan.tidhar.1} @city.ac.uk

Mark Plumbley
Simon Dixon
Mathieu Barthet
{mark.plumbley,
simon.dixon,
mathieu.barthet}
@eecs.qmul.ac.uk

Nicolas Gold
Samer Abdallah
{n.gold, s.abdallah}
@ucl.ac.uk

Mahendra Mahey
Adam Tovell
{Mahendra.Mahey
Adam Tovell} @bl.uk

ABSTRACT

Digital music libraries and collections are growing quickly and are increasingly made available for research. We argue that the use of large data collections will enable a better understanding of music performance and music in general, which will benefit areas such as music search and recommendation, music archiving and indexing, music production and education. However, to achieve these goals it is necessary to develop new musicological research methods, to create and adapt the necessary technological infrastructure, and to find ways of working with legal limitations. Most of the necessary basic technologies exist, but they need to be brought together and applied to musicology. We aim to address these challenges in the Digital Music Lab project, and we feel that with suitable methods and technology Big Music Data can provide new opportunities to musicology.

1. INTRODUCTION

Music research, particularly in fields like systematic musicology, ethnomusicology, or music psychology has developed as "data oriented empirical research" [16], which benefits from the development of computing methods and infrastructure, and both quantitative and qualitative methodologies are increasingly found across the broad field of music studies [3]. Particularly in ethnomusicology there has been a recent growing interest in computational methods and their application to audio data collections (see [11], [2]). Another area where empirical research has a long tradition is the study of performance of Western music (see [9], [10]). However, research in this area has been limited to small datasets, by current standards, because of technological and legal reasons.

Musicology has traditionally focused on analyses of musical texts for its understandings of musical meaning. In the last few decades, however, it has broadened this approach, taking an 'interpretive' turn not unrelated to similar movements in the social sciences. Part of this move has seen an increased focus on studying music in performance, rather than focusing only on musical scores. The study of musical recordings has become an important part of music studies (e.g.,[4]). This has led to a range of computational approaches being developed which seek to identify and explain the different performance aesthetics and stylistic changes within or between given repertories, or between performances of the same material. This return to a form of "comparative musicology" [5] has been made possible by developments in computing technology and the evolution of specific software programs. Nevertheless, much of this work has concentrated on detailed analyses of small data sets.

1.1 Music Information Retrieval and Big Music Data

Researchers in Music Information Retrieval (MIR) have started to explore large datasets, particularly in commercial recommendation and playlisting systems (e.g. The Echo Nest, Spotify[1]), but that work is largely separate from research into music as such. There are differences in the terminologies, methods, and goals between MIR and musicology [13] as well as technological and legal barriers that prevent musicologists from taking advantage of large datasets that are available now. It is therefore necessary to support music research by bridging the gap to MIR and enabling access to large music audio data collections and providing powerful analysis and visualisation tools, avoiding legal restrictions and overcoming technical limitations.

[1] `http://the.echonest.com`, `http://www.spotify.com/`

2. CHALLENGES OF BIG MUSIC DATA

Most music research has so far required manual transcription and alignment, which limits the number and scale of investigations as well as the statistical significance of results.

2.1 Music Transcription and Score Alignment

To perform musicological analyses on large collections of audio data, we need to apply automatic music transcription (AMT, the process of converting an acoustic musical signal into some form of musical notation). In recent years, the problem of automatic music transcription has gained considerable research interest due to the numerous applications associated with the area, such as automatic search and annotation of musical information, interactive music systems, and computational musicology [12]. Even for expert musicians, transcribing music is not a trivial task, and while the problem of automatically transcribing monophonic music is considered to be solved, the creation of a system able to transcribe multiple-instrument polyphonic music still remains an open challenge [15]. However, it has been shown that genre-specific, instrument-specific, or user-informed automatic transcription systems can perform with a high degree of accuracy [1]. In addition, high-level music descriptors such as chords and keys can be reliably extracted from an intermediate automatic transcription step[14].

2.2 Technologies

The progress in automatic music transcription, music information retrieval, and computational musicology as well as in large-scale data processing architectures brings now new opportunities for music research that up until now required manual methods, thus preventing large-scale audio analysis. Although most of the technologies needed for this type of research exist, the methodologies need to be developed, the technical infrastructure and tools need to be put in place, and legal issues need to be taken into account. In Music Information Retrieval research, some of these points have been addressed by the Million Song Dataset[2], which provides low-level audio analyses and metadata for one million songs, but not the audio. However, there are significant limitations: the set consists of popular music, the audio analysis that was performed is limited to low-level features and there is no way for researchers to apply their own algorithms to the audio data.

2.3 Legal restrictions

Data analysis in general is subject to agreement by copyright holders, but in many jurisdictions special copyright regulations exist, that enable research and publication of derived data. E.g., a copyright exception for text and data mining is part of the Intellectual Property Bill that was in the UK recently in response to the Hargreaves Review of Intellectual Property and Growth 2011[3]. This could even apply to the analysis of internet radio streams, albeit posing some additional technical challenges (segmentation, audio fingerprinting) but providing a vast source of audio data. Even with copyright clearance, the analysis of music data, especially audio, requires significant technical expertise and resources. An infrastructure is needed than enables researchers to make use of previously computed results as well as defining their own analyses. By combining automatic transcription, symbolic and audio analysis with access to large data collections and collection-level analysis tools, there is an opportunity to expand music research and support a research community.

3. APPROACH

The authors of this paper collaborate in the AHRC funded project *Digital Music Lab - Analysing Big Music Data* [4] to develop methods and technologies to support the use of Big Data in musicology. The Digital Music Lab project will develop a software infrastructure for exploring and analysing large-scale audio collections, particularly with regard to music performance and its relation to musical structure.

3.1 Methods

With regards to methods, we are exploring relevant questions in musicology that can be specifically addressed with large data collections. Questions that benefit from large data collections are e.g. how music performance develops over time, how it differs between regions, and where similarities are between different cultures. Studies of cultures of music performance are examples of research that can hardly be conducted without large amounts of data. With current resources and technologies available to music researchers, these studies are hard to realise this although most of the technologies (e.g. transcription, audio analysis) and the datasets (e.g. collections at the BL, archive.org) that are required exist. What is missing is a methodology for working on large amounts of data in music performance analysis, the tools and infrastructure that bring needed technologies together, and access to data collections in a way that is practical and does not infringe copyright. The higher degree of automation in such a system requires different approaches to music performance research with different trade-offs on large and small scales of analysis.

3.2 Technologies

The technologies needed for Big Data applications have been rapidly evolving in the last few years. The most popular approach is parallelisation with Map-Reduce[7], using the Hadoop framework[5]. We are embedding VAMP plugins and other audio feature extraction into the Map step and perform collection level analysis in the Reduce step of Hadoop. We are working on visualisations for to support non-technical users in defining queries and analysing the results.

3.3 Services

A major output of the project will be a software service infrastructure with two prototype installations. One installation will enable researchers, musicians and general users to explore, analyse and extract information from audio data stored in the British Library (BL), which cannot be used outside the BL for copyright reasons. Another installation will be hosted by the Centre for Digital Music at Queen Mary University of London and provide facilities to analyse the audio collections by *I Like Music*[6], the CHARM[7]

[2] http://labrosa.ee.columbia.edu/millionsong/
[3] http://www.ipo.gov.uk/types/hargreaves.htm
[4] http://dml.city.ac.uk
[5] http://hadoop.apache.org
[6] http://www.ilikemusic.com/
[7] http://www.charm.kcl.ac.uk/sound/sound.html

dataset, the Isophonics[8] datasets and other freely accessible datasets. We will provide researchers with the tools to analyse large-scale music audio, scores and metadata. The combination of state of the art music transcription [1] and music analysis on the audio and the symbolic level (e.g. [8], [6]) with collection level analyses will allow for exploration and quantitative research on music that has not been not possible at this scale up until now.

3.4 Copyright

To enable research on copyright restricted data, we are developing a web service, that allows users to run experiments without access to he audio data. This has practical advantages, as it will enable users to do research remotely as well as allow libraries and other collection holders to put their data to use without infringing copyright. The DML system will be distributed on virtual machines, that institutions can install to allow access to their own data and to allow distributed experiments across multiple collections.

4. CONCLUSIONS

Big Music Data collections are being compiled and becoming available. The use of large collections for musicology is promising, but requires methodological adaptations, technological development and strategies for avoiding copyright infringements. We are developing the required methods, tools and strategies, and we feel that this approach has the potential to provide useful extensions to music research with quantitative methods and Big Music Data tools.

References

[1] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, pages 1–28, 2013.

[2] S. Canazza, A. Camurri, and I. Fujinaga. Special section: Ethnic music audio documents: From the preservation to the fruition. *Signal Processing*, 90(4):977–1334, 2010.

[3] E. Clarke and N. Cook, editors. *Empirical Musicology: Aims, Methods, Prospects*. Oxford University Press, 2004.

[4] N. Cook. The ghost in the machine: Towards a musicology of recordings. *Musicae Scientiae*, 14(2):3–21, 2010.

[5] S. Cottrell. The rise and rise of phonomusicology. In A. Bayley, editor, *Recorded Music: Performance, Culture and Technology*. Cambridge University Press, 2009.

[6] R. de Valk, T. Weyde, and E. Benetos. A machine learning approach to voice separation in lute tablature. In *14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

[7] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, San Francisco, CA, 2004.

[8] `http://www.isophonics.net/datasets`

[8] N. Degara, M. E. P. Davies, A. Pena, and M. D. Plumbley. Onset event decoding exploiting the rhythmic structure of polyphonic music. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1228–1239, 2011.

[9] A. Gabrielson. Music performance research at the millennium. *Psychology of Music*, 31(3):221–272, 2003.

[10] W. Goebl and G. Widmer. On the use of computational methods for expressive music performance. In T. Crawford and L. Gibson, editors, *Modern Methods for Musicology: Prospects, Proposals, and Realities*, pages 93–113. Ashgate Publishing, London, 2009.

[11] E. Gómez, P. Herrera, and F. e. Gómez-Martin. Special issue: Computational ethnomusicology. *Journal of New Music Research*, 42(2), 2013.

[12] A. Klapuri and M. Davy, editors. *Signal Processing Methods for Music Transcription*. Springer-Verlag, New York, 2006.

[13] A. Marsden. What was the question? music analysis and the computer. In T. Crawford and L. Gibson, editors, *Modern Methods for Musicology*, pages 137–147. Ashgate, 2009.

[14] L. Mearns, E. Benetos, and S. Dixon. Automatically detecting key modulations in j.s. bach chorale recordings. In *Proceedings of the Sound and Music Computing Conference*, Padova, 2011.

[15] M. Müller, D. Ellis, A. Klapuri, and G. Richard. Signal processing for music analysis. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1088–1110, 2011.

[16] R. Parncutt. Systematic musicology and the history and future of western musical scholarship. *Journal of Interdisciplinary Music Studies*, 1:1–32, 2007.