FAIR, Open, Linked: Introducing the Special Issue on Open Science in Musicology

FABIAN C. MOSS[1]
École Polytechnique Fédérale de Lausanne, Switzerland

MARKUS NEUWIRTH
Anton Bruckner University, Linz, Austria

Published 2021 December 10; https://doi.org/10.18061/emr.v16i1.8246

EMPIRICAL musicology crucially relies on the creation, publication, distribution, and analysis of data. Despite the progress made over the past decades in this vibrating field, numerous issues regarding the accessibility, sharing, and linkage of data, the reproducibility of research findings, and the general role of transparency remain challenging.[2] In many disciplines, these issues are addressed under the umbrella of the Open Science movement and the adherence to the FAIR (findable, accessible, interoperable, reusable) principles for scientific data management (Wilkinson et al., 2016).

Findable. In order for data to be findable, they must possess a unique and persistent identifier such as a Digital Object Identifier (DOI) or Uniform Resource Identifier (URI) that makes them uniquely identifiable on the web. Carefully constructed metadata with extensive descriptions of the data additionally facilitate finding data sets relevant for one's own research. Optimally, the data and metadata are indexed in centralized databases that can be searched for keywords and contents.

Accessible. Knowing where data is stored is not sufficient to make it usable; it also needs to provide means of accessing it. This can, for instance, be facilitated by providing a customized Application Programming Interface (API) for data retrieval. Even if data access is limited, for instance due to ethical or legal reasons, the metadata should be public nonetheless. Accessibility also concerns the formats in which the data are stored. Proprietary software used, for example, in statistical analyses or encoding might not be available to everyone, which means excluding all those researchers who might not have the resources to buy software licenses.

Interoperable. Data is not monolithic (see Sugimoto, Ekbia, and Mattioli, 2016): They usually arise within certain contexts and refer, explicitly or implicitly, to other kinds of data. Moreover, data created for tackling specific research questions may be useful for other researchers to answer different ones. The combination and linkage of datasets of different provenance requires the data – at least in principle – to be translatable into different representations as well as standardized vocabularies for metadata that are readable by different operating systems. Naturally, this is already the case when communities agreed upon encoding standards. For many applications in empirical musicology, however, this is not yet the case, and metadata is often provided as a simple README text file.

Reusable. Data can only be reused for different purposes if accompanying metadata explicate how they were created, what they represent, and under which license they have been published. Issues of copyright and ethical concerns regarding participant data from experiments are important responsibilities for everyone who publishes data. It is equally important to state these circumstances unequivocally in order to avoid possible ethical and legal implications for people who were not involved in the data creation process. The repeated use of data also serves to ensure that research outputs are reproducible: Ideally, running the same code on the same data leads to the same results, and a thorough description of the data generation process, including the precise description of psychological experiment protocols, makes research replicable: re-implementing the same procedure should generate very similar data (Plesser, 2017).[3]

To advance the state-of-the-art in data-based music research, *Empirical Musicology Review* is devoting this special issue to a wide discussion of questions related to Open Science, Open Data, and the FAIR principles.

The submissions address these issues from various angles and perspectives and showcase how the issues mentioned above have impacted on the creation and analysis of data for particular research purposes. All datasets that are discussed in the articles as well as those introduced in the new submission type *Data Reports* (see below) are published in freely-accessible data repositories such as *Open Science Framework*[4] or *GitHub*[5], or made available through institutional websites.

SUMMARY OF THE ARTICLES

Alexander Refsum Jensenius's article "Best versus Good Enough Practices for Open Music Research" addresses specific challenges music researchers are inevitably confronted with when pursuing an open science agenda, for instance the handling of (multi)media files, privacy, and copyright issues, and proposes to aim for a compromise between perfect adherence to the FAIR principles and feasibility.

David M. Weigl et al. describe in "FAIR Interconnection and Enrichment of Public-Domain Music Resources on the Web" how techniques from Music Information Retrieval are used in the context of the TROMPA project (Towards Richer Online Music Public-domain Archives) to interconnect and enrich music repositories in the public domain, in particular referring to the challenges posed by the FAIR principles.

Mark Gotham's think-piece, entitled "Connecting the Dots", discusses key questions related to the openness of musicological research on several levels (score encodings, analysis encodings etc.), illustrating them by examples from a wider range of his own projects (scores, analyses, and pedagogical material available in digital formats).

In "Enabling FAIR use of Ethnomusicology Data," Alex Hofmann et al. address the issue of lacking standards for ethnomusicological research and discuss how the FAIR principles can be applied in order to close this gap. They propose a number of action items towards a better integration and linkage of existing and future resources. The corresponding commentary by Stefan Münnich critically discusses the notion of 'ethnomusicology' and some of its underlying (Western-centered) biases, pointing out that it is imperative not to reproduce them when moving from 'traditional' to 'digital' ethnomusicology.

The article "The Interpersonal Entrainment in Music Performance Data Collection" by Martin Clayton et al. presents an extensive collection of six different but related sources from diverse cultural backgrounds for the empirical study of entrainment in music. In particular, it provides audio and video recordings and computationally extracted onset timing that can be used for within- and cross-cultural performance analysis.

Ajay Srinivasamurthy et al. introduce "Saraga: Open Datasets for Research on Indian Art Music," two datasets of Hindustani and Carnatic music within the larger *CompMusic* data collection. They contain (partially multi-track) audio recordings, metadata, and annotations of several musical features that may serve as ground truth for a variety of Music Information Retrieval tasks. Lara Pearson comments on this article in "Cultural Specificities in Carnatic and Hindustani Music: Commentary on the Saraga Open Dataset" by discussing some of its strengths and potential points of improvement, in particular regarding the representativity of this corpus.

DATA REPORTS: A NEW SUBMISSION FORMAT

Starting with this special issue, EMR is introducing a new section on Data Reports. To acknowledge the scientific effort and value of creating, cleaning, curating, enabling access, and maintaining data, EMR invites researchers to share their datasets under the general philosophy of the FAIR principles. In general, Data Reports may describe a variety of datasets such as musical metadata, annotations of musical corpora in symbolic or audio formats, automatically extracted musical features, data from psychological experiments etc. The Data Reports published in this special issue cover a wide range of areas within empirical music research, such as melody, microtiming, harmony, and emotion.

Anna Aljanaki et al. introduce the "Multitrack Contrapuntal Music Archive" as a symbolic dataset of independent voices in polyphonic (Baroque) pieces. "The MeloSol Corpus" by David Baker presents transcriptions of Western melodies from a sight-singing textbook. Christopher White puts this in a broader

context and compares Baker's corpus with a number of similar datasets. In their data report on "Drum Groove Corpora", Fred Hosken et al. introduce three datasets of onset timings in drum-kit performances. The "Recorded Brahms Corpus" by Ana Llorens provides encodings of note and beat onsets, durations, and tempo fluctuations in several performances of Brahms's Cello Sonatas. In their "The Mozart Expositional Punctuation Corpus," Omer Raz et al. add a dataset of cadence annotations in Mozart's works based on the punctuation theory of 18th-century theorist Heinrich Christoph Koch. Ben Duane, in his commentary on the Mozart dataset, raises a number of critical conceptual issues (such as the notion that closure is brought about by a process rather than a single event) and uses the corpus to test a particular research hypothesis, namely that cadence instances that qualify as structural cadences in Koch's sense come as a series of events at increasingly shorter distances. Finally, "The PUMS Database" by Lindsay Warrenburg provides a systematic survey of stimuli in music emotion research of the last 90 years.

TOWARDS MORE OPEN MUSIC RESEARCH

Since its inaugural issue from 2006, *Empirical Musicology Review* has been dedicated to fostering open science and scholarly debate. As the editors of this special issue on "Open Science in Musicology," we want to express our hope that our research community steadily moves forward towards a more open and FAIR research practice and engages itself in an active debate about the applicability and feasibility of the requirements specific to music research. The articles in this collection may be understood as a sample of the diversity that exists in our field in terms of viable approaches to reproducible research.

It is our conviction that the major scientific societies and communities as well as the funding bodies can play a crucial role in the coordination and further development of domain-specific standards of open science on a larger scale. This can, for instance, also imply specific institutional measurements, such as the appointment of data stewards or similar roles who lead or give advice in this endeavor. The organization of specific training sessions and workshops would also likely raise the awareness of these issues and, at the same time, would bring about an increasing consolidation among research practices. Similarly, scholarly journals play an important role, as they may want to (re-)consider the open access options they offer as well as the publication fees they charge (such that they do not disadvantage less wealthy institutions and research groups). This special issue offers a modest contribution on this path, and we hope that it will inspire and invigorate discussions towards more open music research.

ACKNOWLEDGEMENTS

We thank EMR editors Daniel Shanahan and Daniel Müllensiefen for their continuous support during the curation of this special issue, the copy editors Christine Ahrends, Lottie Anstee, Gabriele Cecchetti, Annaliese Micallef Grimaud, Matthew Moore, and Jessica Pinkham, and layout editor Diana Kayser, as well as all authors and reviewers for their contributions.

NOTES

[1] Correspondence can be addressed to: Fabian C. Moss, École Polytechnique Fédérale de Lausanne, Digital Humanities Institute, Digital and Cognitive Musicology Lab CH-1015 Lausanne, Switzerland, Email: fabian.moss@epfl.ch.

[2] See also the blog posts "Reproducible research in systematic musicology" (Joshua Bamford; https://sites.google.com/view/sysmus/blog#h.p_moS4C1lO3W-y) and "Open Data in Music and Science" (Tuomas Eerola; https://musicscience.net/2018/05/25/open-data-in-music-and-science/); both accessed on 27 January, 2021.

[3] See also Bittner et al. (2019), McFee et al. (2019), and the tutorial on "Open Source and Reproducible MIR Research" (http://ismir2018.ircam.fr/pages/events-tutorial-14.html; McFee, 2018).

[4] https://osf.io

[5] https://github.com

REFERENCES

Bittner, R. M., Fuentes, M., Rubinstein, D., Jansson, A., Choi, K., & Kell, T. (2019). mirdata: Software for reproducible usage of datasets. In Flexer, A., Peeter, G., Urbano, J., & Volk, A. (Eds.): Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019 (pp. 99–106). Delft, The Netherlands: International Society for Music Information Retrieval Conference.

McFee, B., Kim, J. W., Cartwright, M., Salamon, J., Bittner, R., & Bello, J. P. (2019). Open-source practices for music signal processing research: Recommendations for transparent, sustainable, and reproducible audio research. IEEE Signal Processing Magazine, 36, 128–137. https://doi.org/10.1109/MSP.2018.2875349

Plesser, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. Frontiers in Neuroinformatics, 11, 76. https://doi.org/10.3389/fninf.2017.00076

Sugimoto, C. R., Ekbia, H. R., & Mattioli, M. (Eds.). (2016). Big Data is not a Monolith. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/10309.001.0001

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18