Best versus Good Enough Practices for Open Music Research

ALEXANDER REFSUM JENSENIUS[1]
RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, Department of Musicology, University of Oslo

ABSTRACT: Music researchers work with increasingly large and complex data sets. There are few established data handling practices in the field and several conceptual, technological, and practical challenges. Furthermore, many music researchers are not equipped for (or interested in) the craft of data storage, curation, and archiving. This paper discusses some of the particular challenges that empirical music researchers face when working towards Open Research practices: handling (1) (multi)media files, (2) privacy, and (3) copyright issues. These are exemplified through MusicLab, an event series focused on fostering openness in music research. It is argued that the "best practice" suggested by the FAIR principles is too demanding in many cases, but "good enough practice" may be within reach for many. A four-layer data handling "recipe" is suggested as concrete advice for achieving "good enough practice" in empirical music research.

Submitted 2020 April 30; accepted 2020 October 2.
Published 2021 December 10; https://doi.org/10.18061/emr.v16i1.7646

KEYWORDS: *Open Research, multimedia, privacy, copyright*

THE push for more openness in research has been ongoing for a long time and was originally driven by individual researchers, smaller research groups, and some progressive (sub)-disciplines. The transition has recently gained traction due to the political pressure from various initiatives, such as the funder-driven *Plan S* initiative (cOAlition S, 2018), the *Declaration on Research Assessment* (DORA, 2012), the project *Fostering the Practical Implementation of Open Science in Horizon 2020 and Beyond* (FOSTER, 2019), and others. Much of the recent attention has been devoted to solving the challenges of making more publications openly available, which is often called Open Access. The increased attention to open publications has sparked interest in opening up other parts of the research process as well. These are often discussed under the umbrella term Open Science. The *FOSTER* project suggests that this term includes the following main categories (Pontika et al., 2015): Open Access, Open Data, Open Reproducible Research, Open Science Evaluation, Open Science Policies, and Open Science Tools (see Figure 1 for an overview of the complete *FOSTER* taxonomy). While one may agree or disagree with the naming and organization of the individual points in the *FOSTER* taxonomy, it covers many aspects of today's academic life. In this article, the focus will be on the parts that fall under the Open Data category, and some of the subcategories of the categories Open Reproducible Research and Open Science Tools.

Before moving on, I will take a short terminological detour. This is because of a growing concern about the use of the term "Open Science", which I find to rule out the arts and humanities indirectly. I like to call myself a *music researcher* and a *research musician*, to highlight that my research activities fall into two (often overlapping) categories: knowledge-driven *and* art-driven research. Roughly speaking, one could say that the former is based on scientific questions and methods, while the latter is using artistic questions and methods. Consequently, my *scientific* research fits nicely into various definitions of Open Science, including the one presented by the *FOSTER* project (2019):

> "Open Science is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods."

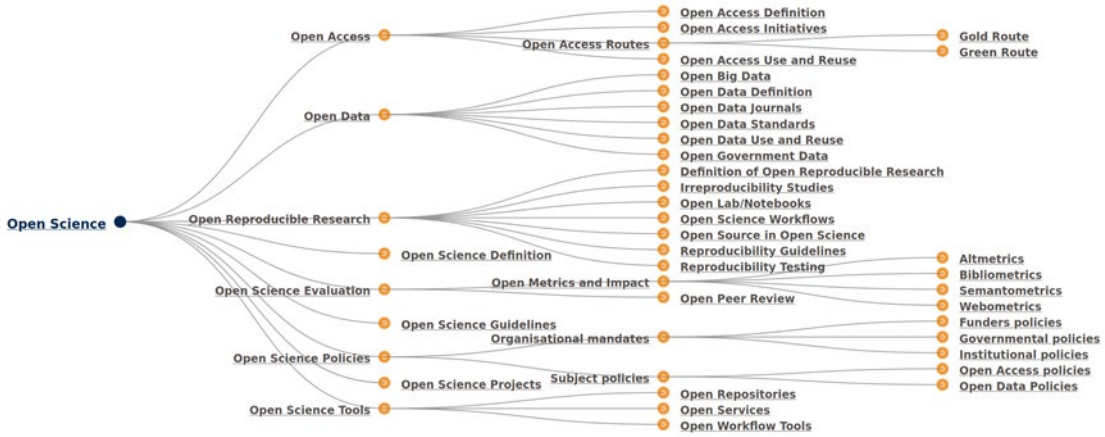

Fig. 1. *The FOSTER taxonomy of Open Science (FOSTER, 2019).*

This definition works less well for *artistic* research. While it was probably never the intention to leave out arts and humanities researchers, the definition mentioned above (and other similar ones using words like "science" and "lab") still feels alienating to many. This is unfortunate because the ideals of Open Science could and should extend to all parts of academia. I have yet to meet anyone who believes that Open Science is only meant for people working in the (hard) sciences. It just happened that the current push for openness began from some of those disciplines. In policy documents, there are attempts at broadening the definition by using sentences like: "… the sciences (including the arts and humanities) …". A better solution would be to use the term "Open Research" instead. This would feel more inclusive for arts and humanities researchers, and it would also include researchers who work outside academia. They, too, may be interested in opening their research, even though they would not call themselves "scientists". Therefore, I will continue to use the term Open Research in the following.

After that minor—but essential from a music research perspective—detour, let us consider some challenges faced when working openly in music research. We will begin by looking at the FAIR principles, which will be taken as a starting point for discussing "best practice" strategies. The argument will here be inspired by the papers *Best practices for scientific computing* (Wilson et al., 2014) and the follow-up paper *Good enough practices in scientific computing* (Wilson et al., 2017). Targeting "good enough practice" is not meant to avoid striving for the best solutions. However, it is an acceptance that if we should be able to do any *music* research at all, we cannot spend all the time developing the best data management solutions.

BEST PRACTICE

The FAIR principles have been suggested as a general "best practice" approach for making data openly available (Wilkinson et al., 2016). The four overarching principles can be summarized as:

- **Findable:** The first step in (re)using data is to find them. Data and their connected metadata should be easy to find for both humans and computers. Humans need to be able to read and understand what the data are data *of* to make use of them. Machine-readable metadata is essential for the automatic discovery of data sets and services.
- **Accessible:** Once the user finds the required data, s/he needs to know how they can be accessed. This may include information about the storage location, authentication, authorization, and licensing.
- **Interoperable:** The data usually need to be integrated with other data and need to interoperate with applications or workflows for analysis, storage, and processing.
- **Reusable:** The ultimate goal of FAIR is to optimize the reuse of data. Then the data and related metadata should be described so that they can be replicated and combined in different settings.

Each of these main principles has several specific sub-principles. They were developed to be general, so no technical implementation was specified. However, to ensure proper implementation, the principles have been followed up with a specification of how to interpret the recommendations (Jacobsen et al., 2020).

It is essential to point out that applying the FAIR principles to a data set is not the same as making the data *open*. Even though openness is an overarching goal, there is an understanding that it is necessary to make data as "open as possible, but as closed as necessary": a phrasing that can be found in many policy documents, for example, by the European Commission (2016). It is generally accepted that Open Data should be FAIR. After all, if the data cannot be found and accessed, they cannot be reused. The opposite is not necessarily the case: FAIR data does not need to be open. Still, the idea is that by applying the FAIR principles also to closed data and making the *metadata* openly available, it is possible to know what types of data exist and how to ask for access.

It should also be noted that making data FAIR is not the same as making them *free*. In open-source software development, freedom is related to running the program, studying and changing the code, and redistributing copies with or without changes. This is what is meant by "free as in speech", not only "free as in beer", as Stallman (2009) explains the difference between open and free software. This point can also be extended to other parts of the research chain beyond software development, including data. At the core of such freedom are copyright and licensing, two topics we will discuss more in a later section.

Even though much has happened over the last few years, we are still far from having solutions in place to share data according to the FAIR principles. On the positive side, an increasing number of general "bucket-based" repositories may be used for archiving data, such as *Zenodo*, *Open Science Framework*, *Dryad*, and *Figshare*, to mention only a few in no specific order. Such repositories can store anything of (more or less) any size. They also provide mechanisms and tools for creating unique identifiers and version control of the data. While such centralized repository systems may be the norm, there are also developments of semantically-enabled (Mitchell et al., 2011) and decentralized (Capadisli, 2020) solutions for data archiving and sharing.

A challenge with general-purpose solutions, such as the ones mentioned above, is the lack of specific metadata, specialized tools, and a community to support the long-term data curation. This is typically done within field-specific archives. In music research—here conceived broadly—some subfields are closer to having systems that comply with FAIR principles than others. Some historical or ethnographic music collections have well-organized databases with proper data curation. These have often been developed within or in close contact with libraries that could secure professional metadata handling. Music Information Retrieval is a sub-field that has been actively involved in developing solutions, as exemplified by the community around the *International Society for Music Information Retrieval* (ISMIR). The ISMIR community maintains a resource for sharing data sets and has pushed for more standardization through the annual MIREX competitions. However, even within this highly knowledgeable sub-field, many datasets do not have FAIR compliance.

One could argue that some of the most "FAIRified" music data sets are those based on symbolic music representations, such as MIDI, MEI, and MusicXML files. One example is the *Essen Folk Song Collection* (Schaffrath, 1995), which has been used by many music researchers over the years. In audio-based music research, there have been several initiatives to help create open data sets, of which *Freesound* is one of the larger and more popular ones (Fonseca et al., 2017). This database is based on crowd-sourcing and contains both musical and non-musical sound. There are also initiatives like *AcousticBrainz* (Porter et al., 2015) and the *Million Song Dataset* (Bertin-Mahieux et al., 2011) that aim to provide extensive metadata on commercially available music. Finally, there are initiatives aimed at collecting other types of information about the music, such as the *MusicBrainz* database (Swartz, 2002). This includes information about the artists, the recording location and time, and so on.

Even though more music-related datasets are made available online, this does not necessarily mean that they are readily usable. In fact, Daquino et al. (2017) show that most score low on interoperability, defined according to the 5 Star Open Data paradigm:

- **Open License (OL):** accessible with an open license (CC-BY, CC0, etc.).
- **Machine Readable (RE):** contains structured data published in a machine-readable format.
- **Open Format (OF):** published in an open standard.
- **URIs (URI):** makes use of Uniform Resource Identifiers (URIs) to identify entities.
- **Linked Data (LD):** published in the Resource Description Framework (RDF).

My scientific research is primarily in the field of embodied music cognition. Here studying people's (bodily) responses to music is at the core, which involves a lot of different data collection methods. Here—and within the field of music psychology at large, I would argue—there are not many openly available data sets, and I cannot recall any that would receive five stars according to the criteria mentioned above. This is not because of an unwillingness to share data among researchers. However, many challenges pile up when one tries to conduct this type of research openly. In the following, some of these will be described in the context of the ongoing innovation project *MusicLab*.

AN ATTEMPT AT OPENING EMBODIED MUSIC COGNITION RESEARCH: THE CASE OF *MUSICLAB*

MusicLab is an innovation project between the RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, and the University Library at the University of Oslo. The aim is to explore new ways of conducting research, research communication, and education within empirical music research. The project is organized around events: a concert in a public venue, which is also the study object. The events also contain an "edutainment" element through panel discussions with researchers and artists, and "data jockeying" in the form of live data analysis of recorded data. There are three aims of *MusicLab*:

- **Research:** The fundamental question is: How do people engage with music? This is investigated by studying the role of the body in music performance and perception using both motion capture and psychophysiological measurements (such as breathing and pulse).
- **Dissemination:** Another aspect of *MusicLab* is the exploration of new research dissemination strategies. For most people, research is something that happens behind closed doors at a university. We perform research in the "field", talk about the process, and show what we find.
- **Innovation:** *MusicLab* is also a test-bed for developing Open Research practice in music research. Rather than keeping the entire research process closed—and only opening up at the time of publication—we want to invite everyone to help with the data collection and analysis. We also want to share the data with the public and show how it can be analyzed.

After running a series of *MusicLab* events and trying to "FAIRify" the data, three topics stand out as particularly challenging: (a) the multimedia and multimodal nature of the data themselves, (b) privacy issues, (c) copyright issues. Not all of these are directly related to FAIR, but they are still obstacles that need to be solved to be able to work openly with the data we collect. In the following sections, the topics will be discussed, and some solutions presented.

Data types and their challenges

Representation of data is the first challenge we have had to deal with in *MusicLab*. This is a topic that does not receive much attention in general FAIR discussions, usually because it is a question about the technical equipment used for data capture and storage. However, it is a point of concern for researchers who have to juggle different types of technologies in real-world data collection scenarios. Without thinking carefully about data types and formats, it is not possible to achieve FAIR compliance. At *MusicLab*, we collect many different data types:

- **Scores:** original scores or adapted scores, possibly containing annotations from multiple performers. They are sometimes collected as scanned papers; other times, a symbolic representation may be available (MIDI or MusicXML).
- **Audio:** multichannel recordings from one device or multiple recordings from different devices. These are usually recorded in raw format and with the same bit rate and sampling frequency. In addition comes compressed audio embedded within video files (MP3, AAC).
- **Video:** recordings of a multi-camera production stream and separate recordings from multiple video cameras. These are usually recorded with the best format offered by the camera in use (MPEG-2, MPEG-4, MOV, AVI), and with different types of compression (MJPEG, H.264), resolution (HD, Full HD, 4K), and frame rate (25-60fps).

- **Motion capture:** typically recorded within a hardware-specific software platform using a proprietary data format and different sampling rates.
- **Sensors:** usually a custom-built system without any particular specification or data format, often recorded as simple text files with a time-code (if available)
- **Questionnaires:** both standard questionnaires and custom-developed, either collected on paper or using a web form.
- **Interviews:** often recorded and annotated, in addition to notes taken by the interviewer.

As can be seen, such a data set is highly complex and multidimensional. There are always several people involved in data collection and many recording devices. In the beginning, we aspired to have a streamlined data collection process, with a centralized server solution, proper time synchronization, and standard file formats. However, since one of the aims of *MusicLab* is to work outside the lab in real-world venues, we have realized that we will have to work with a combination of our own and the venue's equipment. So we have learned to deal with the technological complexity and instead try to develop good routines for centralizing and cleaning the data after the data collection.

When it comes to the audio and video data, it is easy to agree on well-functioning data standards and file types. Still, it is necessary to decide on whether one should use uncompressed formats, such as raw video in AVI containers and AIFF/WAV/FLAC for audio files, or some kind of compression (MJPEG or H.264 for video and MP3, AAC, or OGG for audio). Unless they are stored together in one file, synchronizing audio and video files is not straightforward, either. Storing score information is equally challenging. Many people rely on MIDI-based software, so using MusicXML may not always be an option. Furthermore, linking scores to audio/video recordings is a non-trivial task (Raphael, 2006), although there are successful examples of interoperability using RDF and Linked Data (Meroño-Peñuela et al., 2017).

Things get even more complicated when motion capture and sensor data are added to the mix. Some formats are more interoperable than others, such as C3D (Motion Lab Systems, 2008). However, there are no standards for handling various types of add-on sensors and custom-built systems, and there is no simple way to ensure proper synchronization with audio and video files. More than a decade ago, I proposed the *Gesture Description Interchange Format* (GDIF) as a way to overcome some of these problems (Jensenius, Kvifte & Godøy, 2006). The proposal never went beyond the prototype stage (Jensenius, 2013), mainly due to the complexity of the question. There are also various types of standards and code-books in use for questionnaires and interviews, but these are typically field-specific, and there are few examples developed for our types of research.

Another challenge is that of synchronizing the data. For some researchers, ensuring that the files start and stop at the same time may be sufficient. For others, particularly those looking into musical timing details, frame-based synchronization is the only thing that can be tolerated. However, frame-based synchronization is non-trivial when you are handling data with different types of timestamps, sampling frequencies, and so on.

There currently exists no solution for handling the complexity of the data sets that we collect. The solution that comes closest is the multimodal online database solution *RepoVizz*, which can handle synchronization and display a multitude of signal-based data types (Mayor, Llop, and Maestre, 2011). Some exciting developments are happening within the *TROMPA* project (Weigl et al., 2019), although this is primarily targeted at combining metadata from libraries and collections. One could wish for a massive investment from an external funder that could help solve the problem. However, that is highly unlikely given that the number of empirical music researchers with our type of specialized needs is so small.

Our solution at the moment, then, is to not aim for a "best practice" solution, but instead, opt for a "good enough" solution. This involves aiming to use as many standard and open file formats as possible and develop sufficiently efficient and realistically complex data management routines.

Privacy

The second major issue we have faced in the case of *MusicLab* is handling privacy matters according to legal and ethical standards. With stricter regulations, such as EU's General Data Protection Regulation (GDPR), privacy issues have surfaced in many ethics committees over the last few years.

What we have found particularly challenging is the requirement of participants to give so-called *dynamic consent*. This means that a participant should be able to revoke a given consent at any point in time

and for any reason. This poses some challenges when it comes to storing information about participants in such a way that it would be possible for them to revoke the consent. Names are not unique, and people's mobile phone numbers and e-mail addresses usually change over time. People's personal ID would be the best way to store the information. However, in Norway, the ID number is considered sensitive data, and collecting that would mean going to higher security on the data. This would severely limit the way the data can be handled and used. Also, one of the points of *MusicLab* is to study how audiences (large groups of people) respond to music. It is just not feasible to collect sensitive information from all audience members.

In general, we are mainly interested in collecting anonymous data. However, in many cases, it is impossible to anonymize research data completely. Even a motion capture-based stick figure representation could probably be used to detect people using machine learning techniques. Video files may be "anonymized" by blurring out faces, but a person may still be recognized by the motion of the rest of the body. It is also important to consider that facial information may be crucial for researching emotion and affect.

To overcome some of the privacy challenges, we have at *MusicLab* ended up dividing people present into three groups:

- **Performers:** These people wear sensors and are filmed and captured in many different ways. They sign a written consent with a detailed description of how the data collected is stored and processed.
- **Volunteers:** These wear different types of wearable sensors and reply to questionnaires. All of this information is anonymous. They give written consent when they volunteer.
- **Audience:** These will be filmed from a distance, as part of the larger audience group. The filming is announced in the advertisements for the event, on posters in the venue, and orally from the stage before the concert starts. They do not sign a consent form but are free to walk out if they do not consent to be filmed.

This division into groups simplifies the data collection and data handling. We have also found that it makes it easier to communicate the data processing to everyone present.

One of the aims of *MusicLab* is to make as much data as possible openly available. This ambition does not easily match the GDPR requirement of dynamic consent. It is, in practice, impossible to remove any data that has been made openly available. Of course, we can remove data from our server and the online repositories that we have used. However, we cannot track who has downloaded and stored the files on local computers. This is not an ideal solution, but it is difficult to imagine how it would be possible to combine an Open Research perspective of this type of data and also offer the possibility to revoke consent.

Copyright and Licensing

The third major challenge we have encountered is related to copyright and licensing. To share a data set openly—and in such a way that it is possible to do any meaningful musical analysis on the material—it is necessary also to share the musical material (score and audio) openly and with a permissive license. This is not straightforward. There may be many copyright holders involved: composers, lyricists, performers, record labels, and so on. Also, the metadata in music databases may be incorrect: names may be misspelled, and field names in databases differ (Deahl, 2019). Besides, the audio track is, in many cases, considered “synchronised” when it is paired up with a video recording. In some countries, including Norway, some corporations administer collective copyright agreements for artists. In other countries, it is necessary to negotiate with the individual copyright holders. Finally, there is the challenge of storing provenance information systematically. Here the W3C PROV specification (Grith, 2013) is a solution that can help in structuring information about the legal matters according to FAIR principles. However, this requires that the information is readily available, and it does not solve any of the legal aspects.

New copyright laws, including the new EU Copyright Directive, strengthen the copyright of artists. This is a positive development. However, the stricter regulation of music copyright, coming from the cultural sector, is not aligned with the push for more openness of the research and innovation sector. Music researchers work around copyright limitations in different ways. The *AcousticBrainz* database (Porter et al., 2015) and the *Million Song Dataset* (Bertin-Mahieux et al., 2011) solve it by only storing and sharing metadata about the music. This may work in some ways, and particularly for commercially available music. However, since the databases only provide metadata, it works better for machine-based analysis approaches. For more detailed listening-based analyses and dissemination, the audio track is needed.

Another solution is to only work with music that has been copyright-cleared and given an open license, such as one of the Creative Commons licenses (Lessig, 2004). Initiatives such as *FreeSound* (Fonseca et al., 2017) and *Audio Commons* (Font et al., 2016) fall under this category. However, the challenge with this approach is that music research may be skewed towards studying only a (for now) small subset of musical material. It will also leave out a large part of commercially available music. There is a need for the research and cultural sectors to sit down together and figure out how music research can be done in a way that both adheres to the ideals of Open Research and supports the income of artists.

Recommendations for "Good Enough Practice"

As may be clear by now, I am in no position to offer concrete solutions to any of the challenges mentioned above. The legal issues of privacy and copyright are, obviously, far from what music researchers can solve. Nevertheless, it may help to raise our voice and let others know that this is a concrete problem that we are facing.

The question of handling multimodal and multimedia data is also far too complicated for the average music researcher to grasp. The FAIR principles or the advice presented in the paper *Best practices for scientific computing* (Wilson et al., 2014) are at a level beyond what most empirical music researchers can implement. Still, there is some good advice to learn from. I particularly like the first recommendation: "Write programs for people, not computers". This is excellent advice also for data. The data will be much more accessible if people can understand what they contain. It is vital to work towards machine-readable data as well, but in most cases, no machine will be able to read the data if no human found it first. Another good recommendation by Wilson et al. (2014) is "Let the computer do the work". More music researchers should learn to code. This would help processing in many cases, and it would also make it easier to work towards the reproducibility of our findings.

Perhaps because their first paper was perceived important, yet unreachable for most researchers, Wilson et al. (2017) wrote a follow-up paper called *Good enough practices in scientific computing*. Here the idea was to present some advice that would be possible to follow. Inspired by their list, I have come up with some recommendations of my own. These are based on our experiences at *MusicLab*, and focus less on software development and more on the complexity that music researchers face. My recommendations can be thought of using a kitchen-oriented metaphor, thinking about four steps of cooking: raw, processed, cooked, and preserved.

1. DATA COLLECTION ("RAW")

1a. Create analysis-friendly data. Planning what to record will save time afterward, and will probably lead to better results in the long run. Write a data management plan (DMP).
1b. Plan for mistakes. Things will always happen. Ensure redundancy in critical parts of the data collection chain.
1c. Save the raw data. In most cases, the raw data will be processed in different ways, and it may be necessary to go back to the start.
1d. Agree on a naming convention before recording. Cleaning up the names of files and folders after recording can be tedious. Get it right from the start instead. Use unique identifiers for all equipment (camera1, etc.), procedures (pre-questionnaire1, etc.) and participants (a001, etc.).
1e. Make backups of everything as quickly as possible. Losing data is never fun, and particularly not the raw data.

2. DATA PRE-PROCESSING ("PROCESSED")

2a. Separate raw from processed data. Nothing is as problematic as over-writing the original data in the pre-processing phase. Make the raw data folder read-only once it is organized.
2b. Use open and interoperable file formats. Often the raw data will be based on closed or proprietary formats. The data should be converted to interoperable file formats as early as possible.
2c. Give everything meaningful names. Nothing is as cryptic as 8-character abbreviations that nobody will understand. Document your naming convention.

3. DATA STORAGE ("COOKED")

3a. Organize files into folders. Creating a nested and hierarchical folder structure with meaningful names is a basic, but system-independent and future-proof solution. Even though search engines and machine learning improve, it helps to have a structured organizational approach in the first place.
3b. Make incremental changes. It may be tempting to save the last processed version of your data, but it may be impossible to go back to make corrections or verify the process.
3c. Record all the steps used to process data. This can be in a text file describing the steps taken. If working with GUI-based software, be careful to note down details about the software version, and possibly include screenshots of settings. If working with scripts, document the scripts carefully, so that others can understand them several years from now. If using a code repository (recommended), store current snapshots of the scripts with the data. This makes it possible to validate the analysis.

4. DATA ARCHIVE ("PRESERVED")

4a. Always submit data with manuscripts. Publications based on data should be considered incomplete if the data is not accessible in such a way that it is possible to evaluate the analysis and claims in the paper.
4b. Submit the data to a repository. To ensure the long-term preservation of your data, also independently of publications, it should be uploaded to a reputable DOI-issuing repository so that others can access and cite it.
4c. Let people know about the data. Data collection is time-consuming, and in general, most data is under-analyzed. More data should be analyzed more than once.
4d. Put a license on the data. This should ideally be an open and permissive license (such as those suggested by Creative Commons). However, even if using a closed license, it is important to clearly label the data in a way so that others can understand how to use them.

Many of these points may be seen as obvious. Nevertheless, in my experience, it is remarkable how easy it is to skip many of them when time is limited, and there are many people involved. Agreeing on conventions and workflows has dramatically improved the way we work within the *MusicLab* project, and it has also helped in our lab-based data collection and storage.

The growing attention to developing Open Research practices in general, and the intensification of implementing the FAIR principles more specifically, helps in many ways. While there may not be many specialized resources for music researchers, there are a growing number of general-purpose solutions that can also work for music research. Many libraries hire data stewards and curators that can assist with data handling and archiving. They may not know anything about music research, but they may appreciate the many challenges that music researchers bring to the table. So a piece of final general advice could be: use existing institutional tools and seek help. In our experience, it has helped to work with professional data stewards and curators. They have helped to improve the data structures and metadata, thereby getting closer to true interoperability.

CONCLUSIONS

As I have tried to explain in this paper, empirically-based music researchers have a hard time when it comes to working towards FAIR and Open data practices. This is related to the complexity of the data at hand and the challenges of handling privacy and copyright. The number of empirical music researchers is comparably small, and the external interest is not exceptionally high. Therefore, it is essential to join forces and work as a community towards solving some of the challenges mentioned above. It is unnecessary to reinvent the wheel in many cases, but, to continue the analogy, it may be necessary to find the right combination of vehicles, tires, and screws to make the wheel fit.

However, the most important point is not to let the best be the enemy of the good. We are still far from making all our data FAIR, and we may perhaps never get there. A dream scenario would be a system combining the multimedia storage and visualization tools from *RepoVizz* with the inter-connectivity of *TROMPA*, the CC-spirit of *Audio Commons*, the versioning of *GitHub*, the accessibility and community of

Wikipedia, and the long-term archiving of *Zenodo*. While that may sound far-fetched right now, it could be a reality with some more interoperability. However, to do any *music* research at all, and not delve entirely into data and library science, we may need to accept that a "good enough practice" may suffice.

ACKNOWLEDGMENTS

This work was partially supported by the Research Council of Norway through its Centres of Excellence scheme, project numbers 262762 and 250698. This article was copyedited by Lottie Anstee and layout edited by Diana Kayser.

NOTES

[1] Correspondence can be addressed to: Alexander Refsum Jensenius, University of Oslo, P.O. Box 1133 Blindern, 0318 Oslo, Norway, a.r.jensenius@imv.uio.no.

REFERENCES

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The million song dataset. *Proceedings of the 12th International Society for Music Information Retrieval Conference*. Canada: ISMIR. https://doi.org/10.7916/D8NZ8J07

Capadisli, S. (2019). *Linked Research on the Decentralised Web.* PhD thesis, Rheinischen Friedrich-Wilhelms-Universität. Bonn, Germany. Retrieved from: https://csarven.ca/linked-research-decentralised-web

cOAlition S. (2018). "Plan S" and "cOAlition S" – accelerating the transition to full and immediate Open Access to scientific publications. Retrieved from: https://www.coalition-s.org/plan_s_principles/

Daquino, M., Daga, E., d'Aquin, M., Gangemi, A., Holland, S., Laney, R., Penuela, A. M., & Mulholland, P. (2017). Characterizing the landscape of musical data on the Web: State of the art and challenges. In *Proceedings of the 2nd Workshop on Humanities in the Semantic web*. Vienna, Austria.

Deahl, D. (2019). Metadata is the biggest little problem plaguing the music industry. Retrieved from: https://www.theverge.com/2019/5/29/18531476/music-industry-song-royalties-metadata-credit-problems

DORA. (2012). *San Francisco Declaration on Research Assessment (DORA)*. Retrieved on 6 July 2021 from: https://sfdora.org/read/

European Commission. (2016). *H2020 programme—Guidelines on FAIR data management in Horizon 2020* (v3.0). European Commission Directorate-General for Research & Innovation. Retrieved from: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Fonseca, E., Pons Puig, J., Favory, X., Font Corbera, F., Bogdanov, D., Ferraro, A., Oramas, S., Porter, A., & Serra, X. (2017). Freesound datasets: A platform for the creation of open audio datasets. *Proceedings of the 18th International Society for Music Information Retrieval Conference*. Canada: ISMIR.

Font, F., Brookes, T., Fazekas, G., Guerber, M., La Burthe, A., Plans, D., Plumbley, M. D., Shaashua, M., Wang, W., & Serra, X. (2016). Audio commons: Bringing creative commons audio content to the creative industries. In D. Murphy (Ed.) *Proceedings of the 61st International Conference: Audio for Games*. New York, NY: Audio Engineering Society.

FOSTER. (2019). *FOSTER – The Future of Science is Open*. Retrieved on 6 July 2021 from: https://www.fosteropenscience.eu/

Groth, P., & Moreau, L. (2013). *PROV-Overview: W3C Working Group Note.* Retrieved on 6 July 2021 from: https://www.w3.org/TR/prov-overview/

ISMIR. (2020). *ISMIR Resources: Datasets*. Retrieved on 6 July 2021 from: https://ismir.net/resources/datasets/

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., & Evelo, C. T. (2020). FAIR principles: Interpretations and implementation considerations. *Data Intelligence*, *2*, 11–29. https://doi.org/10.1162/dint_r_00024

Jensenius, A. R. (2013). Structuring music-related movements. In J. Steyn (Ed.), *Structuring Music Through Markup Language: Designs and Architectures* (pp. 135–155). Hershey, PA: IGI. https://doi.org/10.4018/978-1-4666-2497-9.ch007

Jensenius, A. R., Kvifte, T., & Godøy, R. I. (2006). Towards a gesture description interchange format. *Proceedings of the 6th International Conference on New Interfaces for Musical Expression*, 176–179. Paris, France: IRCAM.

Lessig, L. (2004). The creative commons. *Montana Law Review*, *65*(1), 1–13. https://scholarship.law.umt.edu/mlr/vol65/iss1/1

Mayor, O., Llop, J., & Maestre, E. (2011). RepoVizz: A multimodal on-line database and browsing tool for music performance research. *Proceedings of the 12th International Society for Music Information Retrieval Conference*. Canada: ISMIR.

Meroño-Peñuela, A., Hoekstra, R., Gangemi, A., Bloem, P., de Valk, R., Stringer, B., Janssen, B., de Boer, V., Allik, A., & Schlobach, S. (2017). The MIDI linked data cloud. *International Semantic Web Conference*, 156–164. https://doi.org/10.1007/978-3-319-68204-4_16

Mitchell, S., Chen, S., Ahmed, M., Lowe, B., Markes, P., Rejack, N., Corson-Rikert, J., He, B., & Ding, Y. (2011). The VIVO ontology: Enabling networking of scientists. *WebSci '11: Proceedings of the 3rd International Web Science Conference*. New York, NY: Association for Computing Machinery.

Pontika, N., Knoth, P., Cancellieri, M., & Pearce, S. (2015). Fostering open science to research using a taxonomy and an eLearning portal. *Proceedings of the 15th International Conference on Knowledge Technologies and Data-Driven Business, i-KNOW '15*(11), 1–8. https://doi.org/10.1145/2809563.2809571

Porter, A., Bogdanov, D., Kaye, R., Tsukanov, R., & Serra, X. (2015). Acousticbrainz: a community platform for gathering music information obtained from audio. *16th International Society for Music Information Retrieval Conference*. Canada: ISMIR.

Raphael, C. (2006). Aligning music audio with symbolic scores using a hybrid graphical model. *Machine Learning*, *65*(2–3), 389–409. https://doi.org/10.1007/s10994-006-8415-3

Schaffrath, H. (1995). *The Essen Folksong Collection*. Retrieved from: http://essen.themefinder.org/

Stallman, R. (2009). Why "open source" misses the point of free software. *Communications of the ACM*, *52*(6), 31–33. https://doi.org/10.1145/1516046.1516058

Swartz, A. (2002). Musicbrainz: A semantic web service. *IEEE Intelligent Systems*, *17*(1), 76–77. https://doi.org/10.1109/5254.988466

Weigl, D., Goebl, W., Crawford, T., Gkiokas, A., Gutierrez, N. F., Porter, A., Santos, P., Karreman, C., Vroomen, I., Liem, C. C. S., Sarasúa, Á., van Tilburg, M. (2019). Interweaving and enriching digital music collections for scholarship, performance, and enjoyment. In D. Rizo (Ed.), *DLfM '19: 6th International Conference on Digital Libraries for Musicology* (pp. 84–88). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3358664.3358666

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … 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

Wilson, G., Aruliah, D. A., Brown, C. T., Hong, N. P. C., Davis, M., Guy, R. T., Haddock, S. H. D., Huff, K. D., Mitchell, I. M., Plumbley, M. D., Waugh, B., White, E. P., & Wilson, P. (2014). Best practices for scientific computing. *PLOS Biology*, *12*(1), e1001745. https://doi.org/10.1371/journal.pbio.1001745

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLOS Computational Biology*, *13*(6), e1005510. https://doi.org/10.1371/journal.pcbi.1005510