Enabling FAIR use of Ethnomusicology Data – Through Distributed Repositories, Linked Data and Music Information Retrieval

ALEX HOFMANN[1]
University of Music and Performing Arts Vienna, Austria

TOMASZ MIKSA
Technische Universität Wien, Austria

PETER KNEES
Technische Universität Wien, Austria

ASZTRIK BAKOS
Technische Universität Wien, Austria

HANDE SAĞLAM
University of Music and Performing Arts Vienna, Austria

ARDIAN AHMEDAJA
University of Music and Performing Arts Vienna, Austria

BOONSIT YIMWADSANA
Mahidol University, Thailand

CLARE CHAN
Universiti Pendidikan Sultan Idris, Malaysia

ANDREAS RAUBER
Technische Universität Wien, Austria

ABSTRACT: Recordings of musical practices are kept in various public institutions and private depositories around the world. They constitute valuable data for ethnomusicological research and are substantial for the world's musical heritage. At the moment, there are no commonly used systems and standards for organizing, describing or categorizing these data, which makes their use difficult. In this paper, we discuss the required steps to make them findable, accessible, interoperable and reusable (FAIR), and outline action items to reach these goals. We show solutions that help researchers to manage their data over the whole research lifecycle and discuss the benefits of combining technologies from information science, music information retrieval, and linked data, with the aim of giving incentives for the ethnomusicology research community to actively participate in these developments in the future.

Submitted 2020 April 23; accepted 2020 October 13.
Published 2021 December 10; https://doi.org/10.18061/emr.v16i1.7632

KEYWORDS: *ethnomusicology, research data, FAIR principles, music information retrieval, linked data*

1. INTRODUCTION

SINCE Jesse Walter Fewkes's (1890) first phonograph recordings of Native Americans, the amount of ethnomusicological data has grown enormously across a wide range of storage media. Research in ethnomusicology today encompasses all the music of the world in their various social and cultural contexts (Merriam, 1960; Myers, 1992; Nettl, 2005). Research data in ethnomusicology is linked to qualitative and quantitative variables about the musicians, the music they make, the tools they use and the contents and contexts of the musical performances. Here it is the cooperation between the ethnomusicologists and the music-makers that results in the various forms of data that are stored in private or institutional archives. Direct personal contact with the data-owner is often the only way to find, access and interpret the data, while the reuse and sharing of data is often limited, problematic, or even impossible in this field. The current situation is a massive barrier for conducting computer-aided music research when there is a clear demand for computational ethnomusicology (Panteli et al., 2018).

Digital preservation, data management, and stewardship are highly discussed topics in many scientific domains. Researchers are advised to follow FAIR principles to make their data findable, accessible, interoperable, and reusable. Following FAIR principles (Wilkinson et al., 2016) also became an official requirement for research funders and the cornerstone of the European Open Science Cloud (EOSC) that facilitates the access and exchange of scientific data. Several solutions to help researchers manage their data over the whole research lifecycle are available: repositories hosted by trusted institutions store research data safely; research data policies regulate responsibilities of researchers and institutions; Data Management Plans act as awareness tools for researchers when collecting data (Miksa et al., 2019).

We believe that research in ethnomusicology can benefit from FAIR principles and the tools for data management. Application of FAIR principles to ethnomusicology will only be successful when the researchers see benefits for themselves, e.g., less work in managing data, better quality of metadata, prevention of data loss etc. We must also respect specific local customs and take cultural constraints and differences into account. For this reason, applying FAIR principles to such a heterogeneous domain is not a straightforward task. In this paper we describe the necessary steps to make ethnomusicology data FAIR. For each of the principles we discuss technical solutions and present how they can be customized to suit community needs. The discussion in this paper is based on two workshops and a panel discussion at the ICTM World Conference (Sağlam et al., 2019).

The paper is structured as follows. In Section 2 we give an overview of relevant state-of-the-art research and technology. We present our methodology in Section 3, where we compare current practices in data management and elaborate future use-cases for computer-aided music research. In Section 4 we describe advances needed to establish FAIR culture in ethnomusicology. Section 5 describes a proof-of-concept software implementation. Finally, in Sections 6 and 7 we discuss relevant aspects that may contribute to successfully developing joint practices and federated repository infrastructures.

2. RELATED WORK

This section discusses related research and technologies from information science, music information retrieval (MIR) and digital musicology.

Data repositories

Repositories play a key role in sharing and long-term preservation of research data and are therefore crucial for the promotion of Open Science. Many funders and publishers mandate or recommend to deposit research data underlying a publication in a suitable repository. Existing repositories can be found using services such as *re3data.org* or *opendoar.org*. Institutions willing to have their own repository can choose from open-source systems that must be installed and maintained by them. Examples of such systems are Dataverse[2], Invenio[3], and DSpace[4]. Each of the systems provides a website interface in which objects can be searched and accessed. Modes of access can be defined (open/shared/closed), and an API allows for automated exchange of metadata between systems (OAI-PMH protocol[5]). Thus, systems that aggregate information on objects in different repositories can be built.

Metadata

Together with the development of repositories, standards for describing objects with metadata were developed. Dublin Core[6] is a basic set of vocabulary terms that can be used to describe resources, such as books, CDs, and digital objects. Over the years, Dublin Core became a basis for newer standards, like Data Cite Metadata profile[7] that is used to describe resources for which a digital object identifier (DOI) is minted. All these standards contain a minimum set of fields, such as creator, title, etc., that are independent of the domain and medium they are describing. These fields are also relevant for the ethnomusicology domain. However, these standards can ensure only basic FAIRness of data. Domain-specific standards are needed to enable more sophisticated types of queries on specialized data repositories.

Linked Data in Musicology

Combining music collections with Linked Open Data (LOD) technology has gained increased attention (Raimond et al., 2007; Crawford et al., 2014; Pugin, 2015; Weigl and Page, 2017; Page et al., 2017; Pascua, 2018). This technology has the potential to connect single data points with other knowledge sources, and thus creates a Web of human- and machine-readable distributed information, also called Semantic Web (Berners-Lee et al., 2001; Shadbolt et al., 2006). Linking metadata of large music collections to other knowledge sources has been an aim of several musicological research projects such as DOREMUS (Lisena et al., 2018) or the Digital Music Lab (Abdallah et al., 2017). In the latter project, inconsistent metadata were identified to complicate context-related queries. With the aim to fill this gap, this research project is based on creating an alliance of different data holders from the beginning and aims at collaboratively finding domain-specific data descriptors that can be interlinked with other data sources.

Music information retrieval

Automatic analysis of larger audio collections is becoming an established practice in digital musicology. To this end, MIR-based tools are often applied and fine-tuned to the specifics of ethnomusicological data. The aim of MIR is to extract information directly from the audio material, symbolic representations, or any data related to these (Knees and Schedl, 2016). Extraction of information from the audio involves discrimination tasks (e.g., speech–music segmentation, genre, and instrument identification) but also aims to transcribe melodies, rhythms and harmony in music (Lerch, 2012; Müller 2015). Upon successful transcription, (semi-)automatic musicological analyses become feasible, such as pattern discovery, analysis of form, or composer detection. For an overview of the breadth of the field and types of analyses performed see Meredith (2016). For ethnomusicological field recordings a number of algorithms are currently available (e.g., Bozkurt et al. (2014), Marolt et al. (2019); see Panteli et al. (2018) for an overview). Nevertheless, automatic transcription of polyphonic music recordings still needs improvement (Holzapfel et al., 2019). Since many MIR algorithms depend on high-quality ground-truth training data, a joint effort of ethnomusicologists and MIR researchers can lead to technological improvements.

3. METHODOLOGY

In this section we present an overview of the activities we performed to identify domain-specific needs and to derive requirements for the solution presented in Section 4.

Overview

We organized two international workshops with 25 researchers from ethnomusicology and computer science, representing institutions from Thailand, Malaysia, Indonesia, Philippines, Latvia, and Austria. The aims of the first workshop, held in January 2019 at the Mahidol University (Bangkok), were 1) *to gather information on the current data holdings at research institutions and in various collections*, and 2) *to identify future use-cases for how ethnomusicologists want to interact with data repositories*.

In the following working phase, we started from the collected use-cases, and defined requirements for a system architecture that supports FAIR and open data in ethnomusicology. Furthermore, a first sketch for a prototype architecture was developed. For the second workshop (Summer 2019), we aimed to discuss

and validate the derived solution within the closed circle of project members, before presenting it to a wider audience. As a third step, we presented the results in the form of an open panel discussion at the 45th ICTM World Conference to a wider community to receive feedback and initiate an intense exchange on the topic (Sağlam et al., 2019).

Current data holdings and storage concepts

The data holdings presented by the members of the participating institutions contain several thousand hours of audio and video recordings with additional information such as texts and photos. The following data types were mentioned as parts of the collections:

- audio recordings in various formats e.g., old tape recordings, digitized tape recordings, original digital recordings
- videos (analogue, digital)
- additional materials like concert programs, CD/DVD booklets, flyers, sheet music and field notes (paper, digital scans)
- photos (analog, digital scans, digital photos)
- musical instrument collections

All institutions use different, customized systems to organize their data – either a local software database or an online database that provides public access via a web front-end. However, we identified no standardized vocabulary that was used to describe the data, and metadata descriptions were sometimes only available in the local language. Data descriptions ranged from being rather compact with only 11 fields[8] to broad with more than 30 fields[9].

Use-cases and requirements

The workshop participants were encouraged to formulate so-called user stories describing how they typically interact with data during their research phase. We grouped those into use-cases, from which we derive requirements for FAIR data management and music information retrieval applications supporting the research process in the future. The main use-cases and requirements stemming from them were:

R1. SECURE STORAGE AND EASY MANAGEMENT OF GATHERED RESEARCH DATA

The majority of ethnomusicologists keep their fieldwork-data on personal computers, USB sticks, etc. There is often no backup strategy in place which endangers the non-reproducible data to be lost forever. There is a need for a system to secure and preserve the data.

R2. CONTROLLED DATA ACCESS AND SHARING WITH COLLABORATORS AND CONTRIBUTORS

A common reason for not using any central services for storing data is a misconception that data stored to a repository will be open to everyone. It was crucial to clarify that FAIR data, or centrally managed data is not identical with open data. Ethnomusicologists want to be able to restrict access to themselves or to a group of trusted collaborators who obtained permission from them. They are willing to support FAIRness, but not necessarily openness.

R3. IMPORTING EXISTING COLLECTIONS

Ethnomusicologists understand the importance of proper data management and are willing to hand over their collections to a trusted repository that fulfills their requirements. It is expected that the process of importing can be substantially automated or assisted by an expert, hence not the sole responsibility of a researcher.

R4. SEARCH ACROSS DISTRIBUTED DATA COLLECTIONS

Workshop participants showed a great interest in being able to search among research (meta-)data of various research communities, research institutions, and individual researchers. They all agree that having access to more research data will undoubtedly enrich scientific knowledge and provide a wider view on different research discourses, research methodologies, and results.

R5. AUTOMATIC (AUDIO) DATA ANALYSIS FOR METADATA GENERATION

There was a consensus that describing and annotating the data collected in the field is one of the most time-consuming parts of work. The metadata for the fieldwork datasets can usually only be assigned by an expert, because to analyze this kind of material requires years of experience and knowledge on those specific music traditions. There are also technical types of metadata, e.g., music-speech segmentation, for which researchers would welcome a solution that automates this task.

The main takeaway from our activities was that the ethnomusicology research community needs to see the benefits of working with FAIR data for itself. Hence, a proposed software solution must not only fulfill the basic requirements for research data management, it must also allow for future extensions that support the researchers with integrated tools for data mining, artificial intelligence, and MIR. Table 1 summarizes the identified requirements and aligns how each requirement contributes to making ethnomusicology research data compliant with the FAIR principles.

4. PROPOSED SOLUTION

This section outlines the possibilities that are currently available in the domain of Information Science and gives a forecast of how software extensions may support researchers in the future, which may work as an incentive to the community to actively participate in these developments. We describe actions we identified as crucial for addressing the researchers' requirements. For each of the requirements, we provide **action items** for different stakeholders that help to make ethnomusicology data FAIR. There is no single system, standard, or solution that can fulfill all requirements of researchers in ethnomusicology and at the same time improve FAIRness of the data. Tackling the challenges requires advances and changes in different parts of a research lifecycle and at different levels, i.e., from technical and organizational to political and cultural.

Table 1: Identified use-cases and requirements formulated by ethnomusicology researchers, aligned to FAIR principles.

Requirements		**F**indable	**A**ccessible	**I**nteroperable	**R**eusable
Secure storage and easy management of gathered research data	R1	✓	✓		
Controlled data access and sharing with collaborators and contributors	R2		✓		
Clarity on the data rights for sharing and reuse	R2				✓
Importing existing collections	R3	✓			
Description of data using a standardized vocabulary, to search across distributed data collections	R4	✓		✓	✓
Automatic (audio) data analysis for metadata generation	R5	✓			

R1 Secure storage and easy management of gathered research data

Research data in ethnomusicology is often gathered under fieldwork conditions. These involve different environments such as everyday life spaces, e.g., at home, in community centers, in open air presentations.

To store this data securely, the captured material must immediately be stored in a professional repository system at the host institution, to avoid eventual data loss and data modification. A repository system must allow researchers to store and manage their data, since they have the most information about the original data and are most qualified to organize and provide metadata for it. The system must allow researchers to restrict access to data – that is, the system must not force researchers to disclose data to others if they do not wish to do so.

The role of the repository operator is to ensure that the data is properly backed up and preserved, as well as secured from unauthorized access. Managing the repository system should be a responsibility of qualified IT personnel. Thus, securing and storing research data is a joint responsibility of researchers and institutions supporting them.

To provide the researchers with a secure storage and easy management of their gathered datasets, the institutions can choose one of the existing open-source data repository systems (see Section 1). All systems provide the basic features for secure and structured data storage and provide a graphical user interface where researchers can store and describe their data via the web browser interface. Up to this point all data is only locally stored with the host institution that the researcher trusts and is not shared with any third party. Moreover, the structured form of data storage, and the tools provided by the repository system make the basis for publishing information about the data (findability) and for managing data access.

R1 Action items:

- **Institutions:** Provide data repositories that ensure data is properly backed up and managed
- **Researchers:** Secure the collected data by uploading the data into data repositories

Fig. 1. Action items for secure storage and easy management of gathered research data.

R2 Controlled data access and sharing with collaborators and contributors

In the workshops we identified that different rights-holders are involved with ethnomusicology data. As most data involves music recordings, special composer and performer rights may apply, which restricts some data from public sharing. Another reason is that some data were collected in the past and no explicit permission from the original creator exists to share the research data publicly. However, in most cases metadata are not affected by these restrictions and may be of high interest for other researchers and the public.

Therefore, in the case of ethnomusicology data sharing, the challenges can be divided into two categories:

1. managing access to data and metadata (technical challenge),
2. providing clarity on the data rights to enable sharing and reuse (legal and organizational challenge).

The first can be addressed by a data repository, the solution already proposed to requirement R1. Data repositories allow data-owners (researchers) to choose which data or metadata will be visible to the public and how controlled access rights can be applied to the data. Most systems provide different modes of access, such as the data-owner (full rights), collaborator (read only), guest visitor (read metadata only), etc. Solving the second challenge requires the institutions to establish Research Data Management (RDM) policies along with Digital Rights Management (DRM) policies which describe the responsibilities of researchers and institutions with respect to data management, collection, and security. For example, an RDM policy can obligate an institution to manage a data repository and to provide legal support on intellectual

property rights in a certain direction. A DRM policy can obligate the institution to set up access control to data resources stored in the repositories: who to access, when to access, what to access, how to access, etc. Researchers can in turn be required to use the provided infrastructure for managing their data. The RDM and DRM policies can obligate researchers to create Data Management Plans (DMP) and Data Security Plan (DSP) for each project.

DMPs describe which data are used and produced during the research, where data will be archived, which licenses and constraints apply, and to whom credit should be given. DMPs can be seen as awareness tools to identify issues that may be encountered during data collection, e.g., ownership of data. Writing a DMP cannot be a sole responsibility of an ethnomusicology researcher because they may lack necessary technical or legal knowledge.

DSPs describe which data are allowed to be accessed when, how, and by whom. An important part of DSPs is the authentication of users. Technologies related to digital IDs may be used to prevent unauthorized access. Since intellectual property plays a very important role in the usage and distribution of ethnomusicology data, preventing unauthorized access is extremely important.

In the workshops the question of intellectual property was identified as an issue of significant importance for ethnomusicology data, due to the complex interactions of different rights-holders (researchers, performers, composers, cultural communities, group authorships). Hence, this will require future effort and bundled expertise of ethnomusicologists, international lawyers, and maybe other entities such as national cultural heritage agencies of different countries to evaluate the applicability of existing licenses, or to develop new applicable license models for this domain.

R2 Action items:

- **Institutions:**
  - Provide data repositories allowing for different modes of data access
  - Develop RDM and DRM policies to clarify responsibilities
  - Require researchers to create DMPs
  - Operate services providing support in research data management
- **Researchers:**
  - Use data repository services provided by institution
  - Write DMPs and DSPs and clarify up-front issues in data management
  - If data sharing is not possible, consider sharing metadata
- **Ethnomusicology community:**
  - Evaluate existing licenses or develop new applicable license models for ethnomusicology research data
  - Collect best-practice licensing examples

Fig. 2. Action items for controlled data access and sharing with collaborators and contributors.

R3 Importing existing collections

Many data holdings of the individual institutions already contain a large number of data sets. Importing existing collections into a new system must be semi-automatic. Therefore, an API to access the storage for bulk data importing is required. The main challenge lies in the heterogeneity of existing collections that have been collected over years on different mediums, using different descriptions, different sets of data, different data types and formats, languages, annotations, etc. This is a typical digital data collection and preservation problem, where a data integration and preservation plan for each logical collection of items must be developed (Becker et al., 2009). Such a plan describes significant properties of existing collections and evaluates methods that best preserve them and enable them to be used collectively together. These problems are well-known in the cultural heritage domain. The tools and techniques developed in that domain may also be applied, with slight modification, to the ethnomusicology domain for moving the existing objects into data repositories (DPC, 2015).

R3 Action items:

- **Researchers:** Develop preservation plans for their collections
- **Institutions:** Provide technical support and funds to enable populating data repositories with existing data

Fig. 3. Action items for importing existing collections.

R4 Search across distributed data collections

The benefit of using interoperable data collections is the increase in value for each data entry by new data entries and the use of larger and more diverse sets of data. This is the network effect, also often referred to with social network technology. Compatible, structured and machine-readable data descriptions (metadata) will support unified search across distributed data collections and may overcome language and cultural barriers. Using the open Dublin Core standard as a minimum set for metadata fields for each data point, will bring basic compatibility to the ethnomusicology research data. From the collected use-cases, we extracted additional metadata fields that enable domain-specific search queries, such as the musical instruments used in a recording. Table A1 (Appendix A) gives a summary of both, the 15 Dublin Core metadata terms and our additions.

To overcome language-specific data descriptions, the usage of Linked Data Uniform Resource Identifiers (URIs) for both the metadata terms and their properties was discussed in the workshops. As a technical solution with a low entrance barrier, we propose to use Wikidata[10], an open-source, user-editable, document-oriented database that stores items representing topics, concepts, and objects as URIs. Each item has labels in different languages and references to other resources like Wikipedia articles, geographical coordinates, and other knowledge bases. The application of URIs can help to avoid confusion with entities that have similar names but different meanings e.g., 'Bandung' (wd:Q10389), the city in West Java, Indonesia – versus – 'Bandung' (wd:Q3435515) a district in Serang, Indonesia. Table A1 (Appendix A) shows the vocabulary for both the basic Dublin Core fields and the Wikidata additions.

We propose the usage of Wikidata as a temporary solution, where researchers around the world can edit or add missing items themselves, to encourage experiments with a Resource Description Framework (RDF) and language-independent data entries. Once a domain-specific vocabulary has been established over time, a standardization can be derived by reviewing the used vocabulary of a larger number of compatible repositories. Further support for established Linked Data standards can be achieved by mapping the observed vocabulary to existing ontologies e.g., MusicOntology, DOREMUS.

R4 Action items:

- **Researchers:**
  - Describe research data via the standardized metadata fields;
  - Experiment with Wikidata URIs as entities and add missing entities to Wikidata;
- **Institutions:**
  - Provide repositories with metadata templates that are compatible with Dublin Core
  - Support domain-specific extensions
- **Ethnomusicology community:**
  - Review the proposed metadata extensions after a testing period
  - Derive field specific (RDF) vocabulary and/or map to established ontologies

Fig. 4. Action items for search across distributed data collections.

R5 Automatic (audio) data analysis for meta-data generation

A large fraction of data gathered in ethnomusicology research contains audio, video, images, texts, and other materials where automation technology can be applied. With the current state of the art, there are automatic data integration and data analysis tools which are helpful to aggregate data of different formats and support researchers by generating additional metadata for non-text material using various signal processing techniques (e.g., sound pattern recognition, instrument detection, genre classification). Analysis of other types of media (e.g., images, texts, or video) may also benefit from the developments in research fields such as artificial intelligence (AI) and machine learning (ML), in particular image information retrieval, computer vision, signal processing and natural language processing. To develop, analyze and test algorithms, ground truth data in the form of the raw data plus human annotations are required. By providing their high-quality annotated data, the musicology community can contribute to computer science research and vice versa.

Algorithms are constantly improved by the computer science community; hence, an analysis tool chain should run independently and in parallel with the repository and be designed in a way that the algorithms can be updated or exchanged (plug-ins). The results of the analyses must be stored in a machine- and human-readable format. The repository needs to provide viewable and editable access for the machine-created information, so that the data-owner can approve and eventually correct it prior to publication.

Audio recordings in ethnomusicology are mostly done during fieldwork which may have implications on the quality of the recordings. Depending on the position of the microphones to the performers, large portions of reverberation and other stray noise may be in the sound. Hence, the acoustic conditions are not comparable to those of studio recordings. Consequently, MIR algorithms that are specifically suited for such audio material are preferably chosen.

R5 Action items:

- **Researchers:** Publish audio data together with annotations (ground truth data) in a machine readable a format, to support MIR research and algorithms
- **Institutions:** Provide technical support for integration of micro-services for automatic (audio) data analyses and provide GUI support for display and manual corrections
- **Ethnomusicology community:** Collaborate with MIR community to define common goals, identify challenges and compile (larger) example datasets

Fig. 5. Action items for automatic (audio) data analysis for meta-data generation.

5. SOFTWARE PROTOTYPE

In this section we describe a proof-of-concept implementation of compatible, distributed data repositories for ethnomusicology. We used the system to demonstrate to ethnomusicology researchers how the proposed solution (cf. Section 4) can be realized and be a basis for discussions. Due to the distributed nature of the system, realization of a production ready system in the future requires a joint community effort and commitment from various institutions and researchers. We provide screencasts, demonstrating the functionalities that are described below (Bakos, 2020). In the remainder of this section, we explain the architecture of the system and provide examples of how typical data management tasks are supported.

System architecture

The system is designed as a network of local nodes each running at a different host institution (see Figure 6). Each node comprises a secure local data repository and an automatic music analyzer tool and provides different interfaces. A web GUI allows researchers to enter, edit, manage, and search for research data within their own local data repositories. A web API that supports bulk data imports and exports based on an agreed data exchange standard. An important feature of this web API is that each institution can share their metadata entries publicly. This feature is relevant for fast, centralized search engines that recurrently harvest the metadata and provide cross-network search results.

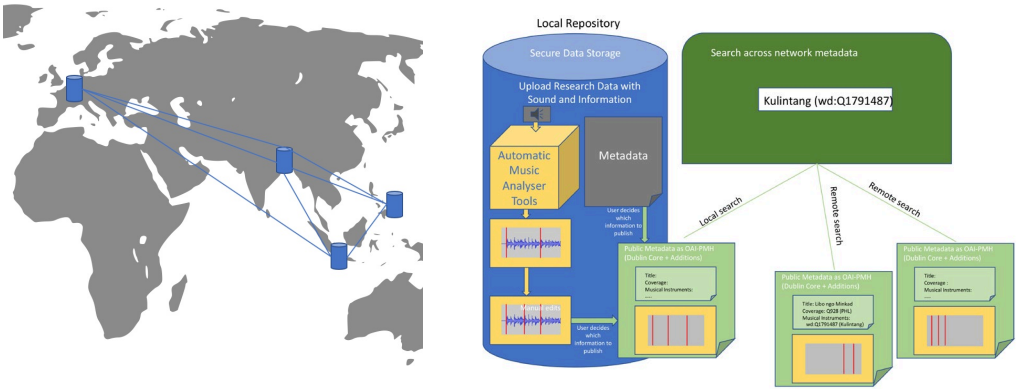

Fig. 6: Exemplary network of local nodes, connected via API endpoints using standardized metadata for information exchange (left). One local node, with storage system, music analyzer, and cross-network search engine (right).

The reason we propose a distributed network of repositories and not a single repository is the fact that researchers are skeptical when it comes to handing over their data to someone else according to the legal restrictions related to intellectual properties. For this reason, establishing individual data repositories in each region and providing a common metadata search interface displaying results from all participating repositories is a logical and feasible solution.

System functionality

To develop a proof of concept, we customized the open-source research data repository software *Dataverse*.

MAKING DATA FAIR

Each node allows users to create so-called dataverses that are containers for datasets. Each dataverse has a name, persistent identifier and information about its owner, such as name, affiliation and contact details. Furthermore, users can decide which metadata standards are applicable for the given dataverse.

The system also allows users to create metadata templates that can be used when adding new datasets to a dataverse (Figure 7). Templates can be pre-filled with information that is the same for all of the objects; for example, contact person, author, or any other metadata fields. Thus, the workload is reduced and there is an incentive for researchers to use the system. Once the dataverse is created, users can store their data to their local data repositories using a GUI in the web browser. Each dataset gets a persistent identifier (DOI), but access to the file is restricted. This means that dataset owners can continue to work on the dataset, (improve metadata, add further files, clarify terms of publishing, etc.) while ensuring the dataset is securely stored. When the dataset is ready for publishing the visibility of the dataset can easily be changed to "public". Each dataset that is published has a license, to support reuse.

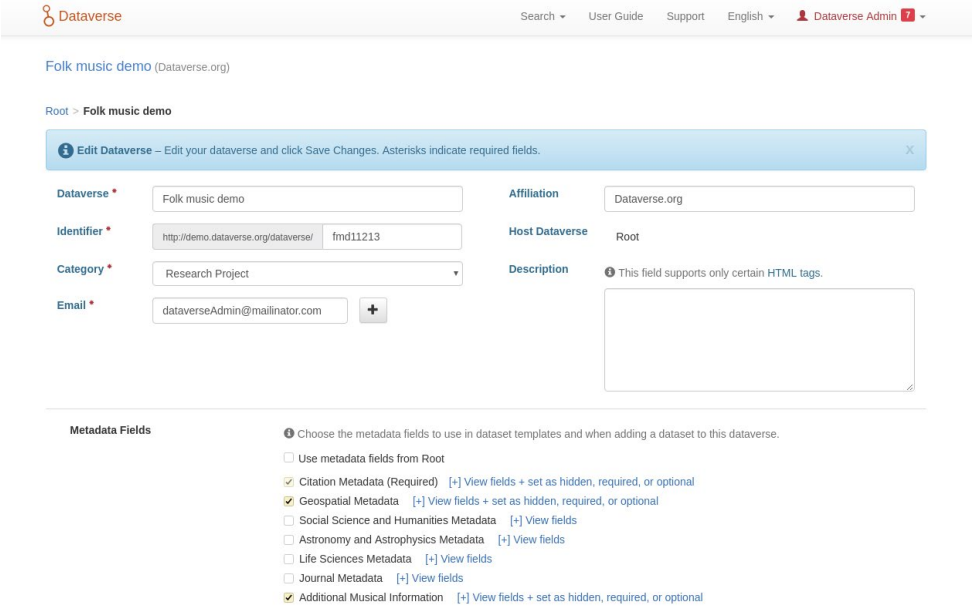

Fig. 7: Screenshot of web user interface for creating a dataverse to store data collected in fieldwork. A suitable metadata scheme can be selected.

The owner of the dataverse can select whether they are the only person having access to it, or whether other researchers can add to or view the dataverse. Figure 8 shows different roles and user groups that the owner can choose for their collaborators. Thus, the system supports collaboration within projects.

If the owner restricts files within a dataset, they must provide terms of access for the dataset. The files are listed as part of a dataset, but their contents cannot be seen. Other researchers can request access and the owner can decide whether the process is granted or not. Here we made the workflow for requesting access to restricted data explicit. This is a big improvement over existing practice, where researchers have to find contact details such as emails, phone numbers, etc.

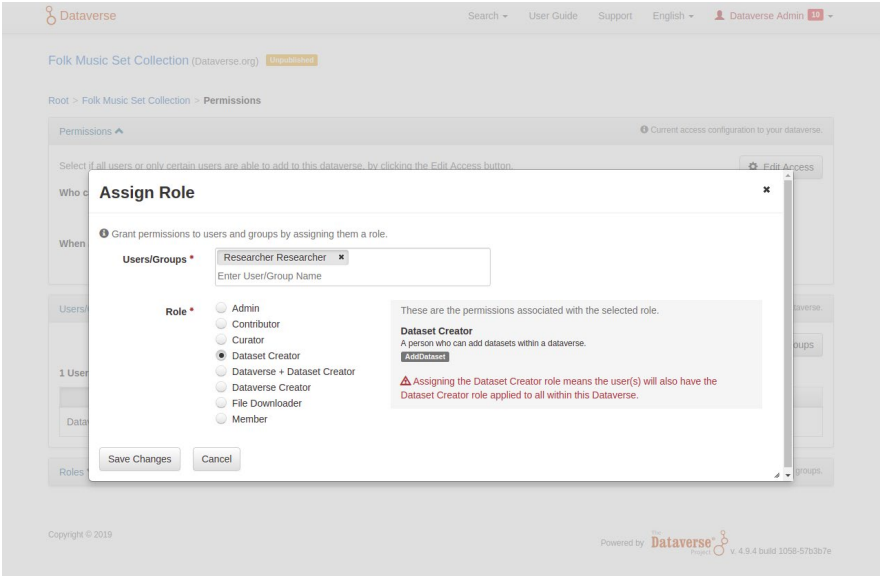

Fig. 8: Screenshot of web user interface with different roles and permissions that a dataset owner can provide for others.

Thus, data stored in the system becomes FAIR, because:

- The use of identifiers and metadata supports its findability and interoperability. The end users can easily locate the relevant data and quickly examine whether the data is of interest.
- Different modes of access support accessibility by providing clear options on whether and how the data can be accessed.
- The possibility to define custom terms for access and licensing mechanisms supports reuse, by making explicit what data is available, and under which conditions

Note: although good data management practices are a pre-requisite, the repository itself will not make data FAIR. The data points itself must be properly documented and organized from the moment that they were collected.

LINKED DATA TO FURTHER IMPROVE INTEROPERABILITY

We extended the basic Dataverse implementation with a mechanism that provides links to Wikidata (see Figure 9). When a user fills in metadata, the system searches the Wikidata URI, and adds the URI to the metadata page of the dataset. The URI unambiguously describes this object while also providing additional information such as language labels. Thus, despite the lack of common standards for metadata in ethnomusicology, we propose a mechanism that supports unification of terms and improves interoperability. However, this process requires researchers to carefully plan data integration together to build an agreed-upon data exchange and repository design standard.

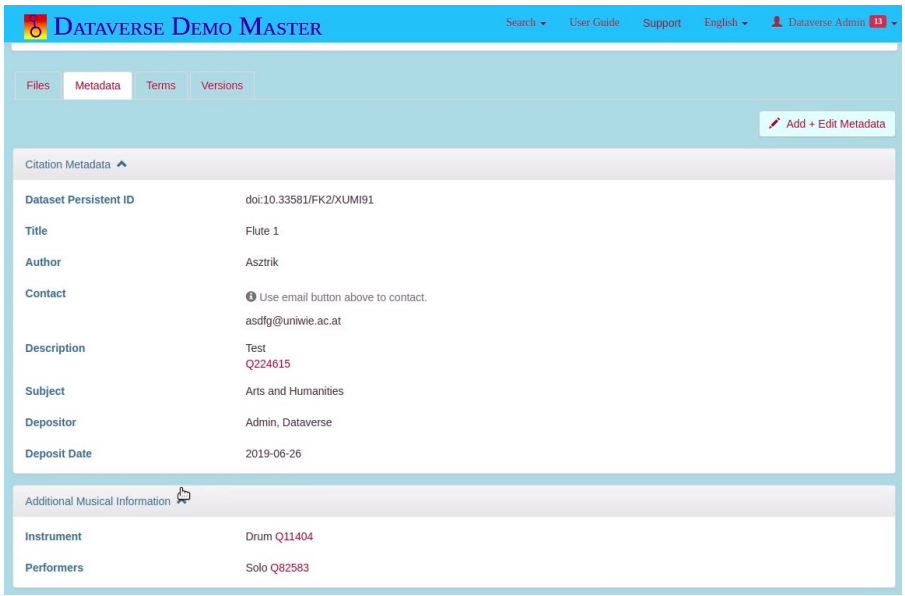

Fig. 9: Screenshot of web user interface on a public node, showing Wikidata entries (e.g., wd:Q11404) in metadata to improve interoperability and language independence.

CENTRALIZED SEARCH TO FURTHER IMPROVE FINDABILITY

We used the OAI-PMH protocol to exchange metadata between repositories. For demonstration we created an additional node that only aggregates information on datasets from other nodes. These nodes are Dataverse instances operated by us, as well as real existing repositories, such as AUSSDA[11]. We used a mix of repositories to demonstrate to ethnomusicology researchers that their data remains with their local repository, but metadata information is displayed also in other systems. When researchers select a dataset from a list of

results, they are always redirected to the respective local repository. The data never leaves the original repository, which is one of the fundamental requirements that makes the local nodes trustworthy in the view of the ethnomusicologists. We provide a screencast demonstrating the process of harvesting of metadata into the search node and querying its content (Bakos, 2020).

MUSIC INFORMATION RETRIEVAL TO AUTOMATE TASKS

We implemented a prototypical Music information retrieval service, based on Python and the Falcon Framework[12] and made the code publicly available (Hofmann and Knees, 2020). The service can be accessed via a REST-API, but for security reasons we recommend running the MIR Service locally in parallel with the repository system. When the service receives an audio file, the automatic analysis is triggered and the analysis results are returned as a structured JSON-file. As an example of how to integrate such a service, we configured the Dataverse repository so that for each uploaded audio file, the service is called and the generated annotation file is automatically stored along with the original data. As a proof of concept, a segmentation of speech and music was implemented, with the original settings provided by Marolt *et al.* (2019), that achieves an F1 measure up to 0.63. Considering this moderate segmentation accuracy, the need to support manual correction of the annotations by ethnomusicology experts becomes evident in order to verify the quality of the data, which at the same time creates new ground truth datasets that may inform future analysis algorithms.

6. DISCUSSION

With the aim to establish a culture for FAIR and open research data in ethnomusicology, we directly interacted with a group of stakeholders to collect data about the current situation, and to develop new use-cases. From these we derived requirements and discussed our findings with the wider community.

Pre-FAIR

Digitization had a strong impact on ethnomusicology and most data holdings are either in a digital form or are currently being digitized. However, we observed many individual storage solutions across the domain, lacking a required standardization that would allow instant data compatibility and exchange. Although many researchers support the idea of opening their collections and sharing their research data, the domain is missing uniform metadata, ontologies, and taxonomies. Dealing with multi-language content further complicates data interoperability. A second problem is the lack of clear research data policies at the host institutions. Developing such policies will support researchers in the future when clearing the data rights and support sharing and reuse of data.

FAIR

Within this project, we developed a list of requirements specifically suited for ethnomusicology research data management (Table 1). These support the FAIR principles and furthermore consider the specifics of the field, where working with music recordings may involve different rights-holders and an exchange may be restricted because of copyright-protected material. Nevertheless, even for restricted original data, we strongly encourage the publication of metadata to support the *findability* of the data.

As the field is in the very beginning of establishing a standardized vocabulary, this paper proposes to build upon a basic set of interoperable metadata fields that are based on the established Dublin Core standard. Derived from query examples, developed by the researchers in the workshops, we propose minor additions of field-specific metadata. Based on the methodology of related digital musicology projects, we proposed experiments with Linked Data that prepare for future experiments with semantic web technology (e.g., RDF). We explored the possibility of using Wikidata URIs, that can be created and edited by the researchers themselves, instead of defining a new ontology for this project at such an early stage of development.

Reusability of ethnomusicological datasets presented a challenge; due to copyright restrictions, not all the research data can be shared publicly. Hence, an important customization for research repositories in ethnomusicology is controlled data access management.

Beyond FAIR

There was strong interest in the discussions on the currently available tools in computer science and how these can be applied to ethnomusicology research. Hence, this paper gives an overview of the existing technologies and aims to point out how interdisciplinary collaborations of ethnomusicologists and computer scientists can lead to advances in both domains. With the aim to promote FAIR principles on one side and knowing the burden of complicated data rights on the other side, we emphasize how high-quality open data will pave new ways of research in this field.

Segmentation tasks can already be accomplished with the tools available today, and in the near future automatic instrument detection will be possible. However, here it will be the interplay of researchers from multiple domains that advances the technology. Machine learning-based algorithms can only be improved with high-quality ground truth data. Thus, standardized annotations and correct descriptions of larger ethnomusicological corpora will become valuable sources for the improvement of annotation algorithms. These algorithms will then support future research in ethnomusicology.

7. CONCLUSION

With the initial motivation to combine MIR research with ethnomusicological data, we reached out to different institutions and collected information on their data holdings. The affiliated ethnomusicologists showed strong interest in the application and further development of computational tools for their field, but the sharing of their research data turned out to be problematic. To undertake such joint research projects in the future, significant effort is needed to make ethnomusicological research data FAIR.

In this paper we outlined the required steps and *action items* to achieve FAIRness. We furthermore discussed how Linked Data technology and MIR may enrich the data holdings and support the researchers in the future. We hope this outlook motivates further undertakings in the direction of FAIRness in this domain and levels the ground for future MIR developments. As fieldwork data is created through complex interactions of different actors, it will require the ethnomusicology community to isolate the problems and resolve the challenges as a community step by step. In some cases, collaborations with other communities, such as information scientists or international lawyers, may be required.

We presented technical solutions and experiments with prototypical implementations to receive feedback from the ethnomusicology community. These preliminary steps, aimed to show opportunities for transdisciplinary research and to initiate discussions on FAIR and open data. Overall, we observed that the ethnomusicology domain is at the very beginning of such an initiative and several arguments urge for resolving the discovered problems. This could be either because of regulations coming from funding bodies, or it could be the aim of the researchers themselves to share their data. In the second case, where the initiative comes from an increasing number of researchers, this presents the opportunity to actively develop new ideas and thereby shape the future directions of data management in this domain.

ACKNOWLEDGEMENTS

Research reported in this publication was jointly supported by the ASEAN-European Academic University Network (ASEA-UNINET), the Austrian Federal Ministry of Education, Science and Research and the Austrian Agency for International Cooperation in Education and Research (OeAD-GmbH). The authors are thankful to Dr. Joseph Bowman and Anita Taschler for their support when organizing the workshops, and to all participants in the workshops, and at the panel discussions, that were part of this project. This article was copyedited by Matthew Moore and layout edited by Diana Kayser.

NOTES

[1] Correspondence can be addressed to: Alex Hofmann, Department of Music Acoustics, University of Music and Performing Arts Vienna, Austria, email: hofmann-alex@mdw.ac.at

[2] https://dataverse.org

[3] https://invenio-software.org/products/rdm/

[4] https://www.dspace.com/

[5] https://www.openarchives.org/pmh/

[6] https://dublincore.org

[7] https://schema.datacite.org

[8] 11 data fields: data identification number, old catalogue ID, original format, e-format, location of original, e-location, researcher, year, group/country, image captured, description/notes

[9] Projektname, Titel, Hauptsachtitel, Aufnahmeort, Schlagwörter, Produzent_in, Aufnahmedatum, Gewährsperson, Musikgruppen/Körperschaften, Ausführende/Mitwirkende/Vortragende/Regie, Musikgattungen, Traditionsort, Musikinstrumente, Aufnahmegerät, Aufnahmesituation, Dauer, Protokollbearbeiter_in, etc..

[10] https://www.wikidata.org

[11] https://aussda.at

[12] https://falconframework.org/

REFERENCES

Abdallah, S., Benetos, E., Gold, N., Hargreaves, S., Weyde, T. & Wolff, D. (2017). The digital music lab: A big data infrastructure for digital musicology. *Journal on Computing and Cultural Heritage*, *10*(1), 2. https://doi.org/10.1145/2983918

Bakos, A. (2020). Enabling FAIR use of ethnomusicology data. Zenodo. https://doi.org/10.5281/zenodo.3751570

Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A. & Hofman, H. (2009). Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. *Int. J. Digit. Libr.*, *10*(4), 133-157. https://doi.org/10.1007/s00799-009-0057-1

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34–43. issn:00368733. https://doi.org/10.1038/scientificamerican0501-34

Bozkurt, B., Ayangil, R. & Holzapfel, A. (2014). Computational analysis of Turkish Makam music: Review of state-of-the-art and challenges. *Journal of New Music Research*, *43*(1), 3–23. https://doi.org/10.1080/09298215.2013.865760

Crawford, T., Fields, B., Lewis, D. & Page, K. (2014). Explorations in Linked Data practice for early music corpora. *IEEE/ACM Joint Conference on Digital Libraries*, 309–312. https://doi.org/10.1109/JCDL.2014.6970184

Digital Preservation Coalition (2015). *Digital Preservation Handbook*, 2nd edition. Digital Preservation Coalition. https://dpconline.org/handbook.

Fewkes, J. W. (1890). On the use of the phonograph in the study of the languages of American Indians. *Science, ns-15*(378), 267–269. https://doi.org/10.1126/science.ns-15.378.267.b

Hofmann, A. & Knees, P. (2020). Ethnomusicology music information retrieval analyser v0.2. Zenodo. (code: https://github.com/ketchupok/ethmusmir). https://doi.org/10.5281/zenodo.3751289

Holzapfel, A., Benetos, E. & others (2019). Automatic music transcription and ethnomusicology: A user study. In *Proceedings of the International Society for Music Information Retrieval* (pp. 678–684). https://qmro.qmul.ac.uk/xmlui/handle/123456789/59182.

Lerch, A. (2012). *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics.* Wiley-IEEE Press. isbn:9781118266823. https://doi.org/10.1002/9781118393550

Lisena, P., Achichi, M., Choffé, P., Cecconi, C., Todorov, K., Jacquemin, B. & Troncy, R. (2018). Improving (re-) usability of musical datasets: An overview of the DOREMUS project. *Bibliothek Forschung und Praxis*, *42*(2), 194–205. https://doi.org/10.1515/bfp-2018-0023

Marolt, M., Bohak, C., Kavcic, A. & Pesek, M. (2019). Automatic Segmentation of Ethnomusicological Field Recordings. *Applied Sciences*, *9*(3), 439. https://doi.org/10.3390/app9030439

Meredith, D. (2016). *Computational Music Analysis.* Springer. isbn:978-3-319-25929. https://doi.org/10.1007/978-3-319-25931-4

Merriam, A. P. (1960). Ethnomusicology discussion and definition of the field. *Ethnomusicology*, *4*(3), 107–114. https://doi.org/10.2307/924498

Miksa, T., Simms, S., Mietchen, D. & Jones, S. (2019). Ten principles for machine-actionable data management plans. *PLOS Computational Biology*, *15*(3), 1–15. https://doi.org/10.1371/journal.pcbi.1006750

Müller, M. (2015). *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications.* Springer.

Myers, H. (1992). *Ethnomusicology: An Introduction.* WW Norton. isbn:9780393033779.

Nettl, B. (2005). The harmless drudge: Defining ethnomusicology. *The Study of Ethnomusicology: Thirty-one Issues and Concepts* (pp. 3–15), University of Illinois Press.

Page, K. R., Bechhofer, S., Fazekas, G., Weigl, D. M. & Wilmering, T. (2017). Realising a Layered Digital Library: Exploration and Analysis of the Live Music Archive through Linked Data. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 1–10). https://doi.org/10.1109/JCDL.2017.7991563

Panteli, M., Benetos, E. & Dixon, S. (2018). A review of manual and computational approaches for the study of world music corpora. *Journal of New Music Research*, *47*(2), 176–189. https://doi.org/10.1080/09298215.2017.1418896

Pascua, S. M. (2018). Linked Data: Metadata schemata of the music scores of Jose Maceda Collection. In *IFLA WLIC Kuala Lumpur*. http://library.ifla.org/id/eprint/2203 retrieved: 12.4.20.

Pugin, L. (2015). The challenge of data in digital musicology. *Frontiers in Digital Humanities*, *2*, 1–4. https://doi.org/10.3389/fdigh.2015.00004

Raimond, Y., Abdallah, S. A., Sandler, M. B. & Giasson, F. (2007). The Music Ontology. In *ISMIR, Vol. 2007* (pp. 1–6).

Saglam, H., Ahmedaja, A., Hofmann, A., Knees, P. & Miksa, T. (2019). Towards an alliance for distributed music data (Abstract). In *Abstracts of The 45th International Council for Traditional Music World Conference* (pp. 130–133).

Shadbolt, N., Berners-Lee, T. & Hall, W. (2006). The semantic web revisited. *IEEE intelligent systems*, *21*(3), 96–101. https://doi.org/10.1109/MIS.2006.62

Weigl, D. & Page, K. (2017). A framework for distributed semantic annotation of musical score: Take it to the bridge! In *Proceedings of the International Society for Music Information Retrieval*. uri:945287f6-5dd3-4424-940c-b919b8ad2768.

Wilkinson, M., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L. O., Bourne, P., Bouwman, J., J. Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R. & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.18

APPENDIX A

Table A1: Metadata for ethnomusicology research data, comprising 15 Dublin Core terms and domain-specific additions.

Metadata term	RDF vocabulary	Field description
Title	`dc:title`	The name given to the resource/recording.
Creator	`dc:creator`	An entity primarily responsible for making the resource. E.g., a person (researcher), an organization, or a service.
Subject	`dc:subject`	The topic of the resource. Represented using keywords, key phrases, or classification codes as URIs.
Description	`dc:description`	A brief summary about the content, e.g., abstract, table of contents, graphical representation, or free text.
Publisher	`dc:publisher`	An entity responsible for making the resource available. E.g., a person, an organization, or a service.
Contributor	`dc:contributor`	An entity responsible for making contributions to the resource.
Date	`dc:date`	A point or period of time associated with an event in the lifecycle of the resource in ISO 8601 format: YYYY-MM-DD
Type	`dc:type`	The nature of the resource. Based on Wikidata URIs (Audio=wdt:P51) (Video=wdt:P10) (Image=wdt:P18)
Format	`dc:format`	The file format, physical medium, or dimensions of the resource.
Identifier	`dc:identifier`	Reference to the resource within a given context.
Source	`dc:source`	A related resource from which the resource is derived.
Language	`dc:language`	Language of the resource.
Relation	`dc:relation`	A related resource given as URI.
Coverage	`dc:coverage`	The spatial or temporal topic of the resource like a place or location given as a WikiData entry. (e.g., recording venue)

Table A1 (continued)

Metadata term	RDF vocabulary	Field description
Rights	dc:rights	Information about rights held in and over the resource, e.g various property rights associated with the resource, including intellectual property rights.
Measuring instrument/technical Equipment	wd:Q2041172	Recording equipment, used to gather the data (manufacturer, model)
Number of audio channels	wd:Q71821715	Amount of recorded channels in this recording
Duration of recording	wd:Q2199864	Entire length of the recording.
Musical Ensemble (type)	wd:Q2088357	Type/format of the musical groups/person(s), given as WikiData URIs (e.g., wd:Q43304311).
Musical Instruments	wd:Q34379	Musical instruments used in the recording, based on WikiData URIs (e.g., wd:Q6443033)
Area of Music Origin	wdt:P495	Geographical area where this music tradition is practiced, given as Wikidata URI.
Origin of Performers/Formation	wdt:P740	Geographical place (Countries/Cities) of the performers, as WikiData URIs.
Community	wd:Q177634	Cultural belonging of the performers, as WikiData URIs (e.g., religion, ethnical groups)
Storage Location / Collection	wdt:P276	Preservation location of the original recording
Keywords	wd:Q17152639	Categorization of music traditions on the basis the recording
Music Genre	wd:Q188451	Musical conventions or shared tradition the music belongs to.