A Large-Scale Semantic Library of MIDI Linked Data

Albert Merño-Peñuela
Dept. of Computer Science,
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands
albert.merono@vu.nl

Anna Kent-Muller
Dept. of Music,
University of Southampton
Southampton, United Kingdom
alkm1g12@soton.ac.uk

Reinier de Valk
Jukedeck Ltd.
London, United Kingdom
reinier@jukedeck.com

Marilena Daquino
Dept. of Classical Philology and
Italian Studies, University of Bologna
Bologna, Italy
marilena.daquino2@unibo.it

Enrico Daga
Knowledge Media Institute,
The Open University
Milton Keynes, United Kingdom
enrico.daga@open.ac.uk

ABSTRACT

Over recent decades, the natural sciences have moved from formulating hypotheses through the observation of phenomena to generating them automatically through the analysis of large cross-disciplinary datasets, collected and maintained within large collaborative projects. Recently, it was suggested that musicology should embrace the same paradigm shift, and move to a more collaborative and data-oriented culture. In this paper, we describe the MIDI Linked Data Cloud, an RDF graph of 10 billion MIDI statements linked to contextual metadata. We show examples of its potential application for digital libraries for musicology, and we argue that the use of Linked Data for integrating symbolic music notations and contextual metadata constitutes technical foundations for Web-scale musicology projects.

KEYWORDS

MIDI, Linked Data, Semantic Digital Library

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

DLfM 2018, September 2018, Paris, France
© 2018 Copyright held by the owner/author(s).
ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

1 MOTIVATION AND BACKGROUND

Over recent decades, the natural sciences have moved from formulating hypotheses through the observation of phenomena to their generation through automated analysis of large and cross-disciplinary datasets, collected and maintained within large collaborative projects. Various humanities disciplines, too, have recently begun to apply this interpretative and hermeneutic approach of scientific enquiry to the evidence collected in specialised databases [2, 12]. Increasingly adhering to this data-driven paradigm is the field of musicology, one of the fundamental aims of which is to study the symbolically formalised objects in musical scores and their contextual information, both at small and large scale [3], and which intersected early with computer science, giving birth to computational musicology [11, 20].

However, musicological research is usually performed on well-curated datasets of limited size, hosted by isolated digital libraries. Although a significant number of valuable resources are already available on the Web, they are exposed in heterogeneous ways, often not in a machine-readable form [7]. Recently, a number of projects have suggested from various perspectives that musicology should embrace the same paradigm shift that occurred in other scientific fields, like human biology with the Human Genome Project [5, 8], and move to a more collaborative and data-oriented culture stressing data integration [1, 14, 24, 27]. Specifically, Big Musicology [13] envisions that large musicology data integration platforms could generate hypotheses, uncover patterns and relations—of genre, style, and compositional influence—, facilitate knowledge discovery, and support understanding between close reading and distant reading questions. In order to address the integration of distributed musicological knowledge on the Web, a number of projects use Semantic Web and Linked Data methods and technologies [10]. DBpedia, for example, contains general metadata about popular bands, albums, and songs; 1 MusicBrainz [28] offers fine-grained descriptions of albums, songwriters, versions, and recordings; and AcousticBrainz describes acoustic characteristics of music and includes low-level spectral information. 2 Other examples of the use of Semantic Web technologies to further musicological research are found in [4, 6, 22, 23, 26, 29]. Despite their pioneering contributions, in these efforts Linked Data is primarily used to represent music metadata and workflows, but not notation.

The ability to access both large-scale notation content and metadata at the same time, via Linked Data, the Resource Description Framework (RDF), ontologies, and the RDF query language SPARQL, could greatly contribute to the construction of Big Musicology [13], by enabling the explanation of trends found at distance through pointing—using URIs, unique and global identifiers—at close-reading observations such as specific notes, bars, and instruments.

1 https://www.dbpedia.org/
2 https://acousticbrainz.org/
THE MIDI LINKED DATA CLOUD

Countless MIDI files, mostly but not exclusively encoding popular music, are available on the Web. The format’s simplicity and openness allow for the quick development of manipulation programs and conversions into other formats, making it favoured among musicians and researchers alike. The MIDI Linked Data Cloud (MIDI-LD) [19] is an RDF graph containing 10,215,557,355 triples linking the contents of 308,443 MIDI files gathered from the Web. This dataset follows standards and best practices within the Linked Data community [10], and contributes integrated, interoperable, and interconnected music notation for projects like Big Musicology. The MIDI-LD dataset is built with an algorithm that maps MIDI events onto RDF triples (subject-predicate-object statements), using the lightweight MIDI ontology summarised in Figure 1. In this ontology, a midi:Piece contains all MIDI data from a file organised in midi:Tracks, each containing a number of midi:Events. A midi:Event abstractly represents any MIDI event; specific event types, like note onset, note offset, or instrument change, are represented by its subclasses (midi:NoteOnEvent, midi:NoteOffEvent, midi:ProgramChangeEvent). These can have their own distinctive attributes (e.g., a midi:NoteOnEvent has a pitch and a velocity), but all event types have a midi:tickOffset setting them temporally within the track. Instances of midi:Piece are linked to the original files they were derived from (instances of midi:MIDIFile) through prov:wasDerivedFrom. Instances of midi:MIDIFile are linked to the class mo:Track of the Music Ontology by giving the latter the predicate mo:available_as [25]. Listing 1 shows an example of a MIDI file represented in RDF.

The RDF graph is made accessible through the following services:

- A search engine and Linked Data browser to find MIDI content and metadata;4
- A SPARQL endpoint, allowing users to write their own MIDI SPARQL queries;
- A RESTful API, providing a usable interface for users and applications to get relevant MIDI data using HTTP requests;
- Various documentation pages and user manuals;
- A programming library and a Web service, midi2rdf [18], for users to create their own MIDI RDF graphs;

### Listing 1: Excerpt of a MIDI file represented in RDF.

```sql
PREFIX midi: <http://purl.org/midi-ld/midi#>
PREFIX dbr: <http://dbpedia.org/resource/>
SELECT ? pattern WHERE {
  ?event midi:denominator 4
  ?event midi:numerator 4
  ?event midi:note midi:pitch 7
  ?event midi:velocity 115
  ?event midi:metricWeight 1.0
}
```

### Listing 2: SPARQL query for MIDI files that reference Romeo and Juliet in common time.

- The SPARQL-DJ [16], a mixing engine for creating MIDI mashups from existing MIDI Linked Data tracks;
- An interface [17] to add data and links to MIDI-LD via user annotations, MIDI similarity and named entity recognition.

This last service establishes a bridge between notation and metadata, connecting MIDI events, similar songs, user-provided metadata, and automatically recognised named entities. For example, the SPARQL query in Listing 2 looks up all MIDI files that reference the topic Romeo and Juliet in common time (i.e., 4\2 metre). While the former comes from named entities contained in non-MIDI metadata, the latter comes from MIDI event information. The query returns two results: a movie soundtrack and a Dire Straits song. We propose this querying approach, which combines MIDI and metadata, and uses URIs to identify resources in both, as a method well-aligned with library organisation and retrieval principles.

### 3 FUTURE CHALLENGES

In this paper, we describe the use of Linked Data to represent MIDI in a machine-readable and Web-interoperable way, linking it to contextual metadata, and showing querying possibilities for digital music libraries. Although seemingly counterintuitive, our aspiration is to also embrace lesser quality data, and develop a community of contributors, processes, and tools for its enhancement [9]. Such processes will include the interlinking of the content with other datasets on the Web through the development of pipelines for link discovery and content curation. As a result, the quality of data can be improved, in a distributed way, to include well-curated and purpose-oriented collections. A next step for MIDI-LD towards Big Musicology is to aggregate the data of the musoW catalogue [7] and interlink it with other hubs of music Linked Data [15, 21]. This will steer Big Musicology towards large-scale format interoperability, and bring together MusicXML, MEI, “kern, MIDI, and other digital formats in an integrated dataspace of interlinked music notation.
REFERENCES


