A Large-Scale Semantic Library of MIDI Linked Data

Albert Meroño-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

ACM Reference Format:
Albert Meroño-Peñuela, Anna Kent-Muller, Reinier de Valk, Marilena Daquino, and Enrico Daga. 2018. A Large-Scale Semantic Library of MIDI Linked Data. In *Proceedings of 5th International Conference on Digital Libraries for Musicology (DLfM 2018).* ACM, New York, NY, USA, Article 4, 3 pages. 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.

DLfM 2018, September 2018, Paris, France

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

[1] https://www.dbpedia.org/
[2] https://acousticbrainz.org/

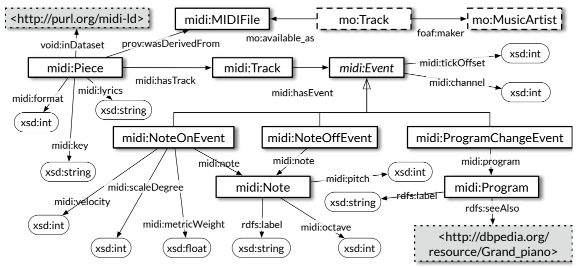

Figure 1: Excerpt of the MIDI ontology.

2 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.[3] 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` locating 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. IRIs of `midi:Piece` instances have the form `midi-r:piece/<hash>/`, where `<hash>` is the unique `MD5` hash of the original MIDI. Additional links to, for example, key, chords, and instruments further enrich this representation.

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;

```
midi-p:cb87a5bb1a44fa72e10d519605a117c4 a midi:Piece ;
    midi:format 1 ;    midi:key "E minor" ;
    midi:hasTrack midi-p:cb87a5b/track00,
        midi-p:cb87a5b/track01, ... .
midi-p:cb87a5b/track01 a midi:Track ;
    midi:hasEvent midi-p:cb87a5b/track01/event0000,
        midi-p:cb87a5b/track01/event0001, ... .
midi-p:cb87a5b/track01/event0006 a midi:NoteOnEvent ;
    midi:channel 9 ;      midi:note midi-note:36 ;
    midi:scaleDegree 6 ;  midi:tick 0 ;
    midi:velocity 115 ;   midi:metricWeight 1.0 .
```

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

```
PREFIX midi: <http://purl.org/midi-ld/midi#>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX dbr: <http://dbpedia.org/resource/>
SELECT ?pattern WHERE {
  ?pattern a midi:Pattern .
  ?pattern dc:subject dbr:Romeo_and_Juliet .
  ?pattern midi:hasTrack ?track .
  ?track midi:hasEvent ?event .
  ?event midi:numerator 4 .
  ?event midi:denominator 4 .
}
```

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., $\frac{4}{4}$ 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.[5] 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.

[3] https://midi-ld.github.io/
[4] https://www.github.com/Data2Semantics/brwsr/
[5] See http://www.purl.org/midi-ld/pattern/13fa17dc74232f7cb710a4d8d9f796b2 and http://www.purl.org/midi-ld/pattern/7a08a4b1efd5ff7afd6c1066b4a8dd94.

REFERENCES

[1] Samer Abdallah, Emmanouil Benetos, Nicolas Gold, Steven Hargreaves, Tillman Weyde, and Daniel Wolff. 2016. Digital Music Lab: A framework for analysing big music data. In *Proceedings of the 24th European Signal Processing Conference, Budapest, Hungary*. 1118–1122.

[2] Alessandro Adamou, Simon Brown, Helen Barlow, Carlo Allocca, and Mathieu d'Aquin. 2018. Crowdsourcing Linked Data on listening experiences through reuse and enhancement of library data. *International Journal on Digital Libraries* (2018), 1–19.

[3] Guido Adler. 1885. Umfang, Methode und Ziel der Musikwissenschaft. *Vierteljahrsschrift für Musikwissenschaft* 1 (1885), 5–20.

[4] Sean Bechhofer, Kevin Page, David M. Weigl, György Fazekas, and Thomas Wilmering. 2017. Linked Data publication of live music archives and analyses. In *The Semantic Web—ISWC 2017: 16th International Semantic Web Conference*, Claudia d'Amato et al. (Eds.). Vol. 2. Springer, Cham, 29–37.

[5] Francis S. Collins, Michael Morgan, and Aristides Patrinos. 2003. The Human Genome Project: Lessons from large-scale biology. *Science* 300, 5617 (2003), 286–290.

[6] Tim Crawford, Ben Fields, David Lewis, and Kevin Page. 2014. Explorations in Linked Data practice for early music corpora. In *Proceedings of the 14th ACM/IEEE Joint Conference on Digital Libraries, London, UK*. 309–312.

[7] Marilena Daquino, Enrico Daga, Mathieu d'Aquin, Aldo Gangemi, Simon Holland, Robin Laney, Albert Meroño-Peñuela, and Paul Mulholland. 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*. 57–68.

[8] Charles DeLisi. 1988. The Human Genome Project: The ambitious proposal to map and decipher the complete sequence of human DNA. *American Scientist* 76, 5 (1988), 488–493.

[9] André Freitas and Edward Curry. 2016. Big data curation. In *New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe*, José María Cavanillas et al. (Eds.). Springer, Cham, 87–118.

[10] Tom Heath and Christian Bizer. 2011. *Linked Data: Evolving the Web into a global data space*. Morgan & Claypool.

[11] Walter B. Hewlett and Eleanor Selfridge-Field. 1991. Computing in musicology, 1966–91. *Computers and the Humanities* 25, 6 (1991), 381–392.

[12] Leif Isaksen, Rainer Simon, Elton T. E. Barker, and Pau de Soto Cañamares. 2014. Pelagios and the emerging graph of ancient world data. In *Proceedings of the 6th ACM Conference on Web Science, Bloomington, IN, USA*. 197–201.

[13] Anna Kent-Muller. 2017. Big Musicology: A framework for transformation. In *Proceedings of the 4th International Digital Libraries for Musicology Workshop*.

[14] Richard J. Lewis, Tim Crawford, and David Lewis. 2015. Exploring information retrieval, semantic technologies and workflows for music scholarship: The Transforming Musicology project. *Early Music* 43, 4 (2015), 635–647.

[15] Pasquale Lisena and Raphaël Troncy. 2017. Combining music specific embeddings for computing artist similarity. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, Suzhou, China,* Late-Breaking/Demo session.

[16] Rick Meerwaldt, Albert Meroño-Peñuela, and Stefan Schlobach. 2017. Mixing music as Linked Data: SPARQL-based MIDI mashups. In *Proceedings of the 2nd Workshop on Humanities in the Semantic Web, Vienna, Austria*. 87–98.

[17] Albert Meroño-Peñuela, Reinier de Valk, Enrico Daga, Marilena Daquino, and Anna Kent-Muller. 2018. The Semantic Web MIDI Tape: An interface for interlinking MIDI and context metadata. In *Proceedings of the Workshop on Semantic Applications for Audio and Music, Monterey, CA, USA*. In press.

[18] Albert Meroño-Peñuela and Rinke Hoekstra. 2016. The Song Remains the Same: Lossless conversion and streaming of MIDI to RDF and back. In *The Semantic Web: ESWC 2016 satellite events*, Harald Sack et al. (Eds.). Springer, Cham, 194–199.

[19] Albert Meroño-Peñuela, Rinke Hoekstra, Aldo Gangemi, Peter Bloem, Reinier de Valk, Bas Stringer, Berit Janssen, Victor de Boer, Alo Allik, Stefan Schlobach, and Kevin Page. 2017. The MIDI Linked Data Cloud. In *The Semantic Web—ISWC 2017: 16th International Semantic Web Conference*, Claudia d'Amato et al. (Eds.). Vol. 2. Springer, Cham, 156–164.

[20] John Morehen and Ian Bent. 1979. Computer applications in musicology. *The Musical Times* 120, 1637 (1979), 563–566.

[21] Terhi Nurmikko-Fuller, Daniel Bangert, Alan Dix, David Weigl, and Kevin Page. 2018. Building prototypes aggregating musicological datasets on the Semantic Web. *Bibliothek Forschung und Praxis* 42, 2 (2018), 206–221.

[22] Terhi Nurmikko-Fuller and Kevin R. Page. 2016. A linked research network that is *Transforming Musicology*. In *Proceedings of the 1st Workshop on Humanities in the Semantic Web, Heraklion, Greece*. 73–78.

[23] Kevin R. Page, Sean Bechhofer, György Fazekas, David M. Weigl, and Thomas Wilmering. 2017. Realising a layered digital library: Exploration and analysis of the Live Music Archive through Linked Data. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries, Toronto, ON, Canada*. 89–98.

[24] Kevin R. Page, Ben Fields, David De Roure, Tim Crawford, and J. Stephen Downie. 2013. Capturing the workflows of music information retrieval for repeatability and reuse. *Journal of Intelligent Information Systems* 41, 3 (2013), 435–459.

[25] Yves Raimond, Samer Abdallah, Mark Sandler, and Frederick Giasson. 2007. The Music Ontology. In *Proceedings of the 8th International Conference on Music Information Retrieval, Vienna, Austria*. 417–422.

[26] David De Roure, Kevin R. Page, Benjamin Fields, Tim Crawford, J. Stephen Downie, and Ichiro Fujinaga. 2011. An e-Research approach to Web-scale music analyis. *Philosophical Transactions of the Royal Society A* 369, 1949 (2011), 3300–3317.

[27] Jordan B. L. Smith, J. Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. 2011. Design and creation of a large-scale database of structural annotations. In *Proceedings of the 12th International Society for Music Information Retrieval Conference, Miami, FL, USA*. 555–560.

[28] Aaron Swartz. 2002. MusicBrainz: A Semantic Web Service. *IEEE Intelligent Systems* 17, 1 (2002), 76–77.

[29] David M. Weigl and Kevin R. Page. 2017. A framework for distributed semantic annotation of musical score: "Take it to the bridge!". In *Proceedings of the 18th International Society for Music Information Retrieval Conference, Suzhou, China*. 221–228.