The MIDI Linked Data Cloud

Albert Meroño-Peñuela[1(✉)], Rinke Hoekstra[1,6], Aldo Gangemi[2,8,9], Peter Bloem[1], Reinier de Valk[5], Bas Stringer[1], Berit Janssen[3], Victor de Boer[1], Alo Allik[4], Stefan Schlobach[1], and Kevin Page[7]

[1] Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
albert.merono@vu.nl
[2] ISCT-CNR, Consiglio Nazionale Delle Ricerche, Rome, Italy
[3] Meertens Instituut, KNAW, Amsterdam, Netherlands
[4] Queen Mary University of London, London, UK
[5] Data Archiving and Networked Services, KNAW, Hague, Netherlands
[6] Faculty of Law, University of Amsterdam, Amsterdam, Netherlands
[7] Oxford e-Research Centre, Oxford, UK
[8] University of Bologna, Bologna, Italy
[9] Paris Nord University, Villetaneuse, France

Abstract. The study of music is highly interdisciplinary, and thus requires the combination of datasets from multiple musical domains, such as catalog metadata (authors, song titles, dates), industrial records (labels, producers, sales), and music notation (scores). While today an abundance of music metadata exists on the Linked Open Data cloud, linked datasets containing interoperable symbolic descriptions of music itself, i.e. music notation with note and instrument level information, are scarce. In this paper, we describe the MIDI Linked Data Cloud dataset, which represents multiple collections of digital music in the MIDI standard format as Linked Data using the novel `midi2rdf` algorithm. At the time of writing, our proposed dataset comprises 10,215,557,355 triples of 308,443 interconnected MIDI files, and provides Web-compatible descriptions of their MIDI events. We provide a comprehensive description of the dataset, and reflect on its applications for research in the Semantic Web and Music Information Retrieval communities.

Keywords: MIDI · Linked data · Music interoperability

1 Introduction

Musicology is the scholarly study of *music* [2]. Most of its subdisciplines are highly interdisciplinary, and require combinations of datasets from different domains in order to succeed. For example, a musicologist studying the popularity of a dance style may need to combine audio, video, notation, and market sales datasets. This depicts an ideal scenario for the deployment of Linked Data. However, only certain types of musical *metadata* are available as Linked Data. DBpedia contains general metadata about popular bands, albums and songs,

C. d'Amato et al. (Eds.): ISWC 2017, Part II, LNCS 10588, pp. 156–164, 2017.
DOI: 10.1007/978-3-319-68204-4_16

but not about their musical characteristics. MusicBrainz [10] offers more fine-grained descriptions for albums, songwriters, versions and recordings, linking those to AcoustID,[1] which assigns unique fingerprints to audio files based on their content. AccousticBrainz[2] describes "acoustic characteristics of music and includes low-level spectral information".

Despite these successful initiatives, our community has given little attention to musical *notation*, traditionally the source of musical knowledge. Current formats for musical notation are diverse, and not directly interoperable. MusicXML [1], and the Notation Interchange File Format[3] (NIFF) represent Western musical notation and are used in various scorewriting applications. The Music Encoding Initiative[4] (MEI) formalizes notation using a core set of rules. The Musical Instrument Digital Interface (MIDI) [11] standard allows electronic musical devices to communicate by exchanging messages that can be interpreted as music notation. However, the mutual compatibility of these formats is burdened by different adoptions across applications and different features [7].

Linked Data can potentially benefit music notation in at least two important aspects: *notation interoperability*, since current notation formats would use RDF to encode musical information; and *entity interlinking*, since musically related entities (groups of notes linked to a motif; instruments linked to DBpedia) could easily be connected. In previous work, we have shown that the MIDI format can be losslessly represented as Linked Data , using `midi2rdf` [5]. In this paper, we describe the *MIDI Linked Data Cloud*, a first step towards interoperable, and interconnected music notation knowledge on the Web. By following best practices within the Semantic Web community, we publish a dataset of 10,215,557,355 triples, representing 308,443 interconnected MIDI scores. We argue that this dataset opens up new research challenges, and can be used to support specific evaluation tasks associated to these challenges in the Semantic Web and Music Information Retrieval communities (Sect. 3).

2 MIDI Linked Data

The MIDI Linked Data Cloud is published at http://purl.org/midi-ld, and provides access to the community, documentation, source code, and dataset. All relevant dataset links and namespaces are shown in Table 1. A GitHub organization hosts all project repositories, including documentation and tutorials, source MIDI collections, and the dataset generation code. The dataset is the result of applying this code to the source MIDI collections, and adding the resources described in this Section. It is accessible as a full dump download, via a SPARQL endpoint, and via a RESTful API. This API is built using SPARQL queries that we publish on GitHub. The dataset includes a VoID description and is registered at Figshare, Datahub, and Zenodo; released under the CC0 1.0 Universal (CC0 1.0) license; and compliant with the FAIR principles [14].

[1] See https://acoustid.org/.
[2] See https://acousticbrainz.org/.
[3] See http://www.music-notation.info/en/formats/NIFF.html.
[4] See http://music-encoding.org/.

Table 1. Links to key resources of the MIDI Linked Data Cloud dataset.

Meroño-Peñuela, A. et al. (2017). *The MIDI Linked Data Cloud.* doi:10.5281/zenodo.579603	
Resource	Link
MIDI Linked Data Cloud dataset	http://purl.org/midi-ld
Portal page	https://midi-ld.github.io/
`midi2rdf`-as-a-Service	http://purl.org/midi-ld/midi2rdf
MIDI Vocabulary namespace	http://purl.org/midi-ld/midi# (prefix `midi`)
MIDI Resource namespace	http://purl.org/midi-ld/ (prefix `midi-r`)
MIDI Notes namespace	http://purl.org/midi-ld/notes/ (prefix `midi-note`)
MIDI Programs namespace	http://purl.org/midi-ld/programs/ (prefix `midi-prog`)
MIDI Chords namespace	http://purl.org/midi-ld/chords/ (prefix `midi-chord`)
MIDI Pieces namespace	http://purl.org/midi-ld/piece/ (prefix `midi-p`)
GitHub organization & code	https://github.com/midi-ld/
Dataset generation code	https://github.com/midi-ld/midi2rdf
Documentation and tutorials	https://github.com/midi-ld/documentation
Source MIDI collections	https://github.com/midi-ld/sources
Sample SPARQL queries	https://github.com/midi-ld/queries
VoID description	http://purl.org/midi-ld/void
Full dump downloads	http://midi-ld.amp.ops.labs.vu.nl/
SPARQL endopint	http://virtuoso-midi.amp.ops.labs.vu.nl/sparql
RESTful API	http://grlc.io/api/midi-ld/queries/
Figshare	https://figshare.com/articles/The_MIDI_Linked_Data_Cloud/4990415
Zenodo	https://zenodo.org/record/579603#.WRluUXV97MU
Datahub	https://datahub.io/dataset/the-midi-linked-data-cloud

Generation, Model, and IRI Strategy. To generate the MIDI Linked Data Cloud dataset we use `midi2rdf` [5]. This algorithm reads MIDI events from a file, and generates an equivalent representation in RDF by mapping MIDI events into the lightweight MIDI ontology shown in Fig. 1 (the complete event hierarchy is included in the documentation). The top MIDI container is `midi:Piece`, which contains all MIDI data organized in `midi:Tracks`, each containing a number of `midi:Events`. `midi:Event` is an abstract class around all possible musical events in MIDI, for example start playing a note (`midi:NoteOnEvent`), stop playing it (`midi:NoteOffEvent`), or changing the instrument (`midi:ProgramChangeEvent`). Concrete events have their own attributes (e.g. a `midi:NoteOnEvent` has a note pitch and veloc-

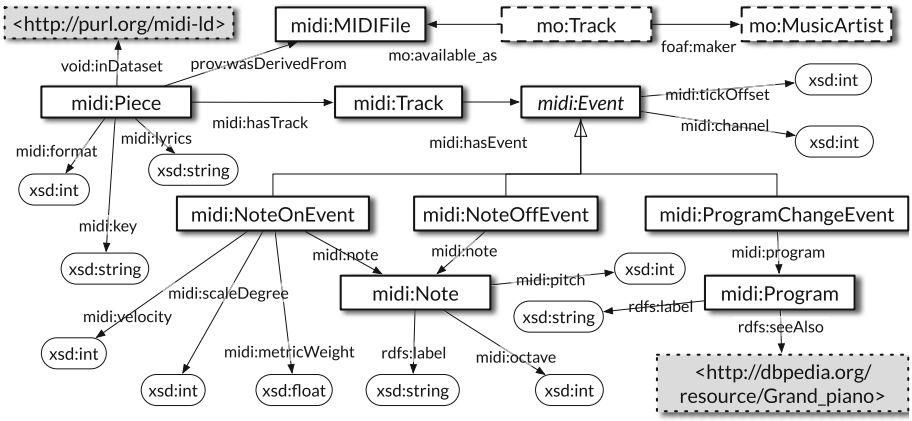

Fig. 1. Excerpt of the MIDI ontology: pieces, tracks, events, and their attributes.

```
midi-p:cb87a5bb1a44fa72e10d519605a117c4 a midi:Piece ;
    midi:format 1 ;
    midi:key"E minor" ;
    midi:hasTrack midi-p:cb87a5bb1a44fa72e10d519605a117c4/track00,
        midi-p:cb87a5bb1a44fa72e10d519605a117c4/track01, ... .
midi-p:cb87a5bb1a44fa72e10d519605a117c4/track01 a midi:Track ;
    midi:hasEvent midi-p:cb87a5bb1a44fa72e10d519605a117c4/track01/event0000,
        midi-p:cb87a5bb1a44fa72e10d519605a117c4/track01/event0001, ... .
midi-p:cb87a5bb1a44fa72e10d519605a117c4/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 1.1. Excerpt of Black Sabbath's *War Pigs* as MIDI Linked Data.

ity), but all event types have a `midi:tickOffset` locating them temporally within the track. Instances of `midi:Track` are linked to the original file (an instance of `midi:MIDIFile`) they were derived from through `prov:wasDerivedFrom`. To enable interoperability and reuse with other datasets, and future extensions, we link the class `mo:Track` of the Music Ontology [8] to the class `midi:MIDIFile` through the property `mo:available_as`. An excerpt of a MIDI file is shown as Turtle in Listing 1.1. IRIs of `midi:Piece` instances have the form midi-r:piece/<hash>/, where `<hash>` is the unique `MD5` hash of the original MIDI. Instances of `midi:Track` and `midi:Event` get IRIs of the form midi-r:piece/<hash>/track<tid> and midi-r:piece/<hash>/track<tid>/event<eid>, where `<tid>` and `<eid>` their respective IDs.

MIDI Sources. The MIDI files in our dataset come from two different sources. The first is a manually curated list of well-known MIDI collections maintained in our GitHub organization. The second is users, who can contribute their own

MIDI files by using the `midi2rdf` algorithm as a service (see Table 1), allowing users to convert their MIDI files to RDF in a browser, and allowing them to get extra links to related Web resources ("this MIDI is my own interpretation of `dbr:Hey_Jude`") and provenance (equipment, purpose, author).

Notes, Programs, and Chords. We publish three additional sets of MIDI resources (see Table 1) that provide a rich description of *MIDI notes* (pitches), *programs* (instruments), and *chords* (simultaneous notes) which in MIDI are expressed simply as integers. MIDI Linked Data notes link to their type, label, the octave they belong to, and their original MIDI pitch value. MIDI programs link to their type, label, and their relevant instrument resource in DBpedia (these have been manually crafted). All tracks link to resources in `midi-note` and `midi-prog`. IRIs in the `midi-chord` namespace are linked to instances of the `midi:Chord` class. Our chord resources (see Table 1) describe a comprehensive set of chords, each of them with a label, quality, the number of pitch classes the chord spans, and one or more intervals –the number of half-steps each pitch class is above the chord's tonic.

Enriching MIDI Files. We enrich the resulting Linked Data with additional features that are not present in the original MIDI files: *provenance*, integrated *lyrics*, and *key-scale-metric* information. To generate provenance, we link the extracted `midi:Piece` with the files they were generated from, the conversion activity that used them, and the agent (`midi2rdf`) associated with such activity. 8,391 MIDI files contained lyrics that were split by syllables, to be used mainly in karaoke software; we join these syllables into an integrated literal, using the `midi:lyrics` property, to facilitate lyrics-based search. Finally, we use the music analysis library `music21`[5] to further enrich the data: the *key* is extracted directly from the MIDI file or automatically detected via e.g. the Krumhansl-Schmuckler algorithm [4]; every note event is represented as the *scale degree* in relation to that key; and we detect and attach *metric* accents for each note (see Listing 1.1).

3 Applications

3.1 Semantic Web Research

Data Integration. The interoperable representation of MIDI as Linked Data can help data integration across music notation databases. Their current formats are incompatible (MIDI, MusicXML, NIFF, MEI, etc.). We envision the study and development of notation *converters*, *vocabularies* and *ontologies* that explicitly specify shared conceptualizations and aid integration. For instance, a simple OWL vocabulary for MusicXML, a format supported by over 210 notation programs,[6], together with a chord structure example[7] can steer integration

[5] http://web.mit.edu/music21.
[6] http://www.ontologydesignpatterns.org/ont/musicml/musicml.owl.
[7] http://www.ontologydesignpatterns.org/ont/musicml/confirmation.ttl.

of MusicXML and MIDI databases. This allows for extended querying in non-proprietary tools and formats, helping to close the gap between produced music signal, song structures, and music metadata. Music notation could also be integrated with related artistic notations such as *dance notation*, for which various machine-readable formats have been proposed, notably LabanXML [3] as RDF. The MIDI Linked Data Cloud can provide a common dataspace for these notations to be linked and integrated, opening up new possibilities for archiving and retrieval, analysis, and choreography and accompanying music generation.

Entity Linking. The MIDI Linked Data Cloud represents musical knowledge with a great level of detail. Concepts such as notes, melodies, chords, and motifs make their appearance as new Semantic Web resources. We must therefore consider the general task of automatically finding new and interesting links among these new entities, and between them and other related entities in the Semantic Web. This has the potential for new challenges in methods for entity linking and link discovery. The novelty our proposed dataset brings to these methods is the potential combination of music *metadata* with *musical content*. A first important task is to generate *quality links between notation and metadata*. For example: how to find relevant links between the score of the Beatles' *Hey Jude* and other Linked Data resources describing this song (e.g. http://dbpedia.org/resource/Hey_Jude)? Relevant metadata for this task, such as the artist name and the song title, can often be extracted from MIDI embedded text[8], further harmonised with the Music Ontology (see Sect. 2), and used to generate links to e.g. MusicBrainz services and the DBpedia musical information graph. Furthermore, through MusicBrainz it becomes possible to retrieve content-based audio features from the AcousticBrainz service, such as its key, tempo or timbre. Such linkage will provide unprecedented queries that combine the full spectrum of musical metadata, features, and notation.

Another group of important links to discover is between *elements of the symbolic notations*, like groups of notes, and other *external symbolic resources*, like chord repositories (see Sect. 3.2). Using our proposed Linked Data chord classification, we devise an algorithm that uses the interval notation to generate links between groups of notes and chords. The results could be used for recognizing chord schemas, representing them in different ways, performing chord substitution adaptation, analysing chord patterns, and consequently searching, mining, matching, and composing them. Linking *notation to audio* is generally more difficult; approaches that map audio to scores [12] are relevant, but could be improved if the target scores are augmented with features correlated to signal. Links in the opposite direction, i.e. from scores to audio, could be generated by using Fast Fourier Transform fingerprinting on sampled MIDI files, but this needs the individual parts (voices) in a piece stored in different MIDI tracks, for which *voice separation* systems are relevant [13]. Finally, an exciting possibility using link prediction would be an *entity classification* task, where we could

[8] See a preliminary algorithm at https://github.com/midi-ld/ner-midi.

remove all `dbr:genre` relations and attempt to predict them, as performed in [9] with predicting critical response to an album in DBpedia.

Semantics and Ontologies. Publishing music notation as Linked Data poses also challenges for semantics and ontologies. Although some of these have been previously addressed (e.g. by the Music, Chord and Timeline ontologies[9]), the *semantics* of musical concepts are still underspecified. For instance, different compositional devices can be used to convey feelings of happiness, sadness, darkness, nostalgia, etc. Similarly, scales, improvisations and motifs transmit different messages based on temporality and interpretation. Explicit, formal specifications of emotions based on musical features could lead to a new generation of creative systems.

3.2 Musicology and Music Information Retrieval (MIR)

Analyses of Chords, Patterns, and Melodies. The MIDI Linked Data Cloud enables a novel, data-driven approach to investigate music in its wider historical, geographical, cultural, economic, and stylistic context. Such combined queries can help musicologists to study influences between composers, investigate the popularity of certain melodic or rhythmic patterns, or understand the use of basic musical building blocks across cultures. It is of obvious musical interest to study the occurrences of resources in the existing vast amount of music repositories, including melody patterns, song collections, chords, and melodies, in music scores. For example, iReal Pro is an open-Web repository of user-contributed chords from thousands of songs, allowing to adapt the key, tempo, and style from these chords. Other repositories contain song melodies, but these are not freely available due to copyright, yet existing MIDI encoding of those melodies could be linked to chords for more functionalities. Another obstacle in reusing these is the lack of established vocabularies for encoding chords, melodies, and their metadata. The MIDI Linked Data Cloud points at addressing these traditional problems of symbolic music analysis: spread and disconnected repositories, incompatible encoding standards, heavily copyrighted databases, and unmanageable high volumes of raw audio data.

Recommender Systems. Analysis of MIDI content as Linked Data could enhance current music discovery and recommendation platforms. Existing systems frequently model similarities between artists as measures for music discovery and recommendation. Similarity models typically rely on collaborative filtering, content-based feature extraction, or a combination of both. Known limitations of these methods (including limited exposure of a collection and lack of high-level descriptions) can be alleviated by adopting Linked Data practices and semantic representations of musical concepts [6], including symbolic analysis of MIDI files. Such analysis could be integrated with other methods, which

[9] http://motools.sourceforge.net/timeline/timeline.html.

take advantage of Linked Data best practices and can enable the identification of musical entities and the discovery of valuable connections between them.

Machine Learning and Music Generation. The most ambitious application of the dataset for machine learning is to learn a *generative model* over knowledge graphs representing music. A generative model represents a class of instances as a probability distribution, allowing us to produce convincing samples of classes of images, natural language, or sound. For knowledge graphs, this task is complicated by our inability to recognize what constitutes a convincing sample. The MIDI Linked Data Cloud presents a solution: a good generative model should produce data that sounds realistic when translated to MIDI samples. This is no simple task: the model should learn that certain triples in the graph represent elements of a stream, and that this stream contains harmonies, a meter and so on. The included metadata should help the model make stylistic choices: a fugue by Bach should not contain tracks with distorted guitars, and a pop song is unlikely to contain more than 8 different tracks.

Linked-Data DJ. Maybe the most simple, though powerful, potential usage of the MIDI Linked Data Cloud is to use formal querying as a language for music mixing. Based on the previously described analytics, a language such as SPARQL can directly be used to filter datasets according to musical properties, such as keys, styles, harmonies, tempo, etc. This provides previously unknown opportunities for mixing and composing new music.

Audiolisation. Visualisations are often very powerful means to help people understand properties of data. Audiolisation, the attempt to use music and sounds to convey meaning, has been applied to algorithms previously but not to data. In order to map audio-expressible features to datasets, the MIDI Linked Data Cloud provides an exciting source for a systematic comparison of both audio and structural features of datasets.

4 Conclusions

This paper presents the MIDI Linked Data Cloud, a linked dataset of more than 10 billion MIDI RDF statements, as a foundation for a common dataspace where musical notation, metadata, and structured music repositories can be linked together on the Web. We have identified its potential for being used in diverse research areas and new challenges in the Semantic Web and Music Information Retrieval communities. We plan to extend the integration, linkage, and ontological methods sustaining this work with in-progress applications for Dutch grants and European funding.

Acknowledgements. We want to express our gratitude to the reviewers for their useful comments; to Frank van Harmelen for his support and motivation; and to Paul Groth for his valuable advice.

References

1. MusicXML 3.0 Specification. Technical report, MakeMusic, Inc. (2015). http://www.musicxml.com/
2. Duckles, V., et al.: "Musicology." Grove Music Online. Oxford Music Online. http://www.oxfordmusiconline.com/subscriber/article/grove/music/46710pg1
3. El Raheb, K., Ioannidis, Y.: A labanotation based ontology for representing dance movement. In: Efthimiou, E., Kouroupetroglou, G., Fotinea, S.-E. (eds.) GW 2011. LNCS, vol. 7206, pp. 106–117. Springer, Heidelberg (2012). doi:10.1007/978-3-642-34182-3_10
4. Krumhansl, C.L.: Cognitive Foundations of Musical Pitch. Oxford University Press, New York (1990)
5. Meroño-Peñuela, A., Hoekstra, R.: The song remains the same: lossless conversion and streaming of MIDI to RDF and back. In: Sack, H., Rizzo, G., Steinmetz, N., Mladenić, D., Auer, S., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9989, pp. 194–199. Springer, Cham (2016). doi:10.1007/978-3-319-47602-5_38
6. Mora-Mcginity, M., Allik, A., Fazekas, G., Sandler, M.: MusicWeb: music discovery with open linked semantic metadata. In: Garoufallou, E., Subirats Coll, I., Stellato, A., Greenberg, J. (eds.) MTSR 2016. CCIS, vol. 672, pp. 291–296. Springer, Cham (2016). doi:10.1007/978-3-319-49157-8_25
7. Raffel, C., Ellis, D.P.W.: Extracting ground truth information from MIDI files: a MIDIfesto. In: ISMIR 2016 (2016)
8. Raimond, Y., Abdallah, S., Sandler, M., Giasson, F.: The music ontology. In: Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, 23–27 September 2007
9. Ristoski, P., Vries, G.K.D., Paulheim, H.: A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web. In: Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F., Flöck, F., Gil, Y. (eds.) ISWC 2016. LNCS, vol. 9982, pp. 186–194. Springer, Cham (2016). doi:10.1007/978-3-319-46547-0_20
10. Swartz, A.: MusicBrainz: a semantic web service. IEEE Intell. Syst. **17**, 76–77 (2002). https://doi.org/10.1109%2F5254.988466
11. The MIDI Manufacturers Association: MIDI 1.0 detailed specification. Technical report, Los Angeles, CA (1996–2014). https://www.midi.org/specifications
12. Thickstun, J., Harchaoui, Z., Kakade, S.: Learning features of music from scratch. arXiv.org Statistics, Machine Learning. https://arxiv.org/abs/1611.09827
13. de Valk, R.: Structuring lute tablature and MIDI data: Machine learning models for voice separation in symbolic music representations. Ph.D. thesis, City University, London (2015)
14. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Nat. Sci. Data **3**(160018) (2016) doi:10.1038/sdata.2016.18