Characterizing the Landscape of Musical Data on the Web: State of the Art and Challenges

Conference or Workshop Item

Characterizing the Landscape of Musical Data on the Web: State of the art and challenges

Marilena Daquino[1] Enrico Daga[1], Mathieu d'Aquin[2], Aldo Gangemi[3], Simon Holland[1], Robin Laney[1], Albert Meroño-Peñuela[4], and Paul Mulholland[1]

[1] The Open University, UK - {marilena.daquino,enrico.daga,simon.holland,robin.laney,paul.mulholland}@open.ac.uk
[2] Insight Centre, NUI Galway, IR - mathieu.daquin@insight-centre.org
[3] National Research Council (CNR), IT - aldo.gangemi@cnr.it
[4] Vrije Universiteit Amsterdam, NL - albert.merono@vu.nl

Abstract. Musical data can be analysed, combined, transformed and exploited for diverse purposes. However, despite the proliferation of digital libraries and repositories for music, infrastructures and tools, such uses of musical data remain scarce. As an initial step to help fill this gap, we present a survey of the landscape of musical data on the Web, available as a Linked Open Dataset: the musoW dataset of catalogued musical resources. We present the dataset and the methodology and criteria for its creation and assessment. We map the identified dimensions and parameters to existing Linked Data vocabularies, present insights gained from SPARQL queries, and identify significant relations between resource features. We present a thematic analysis of the original research questions associated with surveyed resources and identify the extent to which the collected resources are Linked Data-ready.

1 Introduction

Since the early stages of its development, the Web has offered opportunities as a platform to disseminate and exchange information for research and scholarship in the humanities. The digitisation of physical archives, records and other artefacts relevant to humanities research has enabled novel approaches and methods of enquiry that involve computation as a core component, acting on digitized collections of texts, numbers, images, and diagrams [1]. Music research benefits from the same techniques, but offers distinctive additional opportunities due to the powerful affordances for algorithmic analysis, combination, translation and transformation associated with common forms of musical data[5]. For this and related reasons, research in music embraced contributions from Computer Science and AI early [16]. As a result, musical research has benefited from empirical approaches to the study of musical phenomena in which computable formalisations

[5] For example, musical audio can be algorithmically analysed according to cognitive and musicological theories and algorithmically translated in to a wide variety of symbolic notations, and vice versa.

and cognitive models play crucial roles [8]. In recent years, the Web has evolved as an information space consisting not only of linked documents, but also of semantically described resources, following the Linked Data principles: the Web of Data [3]. The opportunities that these developments afford for a variety of musical research activities appear to be substantial. However, the infrastructure to facilitate such opportunities remains scarce and not well understood. To help fill this gap, we survey the status of musical data from the perspective of the Semantic Web, and particularly the emerging Web of Data. We present a survey of the landscape of musical data available on the Web, available as a Linked Open Dataset: the **musoW** dataset of catalogued music resources.

The primary research question is: what is the status of musical data with respect to the Web of Data? Secondarily: to what extent are musical resources ready to be published and linked on the LOD cloud? What types of research and enquiry are musical data meant for, and what direction Semantic Web research should take in order to support them (better)? Through the production of a Linked Open Dataset of musical resources published on the Web, described under a set of key dimensions, we derive a classification of the available data, their nature, form and purpose, and an identification of distinguishing features of the different types of resources. In the light of the gaps of the current landscape with relation to the Web of Data, we identify a set of representative themes in musical research, and formulate hypotheses on how the Semantic Web can help with answering them. Thus we intend to contribute by inspiring possible future directions in Semantic Web developments for the humanities.

We contribute (a) a LOD dataset of catalogued musical resources, as well as a related list of significant SPARQL queries; (b) An analysis of the distinguishing features of each type of data; (c) a set of research themes that are the focus of data oriented musical research, extracted from the corpus; (d) and assessment of the LD-readiness of the resources, by classifying the resources with respect to the 5-Star Web of Data schema.

2 Related Work

In this section we describe existing work addressing the collection of relevant and reusable musical data; the role of musical datasets in interoperable and reusable workflows in Music Information Retrieval (MIR); and the dimensions considered for analysis in existing surveys in the MIR and the Semantic Web communities.

One of the most reused datasets in MIR is the Million Song Dataset [2] (MSD), a "freely-available collection of audio features and metadata for a million contemporary popular music tracks", created to encourage scalability of novel algorithms and provide a benchmark for evaluation. Related to MSD, the Lakh MIDI dataset [14] consists of MIDI files aligned to entries in the MSD. Such alignment is intended to facilitate large-scale music information retrieval, both symbolic and audio content-based. The MusicNet dataset [18] aims at serving "as a source of supervision and evaluation of machine learning methods for music research", and consists of classical music recordings by 10 different com-

posers labelled with instrument/note annotations. Datasets and systems in MIR are sometimes designed without a clear understanding of user requirements. To address this, Lee and Downey [10] conducted a survey in order to "provide an empirical basis" for the development of such datasets and systems, finding that (a) users use collective knowledge (reviews, scores, opinions, etc.) in their music information-seeking; and (b) contextual metadata is of great importance.

Musical data has a key role in the reusability and interoperability of workflows in MIR. Page et al. [13] use the requirements for assisted workflow composition proposed by Gil [7] to study these workflows. A workflow "combines and configures a series of data manipulation and analysis steps into a coherent pipeline" in which data has a primary role [13]. They argue that "for reuse to occur between systems there must also be a mechanism for a mapping of method and workflow between systems, performed through some process of *data exchange*. Aggregation of resources is a common requirement for scientific workflows systems and critical to systems interoperability, reuse, and evaluation in MIR". The Transforming Musicology project [11] aims at developing ontologies for musical concepts and discourse, as well as improving the quality and accessibility of music data on the Web through Linked Data.

In the Semantic Web, dataset descriptions typically deal with only one dataset, and hence domain ontology catalogues and surveys are more relevant to the identification of suitable dimensions. In [12] historical ontologies are classified according to their fit in specific *tasks* in historical research. In [6], authors classify the features of 11 ontology libraries regarding their *scope* and *intended use*, proposing a set of questions to guide the search among them. [9] surveys existing structured languages and ontologies for expressing mathematical knowledge in terms of their coverage of various mathematical representation *requirements*. Surveys of MIR systems (as opposed to datasets) are common, especially regarding methods for analyzing and extracting information from audio and symbolic music notation [17,19]. In [20] authors suggest an evaluation infrastructure based on practices drawn from textual information retrieval. Descriptions of datasets used to evaluate methods are mostly related to benchmarking. For example, in [15] authors have the purpose of creating an "accurate and effective benchmarking system" for MIR systems and consider varying database sizes (from 250 entries to 21,500). The inclusion of methods and software in these surveys influences the dimensions used for analysis. For example, in [15] systems are compared according to their *querying methods*, *extendability*, *ranking* or *partial matching*, which are features found typically in software, but not in datasets. Contrarily, other dimensions, like *file format support* and *database size*, are used to compare their underlying databases. A dimension used in both methods and datasets is *purpose* or *task* [17].

Our survey addresses the shortcomings in existing *musical data collections*, by contributing a more abstract gathering of musical resources available on the Web, and thus with high reusability and a broader scope; in *workflows in MIR*, since we provide a way to find and reuse those resources by means of HTTP dereferencing, as well as a way to facilitate repurposing by specifying what the

original purpose of each of these resources was; and in Semantic Web *dataset surveys*, by borrowing analysis dimensions from at least three sources: the Semantic Web generally, ontologies from humanities research, and terms from MIR research generally.

3 Methodology

In this paper, we assess the status of musical data in the Web of Data, and discuss the potential contributions of the Semantic Web to support music research. To promote reliability, we focus in the first instance on sources derived from musical research and scholarship. The workflow of our assessment methodology is as follows:

1. We design a set of *ad hoc* dimensions to describe the resources of the domain, and we use these dimensions to describe the resources in a table;
2. We survey the following musical resources: repositories, digital libraries, datasets, catalogues, projects, digital editions, services, software, formats, schemas and ontologies;
3. We map dimensions and parameters to well-known Linked Data vocabularies; and we produce **musoW**, a Linked Data dataset describing all these resources (Section 4);
4. We query this dataset in SPARQL to draw an overview of the collection and gain insights (Section 4);
5. We conduct a statistical analysis and identify significant relations between resources features (Section 5);
6. We conduct a thematic analysis of the research questions associated with the above resources (Section 6);
7. We analyse the results in the light of the five-star Linked Data principles using Formal Concept Analysis (Section 7).

In the remainder of this Section we describe the creation of the musoW dataset and the criteria for its analysis.

The survey is designed with the perspective of potential applications of musical data in the Semantic Web, and its targeted users are researchers. To gather our collection, we relied on resources created and used by researchers, and retrievable using online aggregators -which also target researchers. We look for research objects already evaluated in musicology, ensuring their reliability, and establishing a reproducible gathering criterion. We scraped these online aggregators to retrieve names of projects, URLs and descriptions. We only extract resources providing digitizations or transcriptions of scores, audio performance audio, and, eventually, a critical apparatus of notated music. We excluded materials for theoretical studies - such as literature, archives and libraries of resources not available online - and collections of learning materials - e.g. audio and video courses.

To implement the survey we designed a set of 46 dimensions to describe these objects, and we created a table in which such dimensions are columns

whose values are validated by controlled vocabularies. A subset of dimensions is applicable to all the types of resources: *resource ID*, *URL*, *description*, *project affiliation*, *search criterion*, *resource type*, *reused resources* (or connection to other projects), *purpose* (learning or research), *access restrictions*, *licenses*, *situation* or *task*, and t*arget audience*. Another subset applies to data collections only (repositories, digital libraries, datasets, catalogues, and digital editions), and includes: an item *example*, *gathering criteria* of collections (genre, artist, temporal or geographical), *subject terms* from both Music Ontology[6] and a local controlled vocabulary, a *list of services* offered by the resource (data dump, browsable interface, queryable interface, API, SPARQL endpoint), *collection size*, *data size*, what features of symbolic notation (melody, harmony, rhythm, timbre, contour or shape, structure of a song, descriptive metadata) are provided as *structured data* (if applicable), *formats* and their *interoperability*. Most of these dimensions are shared also with schemas, ontologies, services, software and formats, except the ones for describing the scope of contents.

Part of the purpose of the survey is to better understand the existing aims of music researchers. Consequently, as well as describing and linking existing research data, we cached the research questions associated with each surveyed resource, where explicitly available or easily inferred from project documentation on the web[7].

In order to assess the extent to which musical data conforms to the Web of Data principles, characterize the landscape of musical data, and make emerge potential gaps and opportunities for further research, we chose to observe the corpus under four perspectives:

(1) Quantitative. The role of a quantitative analysis is of illustrating the musoW dataset in numbers, by aggregating items with respect to the different dimensions, therefore giving a picture of the musical data landscape. We observed the dimensions of the datasets and formulated a set of questions related to them. We implemented those as SPARQL queries, we report the major findings in Section 4.

(2) Statistical. We performed statistical analysis in order to understand some of the relationships between the dimensions. Our analysis focussed on answering questions related to the size and resource types of the collections and how that related to their scope and musical features.

(3) Thematic. We performed a thematic analysis of the research questions associated with the surveyed resources. This involved coding the statements contained in the research questions and then clustering the codes into a series of emerging themes related to music research.

(4) LD-Readiness. We analysed the data to assess to what extent the collected resources are LD-ready. The 5-Stars Open Data development scheme identifies five key dimensions of open data[8]: Open Licence (OL), Machine read-

[6] http://musicontology.com

[7] When research questions could not be evidenced directly, we provided keywords summarizing our best understanding of the purpose, task or situation.

[8] 5-Star Open Data: `http://5stardata.info/en/`

able (RE), Open format (OF), Adoption of URIs (URI), and Linked Data (LD). We therefore generated these five derived dimensions from the collected data. We analysed the resulting data using Formal Concept Analysis (FCA) with the Contento tool [5].

4 The musoW landscape

The musoW dataset is available online [9] and the content can be queried in SPARQL through the data.open.ac.uk endpoint[10]. To make this analysis reproducible, we publish the SPARQL queries in which it is based. Query identifiers are reported below in squared brackets. To facilitate access to the results of these queries by any application, we also publish an equivalent RESTful API[11]. The collection includes 351 resources: 187 repositories and digital libraries, 44 catalogues and 3 projects, 36 datasets, 21 digital editions, 22 software, 14 services, 12 ontologies, 2 schemas, and 3 formats.

Repositories and **digital libraries** are the most representative resources collecting musical data. They mainly offer digitisations of scores and lyrics (77%) [**DR6**], published as PDF (62%) and/or JPG (40%) [**DR4**]. Audio records are provided by 29% of repositories [**DR7**], as MIDI files (45%) and/or MP3 (29%) [**DR4**]. Only 20% of repositories offer structured data on symbolic music notation [**DR8**], representing melody (95%), rhythm (76%), harmony (74%), structure of a song (46%), timbre or contour (less than 10%) [**DR14**]. The most used formats are here MusicXML (46%), custum XML (23%) and MEI/XML (10%). The more the scale of repositories increases, the less structured formats for representing symbolic notation seem to be used [**DR4**] and the less depth of analysis is provided [**DR13**][12]. We mainly found items belonging to the same musical genre (64%), and/or to the same country (39%) and/or falling within the same period (31%). National projects seem to afford dealing with large amounts of data, while small projects narrow the scope to a single genre. **Datasets** are the second most represented category of resources, mainly available under Creative Commons licenses and in several interoperable formats, such as RDF (44%), JSON (14%), TXT (11%), XML and CSV (less than 10%), and others [**DS2**]. The scope is heterogeneous, and doesn't provide any insight on a particular or shared interest [**DS4**]. Instead, purposes and tasks seem to be the gathering criterion: mainly focusing on research goals (89%) [**DS10**], the aim is to make improvements in music analysis (28%), music information retrieval (22%) - including more specific tasks like genre recognition, score-audio linking, and machine learning. Few ones are targeted to disciplines like musicology and history of music (11%) [**DS15**].

[9] musoW: `https://github.com/enridaga/musow`.
[10] Named Graph: `http://data.open.ac.uk/context/musow`.
[11] API: `http://grlc.io/api/albertmeronyo/mudow-queries`
[12] Moreover, several large-scale repositories offer queryable structured data only for incipits of works, even though we included them in the category of such kind of data providers.

Among the RDF datasets, the focus on descriptive metadata of music is predominant (75%), while 19% represent features extracted from audio data, and only one deals with features extracted from notated music [**LD1**]. To explain this, we look at tasks motivating the realisation of such datasets, finding that there is a common need of publishing a specific kind of data otherwise not available in other data sources (44%) - e.g repositories generally don't offer a data dump - and aggregate it with information from similar datasets (25%), e.g. to enable research in domains like history of music and musicology (25%). Furthermore, music analysis (19%) and music information retrieval (12,5%) seem to find in LOD a testing bench [**LD3**]; but only one dataset is reused by other music related projects [**DS12**]. Finally, data dumps are the most common way to publish data (75%), while only 37,5% offer a SPARQL endpoint [**LD2**]. **Digital editions** generally offer small-scale collections of musical data [**DE4**]. They mainly deal with scores of a single artist (76%) or contemporary related groups of artists (less than 33%) [**DE3**]. Less than 38% offer structured data on symbolic notated music [**DE7**]. Still, the most used formats are JPG (47%), PDF (33%), MEI/XML and MP3 (23%) [**DE2**]. The main goal of such resources is to give a contribution in fields like musicology (86%) and history of music (76%). A shared concern regards visualization of complex information like variants and genetic of music. Few tools have been developed in order to support tasks like annotation and visualisation. MEI/XML files are generally the preferred input [**DE11**]. **Services** and **software** are here mainly considered because of their task and possibilities of reuse. We mainly found tools for annotating music (25%), and enabling further analysis in research fields like musicology (25%), music information retrieval, history of music (19%) and music philology related issues (11%) - e.g. Optical Music Recognition, music style analysis, measure annotation. Secondly, as already revealed when describing digital editions, data visualisation is a shared concern (14%) [**SS3**]. 75% of such tools deal with a structured representation of music features, such as melody and rhythm (100%), harmony (96%)[**SS4**]. 64% of software/services extract music features directly from notated music, rather than audio tracks (33%) [**SS5**]. **Ontologies** and **schemas** do not offer insights of a shared need in knowledge representation at this stage of the analysis. Indeed, 36% deal with the representation of features extracted from audio tracks, 36% from descriptive metadata (e.g. cataloguing information of songs, artists, genres), and 21% from symbolic notation [**SO2**]. There are no evidences of a clear and shared approach to represent music knowledge extracted from audio/scores. In fact, none of the proposed models are reused in other projects than the one where they were born in[13] [**SO4**].

5 Statistical analysis

In order to understand some of the relationships between the dimensions of the survey, a statistical analysis was conducted, focussed on answering the following

[13] Except the Music Ontology, which does not represent any features of notated music or audio files.

questions: 1) Is there a relationship between the size of the collection and the types of resources it holds? 2) Are there any relationships between the musical features represented (e.g. lyrics, rhythm) and size and resource type of the collection? 3) Are there any relationships between the defined scope of the collection (e.g. by time period, artist, genre, geography) and its size and resource type?

The relationship between size and type was analysed. To ensure sufficient cell sizes in the analysis, collection size categories were merged (<100, <1000, >1000) and analysis was restricted to the four most prominent resource types (catalogues, digital libraries, digital editions and repositories). A significant interaction was found between size and type (Fisher exact test, $p < 0.01$). Essentially, digital editions tend to form smaller collections than the others. Multinomial regression analysis was used to test if musical features could predict resource type. Due to cell sizes, this was restricted to the following features: melody, rhythm, lyrics and structure. Lyrics are more likely to be found in software ($B = 2.183$, $p < 0.01$) and datasets ($B = 1.448$, $p < 0.05$) but less likely in digital libraries ($B = 2.125$, $p < 0.05$). Rhythm is more likely to be represented in digital editions ($B = 2.823$, $p < 0.05$), software ($B = 4.530$, $p < 0.01$) and datasets ($B = 3.040$, $p < 0.01$). Structure is more likely to be represented in software ($B = 1.680$, $p < 0.05$) and datasets ($B = 1.711$, $p < 0.01$). Ordinal regression analysis was used to test if musical features could predict collection size. Larger collections are more likely to feature melody (Wald = 4.178, $p < 0.05$). Multinomial regression analysis was also used to test if the defined scope (e.g. by genre or artist) could predict resource type. Digital editions are more likely to be scoped in terms of artist ($B = 2.655$, $p < 0.01$). Software ($B = 2.810$, $p < 0.01$) and datasets ($B = 1.022$, $p < 0.05$) are less likely to be scoped in terms of genre. Ordinal regression analysis was used to test if scope could predict collection size. Smaller collections are more likely to be scoped in terms of artist (Wald = 28.359, $p < 0.01$) or genre (Wald = 7.362, $p < 0.01$) than larger collections.

We can see overall that: (1) there is a relationship between resource type and size; (2) musical features are more or less likely to be represented in a collection depending on its size and resource type; (3) there are relationships between the scope of the collection and its size and resource type.

6 Thematic analysis

For 37 of the projects a textual description of the research question or questions to be answered using the dataset was identified. In order to characterise the range of issues raised in the research questions, a thematic analysis [4] was conducted in which a set of codes for describing the text were formulated bottom-up from multiple readings of the questions. The codes were then clustered around a series of emerging themes.

Of particular interest is the types of musicological inquiry identified from the projects. These are: finding out what a class of objects (such as blues songs) have in common; understanding changes in music over time (e.g. the time the piece was written or the biblical period the piece is about); analysis of different

versions or editions of the same piece and how they vary; analysis of heterogeneous resources associated with the same theme (e.g. documents and data about jazz artists); comparing how people work with digital versus analogue artefacts; and contrasting classes of work (e.g. Chopin versus others).

Projects aimed to develop support for different forms of activity such as research, teaching and performance. Research aims were concerned with supporting different types of music content publishing such as rendering visual scores from some underlying machine readable format and indexing scores according to this format. Research also aimed to publish musical artefacts (such as scores) with some form of associated scholarly interpretation. Some projects had a research goal to construct an archive, but of different types of material such as scores, recordings, ephemera and libretti.

7 LD-Readiness

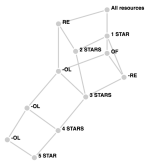
We now report on the evaluation of the collected resources with respect to the 5 Star Open Data paradigm. The 5 Star Open Data scheme includes five level of compliance with the Web of Data.To map the **musoW** catalogue with this scheme we generated five derived dimensions with the following criteria:

OL Open Licence. The resource is publicly accessible with an open access licence (e.g. CC-BY, CC0, OGL), also if only for human consumption.
RE Machine Readable. The resource contains structured data published in a machine readable format (although it can be a proprietary one). Resources being published in any interoperable format or through Web APIs are considerable to be machine readable.
OF Open Format. The resource is published in an open standard (e.g. CSV)[14].
URI URIs. The resource makes use of Uniform Resource Identifiers (URIs) to identify the described entities. We derived this dimensions for all the resources expressed in RDF or related vocabularies (OWL, SKOS).
LD Linked Data. We considered here resources published in RDF using a SPARQL endpoint[15].

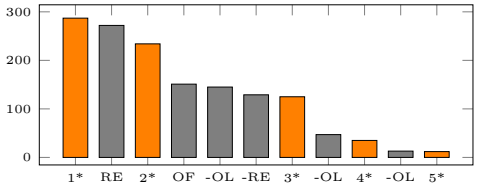
We built a FCA formal context including the catalogue items and the five derived dimensions: OL, RE, OF, URI, LD. Following the FCA approach, we generated a concept lattice and labelled the concepts from one to five stars using the Contento tool [5], obtaining the lattice depicted in Figure 1a. The top of the lattice is the concept including all 327 resources. The first layer includes three concepts: within these we find the 287 resources published with an Open Licence, therefore belonging to the 1-Star group. This concept branches in two directions, one

[14] We inspected the data formats and included here all the resources having well-known formats, for example 'midi', 'musicxml', 'json', 'mei/xml', or 'tei/xml'.
[15] Although we did not verified whether they were actually linked or not to the LOD cloud. However, resources in this group can be considered LD-ready, in the sense that links could be established between those and other LD resources.

intersecting the resources published in a machine readable format (the RE concept, also including some resources without an open licence): the 2-Stars group, 234 resources being published with an open licence in a machine readable format. Following this path we proceed meeting the resources published also in an open standard (3-Stars, 125 resources), and the ones using URIs (4-Stars, 35). The bottom of the lattice includes the 5-Star resources (12) - the ones having a SPARQL endpoint and therefore being ready to be queried and linked to Web of Data. It is interesting to notice that the FCA lattice makes emerge also a good amount of resources that, while adopting open standards or semantic technologies (RDF, SPARQL), are not published with an open license (the concepts tagged '-OL' in picture 1a).



(a) The FCA annotated lattice developed for the LD-readiness analysis



(b) Distribution of resources with respect to the 5* scheme.

8 Discussion

Although some resources are ready to be linked to the Web of Data, the majority of resources are left behind (see Figure 1b). The lack of an open license associated with the data or collection seems to be a generalized issue, a non-technical limitation that nevertheless hinders the reuse-ability of the resources. We notice that the more the scale of repositories increases, the less structured formats for representing symbolic notation seem to be used. This emerged especially in data sources coming from National projects, that can afford dealing with large amounts of resources, and it might be appointed to a heterogeneity of resources' typologies. Dataset are focused on specialized research tasks (e.g. in the context of MIR), but most of them include metadata rather then musical content expressed symbolically. Although software and services for semantic lifting of musical content exists, they are not applied to large repositories or reused outside the original context, often part of small sized digital editions. These observations suggest the need of a reusable and scalable workflow to support the life cycle of musical data on the Web. More importantly, there is a lack of understanding about what kind of life cycle musical data could have on the Web, and whether it would be possible to support it with systematic approaches.

Observing digital editions, we considered that tool support for annotation, exploration and visualization of musical corpora it's still at its infancy, and we

argue that Semantic technologies can have a role in the way musical content can be abstracted and organized for browsing and exploration.

We also notice the opportunity of Linked Data within music on two issues of authority: one related to notes the experts know are wrong; and the other where experts disagree, e.g. because of lack of original score, or poor handwritten originals. This shows an overlap with trust from Web data in general, for which some Linked Data approaches could be of use. In particular, the description and publication of provenance of musical research objects using the PROV vocabulary[16], and the sharing of annotations on top of musical resources using the Open Annotations Data Model[17], could fill the gaps in these issues.

In the light of the scarcity of Linked Data resources, at least concerning the publication of music notation as Linked Data, for which we could only find one resource, there is a long way to go with respect to the reusable, repurposable and interoperable workflows that have been proposed in MIR [13] and musicology [11].This can be the result of a cultural issue, as most of the research in musicology does not happen to be initiated with the data publishing as core objective. However, also inline with the Open Science paradigm, we can foresee that there will be the need of new models to support the diversity of musical knowledge on the Web.

9 Conclusions

In this paper, we surveyed the landscape of musical data on the Web and presented the musoW dataset, a Linked Open Data catalogue of musical resources published on the Web with the purpose of supporting musical research and scholarship. We observed that a large amount of resources are not ready to be part of the Web of Data, and the main obstacles are due to the heterogeneity of large collections, the uncertainty in licensing, and the lack of large scale approaches to semantic lifting of musical resources and data publishing. Ultimately, it is relevant to notice a cultural bias in the distribution of how musical features are represented. In fact, larger collections are more likely to feature melody, reflecting clearly a Western-centric point of view. As all Web material, we can observe this will have issues with representative sampling and quality that could be interesting to investigate further. Furthermore, thanks to the musoW dataset, we were capable of identifying a set of unexplored opportunities for Semantic Web technologies. Future work includes the enhancement of the resources descriptions with the results of the analysis, and support the exploration of the dataset with visualizations. For example, we intend to augment the musoW catalogue with a classification of the resources with respect to research tasks. Finally, we intend to study how pragmatically the musoW dataset can support musical researchers in the discovery and adoption of Web data, for example linking the collection to prototypical workflows for musical enquiry.

[16] W3C PROV-O: `https://www.w3.org/TR/prov-o/`

[17] Open Annotation Data Model (Community draft): `http://www.openannotation.org/spec/core/`

References

1. Berry, D.M.: The computational turn: Thinking about the digital humanities. Culture Machine 12 (2011)
2. Bertin-Mahieux, T., Ellis, D.P., Whitman, B., Lamere, P.: The Million Song Dataset. In: Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011) (2011)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Semantic services, interoperability and web applications: emerging concepts pp. 205–227 (2009)
4. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative research in psychology 3(2), 77–101 (2006)
5. Daga, E., d'Aquin, M., Motta, E., Gangemi, A.: Bottom-up ontology construction with contento (2015)
6. d'Aquin, M., Noy, N.F.: Review: Where to publish and find ontologies? a survey of ontology libraries. Web Semant. 11, 96–111 (Mar 2012)
7. Gil, Y.: Workflow Composition: Semantic Representations for Flexible Automation, pp. 244–257. Springer London, London (2007)
8. Honing, H.: On the growing role of observation, formalization and experimental method in musicology (2006)
9. Lange, C.: Ontologies and languages for representing mathematical knowledge on the Semantic Web. Semantic Web – Interoperability, Usability, Applicability 4(2), 119–158 (2013)
10. Lee, J.H., Downie, J.S.: Survey Of Music Information Needs, Uses, And Seeking Behaviours: Preliminary Findings. In: Proceedings of the 5th International Conference on Music Information Retrieval. Barcelona, Spain (October 10-14 2004)
11. Lewis, R.J., Crawford, T., Lewis, D.: Exploring information retrieval, semantic technologies and workflows for music scholarship: the Transforming Musicology project. Early Music 43(4), 635–647 (2015)
12. Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., van Harmelen, F.: Semantic Technologies for Historical Research: A Survey. Semantic Web – Interoperability, Usability, Applicability 6(6), 539–564 (2015)
13. Page, K.R., Fields, B., Roure, D.D., Crawford, T., Downie, J.S.: Capturing the workflows of music information retrieval for repeatability and reuse. Journal of Intelligent Information Systems 41(3), 435–459 (Dec 2013)
14. Raffel, C.: Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching. Ph.D. thesis, Columbia University (2016)
15. Reiss, J., Sandler, M.: Benchmarking Music Information Retrieval Systems. In: The MIR/MDL Evaluation Project White Paper Collection. vol. 3, pp. 43–48 (2003)
16. Roads, C.: Research in music and artificial intelligence. ACM Computing Surveys (CSUR) 17(2), 163–190 (1985)
17. Schedl, M., Gómez, E., Urbano, J.: Music information retrieval: Recent developments and applications. Foundations and Trends® in Information Retrieval 8(2-3), 127–261 (2014)
18. Thickstun, J., Harchaoui, Z., Kakade, S.: Learning Features of Music from Scratch. ArXiv e-prints (Nov 2016)
19. Typke, R., Wiering, F., Veltkamp, R.C.: A Survey of Music Information Retrieval Systems. In: ISMIR (2005)
20. Urbano, J., Schedl, M., Serra, X.: Evaluation in music information retrieval. Journal of Intelligent Information Systems 41(3), 345–369 (Dec 2013)