Building A Graph of Linked Musical Works

Konstantin Todorov

LIRMM, University of Montpellier, CNRS,
161 rue Ada, 34095 Montepllier, France
todorov@lirmm.fr

The Linked Open Data (LOD) project[1] and the semantic web in general offer technological means for data reuse, increased visibility and data sharing on the web, data federation and facilitated exchange of metadata by the creation of links across resources. Attracted by these possibilities, many major actors from the library world, such as the Library of Congress (LOC) or the French National Library (BnF), have embraced semantic web technologies with the goal to open their archives and catalogs to the web. This process has resulted in a number of openly available and explorable RDF graphs reflecting the rich content of numerous libraries and cultural institutions from all over the world [1].

The DOREMUS project follows this line of research and practice, with a particular interest in classical and traditional music.[2] Three major French cultural institutions—the BnF, Radio France and the Philharmonie de Paris—have joint efforts with data and social science academics in order to develop shared methods to describe semantically their catalogs of music works and events and open them to the web community. A major contribution of the project is the development of the DOREMUS ontology[3], which extends established models for representing bibliographic information and adapts them to the domain of music, thus filling an important representational gap. A number of shared vocabularies about music-specific concepts (such as musical genres or instruments) have been collected or developed, linked and published using the SKOS standard. The data from the catalogs of the three institutions comes in MARC or XML formats. Specific tools for data conversion to RDF following the DOREMUS ontology have been developed. This process results in the construction of several knowledge graphs about music works and events, which have been linked using a specifically developed for this purpose data linking protocol and tool [2].

The datasets that are subject to interlinking in the musical field are highly heterogeneous: a given entity (e.g., a musical work) can be described quite differently across different institutions. In addition to well-known data discrepancies such as lexical, semantic (polysemy, synonymy) and orthographic mismatches of string literals, the use of acronyms and abbreviations or differences in formats and types of numerical values, we have encountered several commonly occurring issues that are specific to musical data. We outline some of them below.

— *Differences in coverage* and particularly lack of information in one of the graphs as compared to a richer description in another. In our case, the works

[1] `https://lod-cloud.net`
[2] `http://www.doremus.org`
[3] `http://data.doremus.org/ontology/`

coming from Radio France are systematically described by a considerably smaller set of attributes, than those found in the catalogs of the other libraries.

— *Different depths in the graphs*, at which we find the value of interest—e.g., the birthplace of a composer can be directly assigned to the entity in one graph, or via a longer property chain in another.

— Presence of *comments in the form of free text* that are difficult to compare, as well as presence of *institution-specific resource identifiers* (bibliographical records ID's) given under the same property name across different datasets, although not comparable.

— Presence of *blocks of highly similar in their descriptions, but yet distinct instances* in each of the graphs—e.g., the set of all piano sonatas by Beethoven, differing from one another in only one or two property values, which makes their disambiguation difficult and is likely to produce false positives.

The limited results obtained with state-of-the-art linking systems [3,4] pushed us to the development of a new linking approach and an open source tool, named *Legato* [5].[4] The proposed solution covers an important number of data heterogeneities and attempts to reduce the user configuration effort. It is based on: (i) *Property filtering*, or automatic data cleaning of "problematic" attributes; (ii) *Instance profiling* allowing to represent each resource by a sub-graph considered relevant for the comparison task; and (iii) *Instance vector representation* allowing to compare resources. To reduce the false positives rate, we apply a (iv) *Post-processing* step based on hierarchical clustering and key ranking techniques aiming to disambiguate highly similar, though not identical instances.

The current talk goes through a brief overview of the state-of-the-art of data linking approaches and systems, outlining major challenges and directions of research in the area. We then focus on data linking issues raised by the specificities of music bibliographical data using the DOREMUS use-case. We expand on the solutions that we propose and the characteristics of the linking tool *Legato*. We present evaluation results on real-world music meta-data benchmarks, as well as other generic reference datasets designed for data linking evaluation purposes (data released by the Ontology Alignment Evaluation Initiatives 2016 and 2017).

Acknowledgments

This work has been partially supported by the French National Research Agency within the DOREMUS project, under grant ANR-14-CE24-0020.

References

1. J. Marden, C. Li-Madeo, N. Whysel, and J. Edelstein, "Linked open data for cultural heritage: evolution of an information technology," in *ICDC*, 2013.
2. M. Achichi, P. Lisena, K. Todorov, R. Troncy, and J. Delahousse, "DOREMUS: A graph of linked musical works," in *The Semantic Web - ISWC*, pp. 3–19, 2018.
3. A. Jentzsch, R. Isele, and C. Bizer, "Silk-generating RDF links while publishing or consuming linked data," in *ISWC*, 2010.

[4] `https://github.com/DOREMUS-ANR/legato`

4. A. N. Ngomo and S. Auer, "LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data," in *IJCAI*, pp. 2312–2317, 2011.
5. M. Achichi, Z. Bellahsene, and K. Todorov, "Legato: Results for OAEI 2017," in *Ontology Matching*, 2017.