Undefined 0 (2012) 1–0
IOS Press

Hello Cleveland! Linked Data Publication of Live Music Archives

Sean Bechhofer [a,*] David De Roure [b] Kevin Page [b]
[a] *School of Computer Science, The University of Manchester, United Kingdom,*
sean.bechhofer@manchester.ac.uk
[b] *Oxford e-Research Centre, University of Oxford, United Kingdom*
{kevin.page,david.deroure}@oerc.ox.ac.uk

Abstract. We describe the publication of a linked data set exposing metadata from the Internet Archive Live Music Archive. The collection provides access to recorded performances and is linked to existing musical and geographical resources. The dataset contains over 17,000,000 triples describing 100,000 performances by 4,000 artists.

Keywords: Linked Dataset, Music, Internet Archive

1. Introduction

The Internet Archive Live Music Archive [12] (further referred to here as LMA) is an online resource providing access to a large community-contributed collection of live recordings. Covering nearly 4,000 artists, chiefly in rock genres, the archive contains over 100,000 live recordings made openly available with the permission of the artists concerned. Audio files are available in a variety of formats, and each recording is accompanied by metadata describing information about dates, venues, set lists, the provenance of the audio files and so on.

From a musicological perspective, the collection is valuable for a number of reasons. First of all, it provides access to the underlying audio files. Thus the LMA provides a corpus that can be used for Music Information Retrieval (MIR) [3] tasks such as genre detection, key detection, segmentation and so on as exemplified by the MIREX series of workshops [5]. It provides multiple recordings by individual artists[1] allowing comparisons across performances. Furthermore, in live situations artists will frequently play works by other artists ("covers"), providing source content for cover detection algorithms [11].

Collection generation and management is a key starting point for research in the digital humanities. An earlier prototype [15] demonstrated how Linked Data can be applied to the MIR research process and the utility of this approach, particularly when gathering and managing corpora of source audio; however, this system re-used pre-existing Linked Data that described the recordings to populate its collections. As computational analysis increases in scale through projects such as SALAMI [4], so too does the value of republishing existing large repositories such as the LMA using Linked Data: as it stands, however, extracting subcollections from the archive is not a straightforward task. Metadata is largely published as free text fields, with heterogeneity in detail and inconsistency in content. Providing structured metadata (with links to external resources) will, we believe, facilitate the activities such as the production of sub-corpora for experiments or evaluation.

We describe an exercise in republishing LMA metadata using a linked data [2] approach. We believe that publishing the collection metadata as linked data brings benefits in supporting query and integration

*Corresponding author

[1] In the case of the Grateful Dead, an act that for many years encouraged audience taping of performances, the LMA contains over 8,000 recorded performances.

with existing sources. Note that we are dealing here purely with the *metadata*, and leave the audio source files untouched (but provide links to the online resources). Enhancing the collection with additional information gained through analysis of those audio files is left to future work.

2. Approach

The collection is published using a *layered* approach. The core metadata describing the resources is essentially published "as is". Raw data provided by LMA is translated to an RDF form, using appropriate vocabulary terms (for example, the label associated with a particular performance is represented using `skos:prefLabel`). Additional information asserting mapping relationships to other collections such as MusicBrainz [14], GeoNames [6] or last.fm [9] is then added. For example, although attempts could have been made to reconcile artist names as used in the collection, this is not achieved through modification of the *core data*. This method allows us to explicitly record provenance information about how the associations were derived, which in turn then allows consumers of the data to make decisions about whether or not to use or trust the relationships asserted. It is thus clear to any consumer of the data whether information has come directly from LMA or is additional information provided via our process. We believe that such an approach is needed for a collection like this, where the data is not simply "asserted truth", but has some subjectivity.

Note also that our approach explicitly avoids the use of `owl:sameAs` triples to relate LMA entities to external entities (e.g. within MusicBrainz or GeoNames) and instead uses a pattern from the Similarity Ontology (see Section 4). This is useful in supporting processes such as artist reconciliation, where matching between recording artists may be a somewhat fuzzy process – groups or bands can be volatile organisations in terms of membership (although they do not appear in the LMA, The Fall's "revolving door of musicians" is a case in point [19]) and we may not always have confidence that descriptions of an artist from different sources match exactly. Data consumers then have the option of using the encoded raw source data or the additional layer of mapped relationships. Furthermore, layering allows for the possibility of further analysis and re-publication of mapping relationships from the source corpus as new techniques and corrections are developed; this could include publication community driven curation and validation. Explicit separation of the "ground facts" from the alignments also allows for revisions and additions to those alignments.

3. Modelling

The collection contains a number of basic entities.

Artist The performer of a concert/gig. There is no attempt to link or split up bands, duos, etc.

Performance A particular concert/gig that has been recorded. Each performance is given a unique identifier by LMA.

Track Performance of a particular track within a Performance

Venue Location where a performance takes place.

Each *Artist*, *Performance*, *Track* and *Venue* is minted a URI in the collection namespace[2] with an appropriate path prefix. A number of ontologies are used for the description of entities.

Music The Music Ontology [16,18] provides terms that describe performances, artists and the relationships between them.

Events The Event Ontology [17] provides terms for describing events.

Similarity The Similarity Ontology [7,8] provides terms for asserting associations between entities. This is used to associate artists in the collection with MusicBrainz ids, and locations with last.fm venues and GeoNames entities.

SKOS SKOS [13] labelling properties are used to label entities.

PROV-O The W3C provenance ontology [10].

VoID The W3C dataset metadata ontology [1].

etree An ontology[3] that defines subclasses of Music Ontology classes and specific properties used in the etree metadata.

The basic modelling pattern used within the data set is shown in Figure 1. In the figures, green, unlabelled links are `rdf:type`. Blue, unlabelled links are `rdfs:subClassOf`. The ontology used to describe the collection is relatively inexpressive, essentially providing classes for performances and venues and properties for the assertion of values and relationships.

[2] `http://etree.linkedmusic.org`
[3] `http://etree.linkedmusic.org/vocab`

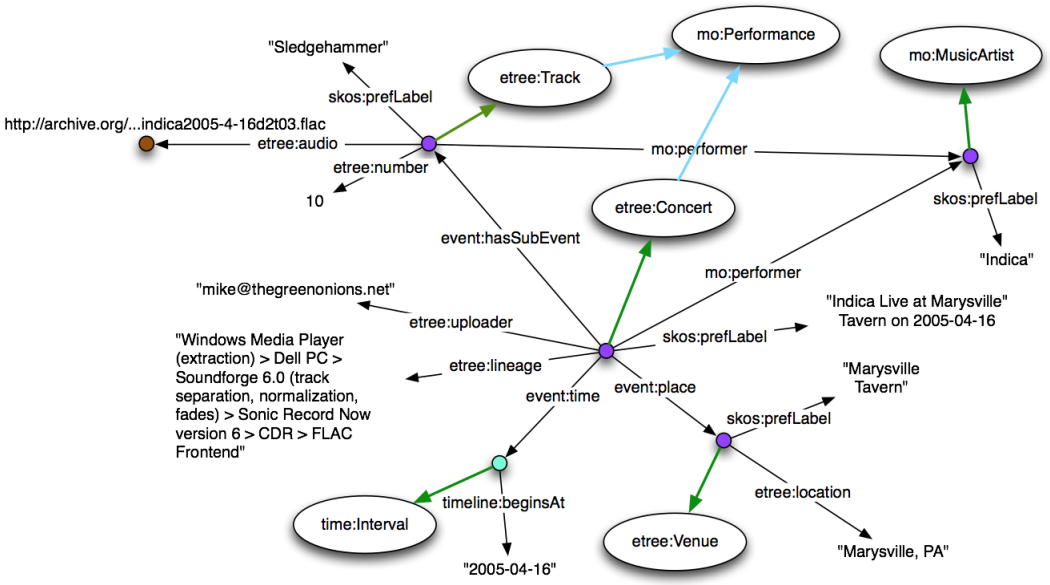

Fig. 1. Basic Model

4. Record Linkage

The collection offers possibilities for record linkage with external datasets. In particular, music artists and geographical locations are entities that are described in a number of external data sources (many of which are also published as Linked Data).

Artist Alignment MusicBrainz [14] provides an "open music encyclopaedia" and provides identifiers for a large number of music artists. MusicBrainz is a clear candidate for linking from a collection like LMA. Queries to MusicBrainz taking exact matches on names provides a simple alignment between our artists and MusicBrainz, covering 1,168 of the 3,981 artists in the collection. In keeping with the strategy outlined in Section 2, the relationships between the artists and MusicBrainz are asserted using the Similarity Ontology. Figure 2 shows how these relationships are asserted.

The Music Ontology considers `mo:MusicArtist` to encapsulate "A person or a group of people [...], whose musical creative work shows sensitivity and imagination" and the current dataset makes no distinction between solo artists and bands/groups of musicians. There is no information in the source corpus that distinguishes between solo artist and group or identifies relationships between, for example, a singer and a band. For our initial purposes, identifying "artist" is sufficient. Mappings to MusicBrainz may allow for further identification of groups or solo artists and a refinement of the types applied (for example asserting that a resource is in fact a `mo:MusicGroup`).

Track Alignment In the current dataset, tracks are not aligned to external data sources. See Section 9 for further discussion.

Geographical Alignment Performances occur at a particular place[4] and can thus potentially be mapped to geographical locations in collection such as GeoNames. Concert performances also tend to take place in specific venues (theatres, concert halls etc) which are described in data sources such as last.fm [9]. Information about venues and general locations is given in the source metadata, with variable granularity and consistency, using the *venue* and *coverage* tags:

venue The name of the venue, e.g. concert hall, club, festival. where the performance was recorded.

coverage The larger geographical area for the location, e.g. city or state.

The raw location information suffers from inconsistencies in presentation (e.g. *Chicago, IL*; *Chicago, Il*; *Chicago, Illinois*; *Chicago* etc.). Location informa-

[4]To the best of our knowledge, the collection does not contain examples of performances recorded by artists collaborating virtually in geographically distributed locations.

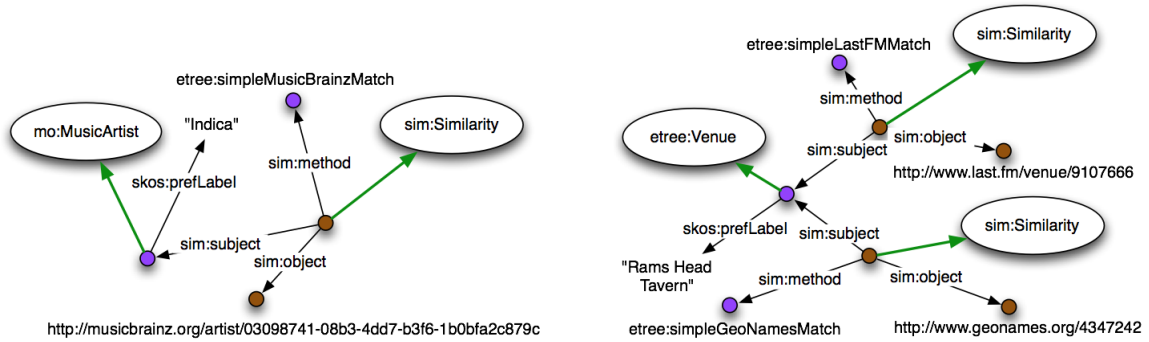

Fig. 2. Similarities with MusicBrainz (left) and Locations (right)

tion may in some cases also be ambiguous, with only city or town name being given (e.g. *Amsterdam* or *Springfield*). As discussed in Section 2, our approach in the collection is to expose the underlying source data and layer additional mappings on top. Thus each performance is associated with a *unique* venue with a name and location. A description that refers to the venue *Academy* in *Manchester* could refer to one of at least four distinct venues and, since there is insufficient information in the raw LMA data to reliably disambiguate, collapsing them is undesirable. Alignments between venues and external sources are thus again asserted using similarities.

Two external data sources provide additional information about venues and geographical locations which is of use here. Both sources provide latitude/longitude information.

GeoNames GeoNames provides identifiers for over eight million place names.

last.fm Last.fm provides a comprehensive list of music venues.

For a performance with a given venue and coverage, candidates for mappings are obtained through queries to the GeoNames and last.fm APIs. If potential candidates are returned from both collections, the geographical locations are cross-compared (both GeoNames and last.fm provide latitude/longitude information). Geographical co-location (up to a threshold of 10 miles) then gives us further confidence in the potential alignment.

Mapping candidates are associated with venues again using an explicit Similarity Ontology relationship. Figure 2 shows how such relationships are asserted.

5. Provenance

Alignments with external sources are represented as *similarities* using the vocabulary provided by the Similarity Ontology [7] (see Figure 2). This provides an object that represents the association and thus allows us to attach additional metadata to those objects asserting the provenance of the relationship. In the current dataset, this includes a link to a URI describing the method that was used to derive the alignment. We do not (as yet) provide explicit links to the *code* that was run in order to produce the alignments, but such an approach may be the topic of further work. Relationships from the W3C's PROV-O ontology [10] are used to assert additional information about the provenance of these mappings.

6. Process

The pipeline for initial data transformation was as follows:

1. Query Internet Archive for performances.
2. Crawl and download XML metadata files (using `wget`).
3. Process XML files using bespoke scripts.
4. Load resulting data into triple store.

This resulted in the core data collection. SPARQL queries against this collection were then used to extract field data for processing (e.g venue and location), with the resulting mapping/association triples added back into the triple store. Conversion to RDF was thereby in itself a useful step in completing the process.

7. Licensing & Access

The dataset is made available under the CC0 1.0 Universal[5] *Creative Commons public domain waiver*

[5] `http://creativecommons.org/publicdomain/zero/1.0/`

licence. Note that this license applies to the published metadata, not the source audio files served by the Internet Archive servers.

The dataset can be accessed via a SPARQL endpoint (see Table 1). Content-negotiated URIs (using a pubby[6] front end) are also available. The service is hosted by OeRC on a domain registered until 2021.

8. Statistics

Overview statistics are provided in Table 2. The collection currently contains information about 3,981 artists, of which 1,168 have mappings to identifiers in the MusicBrainz collection. There are 100,413 performances, with 1,631,604 individual tracks. Of the performances, 3,679 have venues mapped to last.fm venue identifiers, and 91,985 have locations identified with GeoNames. There are 416 distinct last.fm venues and 3,578 geolocations linked from the dataset. The dataset contains 17,555,791 triples.

Interlinking within the datasets is primarily via artists (performances by the same artist) and through the mapped geolocations and venues.

The expressivity of the language used is low. Most of the data here is factual, instance level data. The additional vocabulary specified in the etree ontology[3] provides subclasses of `mo:Performance`, defines a number of data properties, and names an inverse property.

9. Discussion

We believe the dataset as published is a useful resource, providing access to underlying audio files through a standardised query mechanism (SPARQL). As stated in the introduction, collection generation and management is a key starting point for research in the digital humanities. The publication process has also been useful in highlighting a number of issues in such an exercise, including the representation and presentation of alignments. The key value that the current linked dataset offers is the ability to link recordings of live performances to artist and geographical information. Thus we can potentially compare live performances by individual artists across different geographical locations. This could be in terms of metadata – does artist X play the same setlist every night? Such a query could also potentially be answered by similar resources such as setlist.fm[7]. The etree collection, however, also offers the possibility of combining metadata queries with computational analysis of the performance audio data – does artist X play the same songs at the same *tempo* every night, and does that change with geographical location?

An additional aspect here is that the dataset is an artefact which is worthy of further study – it is itself a part of the research process. The conversion and publication supports analysis of the dataset and its contents, with the layered approach as described in Section 2 being key to this.

This is a first step towards a rich dataset describing the resources in LMA and there are a number of additional enhancements that could potentially improve the dataset and enhance its utility.

Individual track matching. Alignment with MusicBrainz is currently at the level of artists. MusicBrainz (and other sources) also include track level metadata describing particular songs or pieces. Providing a mapping from the individual (track) performances in etree to MusicBrainz would then provide access to a corpus of *versions* of particular works. Representation of individual track matching requires the disambiguation between a musical work, a performance of that work and the audio encodings of that work – all of which can be represented in the Music Ontology. Such a matching process is likely to be challenging, however, in particular due to *(i)* the lack of standardisation in the description of track names in the etree source metadata; and *(ii)* the fact that songs played in live performance may not always be songs that feature in an artist's recorded canon.

Additional automated MusicBrainz artist matches. The current dataset uses a simple method to align artists to Musicbrainz. This has resulted in 29% of the artists in the dataset being mapped. A non-exhaustive examination of the data by eye suggests that one explanation for this is the number of "non main-stream" artists represented in the data set. It may be the case, however, that more sophisticated matching can provide further linkage, albeit with reduced confidence. Inclusion of manually curated mappings may also enhance linkage, albeit at an increased cost.

Crowdsourced corrections and mapping layers. Enabling a interface for community contribution to the alignment process. For example, allowing users to

[6] `http://www4.wiwiss.fu-berlin.de/pubby/`

[7] `http://www.setlist.fm/`

Dataset base URI	http://etree.linkedmusic.org
SPARQL endpoint	http://etree.linkedmusic.org/sparql

Table 1
Access details

Type	Total in Dataset	LMA Mapped Entities		Distinct External Entities	
Artists	3,981	MusicBrainz: 1,168 (29%)		MusicBrainz: 1,168	
Performances	100,413	—-			
Venues	100,413	last.fm: 3,679 (4%)	GeoNames: 91,985 (92%)	last.fm: 416	GeoNames: 3,578
Tracks	1,631,604	—-			
Triples	17,555,791	—-			

Table 2
Overview Statistics

identify and confirm the track-level mappings discussed above when they listen to or use the audio data.

Explicit characterisation of alignment processes. As discussed in Section 5, information is provided about the processes used to align entities with external sources. This simply takes the form of a label identifying a method. Further machine readable information describing the methods (and their execution) could also be provided.

Acknowledgements This work was undertaken during a visit to the Oxford e-Research Centre (OeRC) by Sean Bechhofer. He would like to thank the OeRC for hosting him during that time and the University of Manchester School of Computer Science for granting sabbatical leave. The authors would also like to thank the Internet Archive for granting permission to use the etree collection.

References

[1] Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. Describing Linked Datasets with the VoID Vocabulary. W3C Interest Group Note, World Wide Web Consortium, 2012. Available from: http://www.w3.org/TR/void/.

[2] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[3] Donald Byrd and Tim Crawford. Problems of music information retrieval in the real world. *Inf. Process. Manage.*, 38:249–272, March 2002. Available from: http://dl.acm.org/citation.cfm?id=637503.637509.

[4] D. De Roure, J.S. Downie, and I. Fujinaga. Salami: Structural analysis of large amounts of music information. In *UK e-Science All Hands Meeting*, 2010.

[5] J. Stephen Downie. The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine*, 12(12), December 2006. Available from: http://www.dlib.org/dlib/december06/downie/12downie.html.

[6] GeoNames [online]. Available from: http://www.geonames.org/ [cited 16th May, 2012].

[7] Kurt Jacobson, Yves Raimond, and Thomas Gängler. Similarity Ontology [online]. Available from: http://purl.org/ontology/similarity/ [cited 16th May, 2012].

[8] Kurt Jacobson, Yves Raimond, and Mark Sandler. An Ecosystem for Transparent Music Similarity in an Open World. In *International Conference on Music Information Retrieval*, 2009.

[9] last.fm [online]. Available from: http://www.last.fm/ [cited 16th May, 2012].

[10] Timothy Lebo, Satya Sahoo, and Deborah McGuinness. PROV-O: The PROV Ontology. W3C Working Draft, World Wide Web Consortium, 2012. Available from: http://www.w3.org/TR/prov-o/.

[11] Cynthia C.S. Liem and Alan Hanjalic. Cover Song Retrieval: A Comparative Study of System Component Choices. In *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, 2009.

[12] Internet Archive Live Music Archive [online]. Available from: http://archive.org/details/etree [cited 16th May, 2012].

[13] Alistair Miles and Sean Bechhofer. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, World Wide Web Consortium, 2009. Available from: http://www.w3.org/TR/skos-reference/.

[14] MusicBrainz [online]. Available from: http://musicbrainz.org/ [cited 16th May, 2012].

[15] K.R Page, B. Fields, B.J Nagel, G O'Neill, D. De Roure, and T. Crawford. Semantics for music analysis through linked data: How country is my country? In *IEEE Sixth International Conference on e-Science*, pages 41–48. IEEE, 2010.

[16] Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson. The music ontology. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2007.

[17] Yves Raimond and Samer Abdallah. The Event Ontology [online]. Available from: http://motools.sourceforge.net/event/event.html [cited 16th May, 2012].

[18] Yves Raimond and Frédérick Giasson. The Music Ontology [online]. Available from: http://musicontology.com/ [cited 16th May, 2012].

[19] Dave Simpson. Mark E Smith makes his entrance [online]. 2011. Available from: http://www.guardian.co.uk/music/2011/jun/14/mark-e-smith-entrance [cited 16th May, 2012].