HELLO CLEVELAND! LINKED DATA PUBLICATION OF LIVE MUSIC ARCHIVES

Sean Bechhofer[1]*, Kevin Page*[2]*, David De Roure*[2]

[1]School of Computer Science, The University of Manchester
[2]Oxford e-Research Centre, University of Oxford

ABSTRACT

We describe the publication of a linked data set exposing metadata from the Internet Archive Live Music Archive. The dataset contains over 17,000,000 triples describing 100,000 performances by 4,000 artists. Links to other existing musical and geographical resources facilitate query of the collection along a number of axes. We describe both the methods used to annotate and layer the metadata—with a focus on considering the patterns used to represent mappings—and the role that views constructed from such a Linked Data set can play to bring together multidisciplinary multimedia analysis techniques.

1. INTRODUCTION

Semantic Technologies offer the promise of standardised mechanisms for metadata management. The use of vocabularies and ontologies—shared collections of terms—ensures that applications use common terms. The provision of explicit machine-readable characterisations of those terms also helps ensure that interpretation of those annotation is consistent. *Linked Data* advocates the use of principles that essentially boil down to using HTTP URIs for resource identification, and ensuring that dereferencing those URIs provides useful information with links to other resources. This approach allows the use of existing web infrastructure to publish and consume metadata and facilitates the sharing and reuse of data across applications.

We describe an exercise in publishing metadata for a large music collection using Linked Data [2]. This metadata then allows for querying of the collection. Although query can already be done through existing services, we are able to benefit in a number of ways. Existing vocabularies/schema can be used to provide the conceptual models used in the annotations. For example, the Music Ontology provides schema information relating to artists and performances. Links can then be forged to external information sources (artist, venue and geographical information), enriching the metadata and enhancing query capability. This supports the extraction of subcollections based on particular axes of interest (e.g. performances of particular artists in geographical locations). One use of such collections is to generate ground-truth sets based on known criteria which can then be used to train computational analysis and classification tools. Such analysis can then be "fed back" into the collection metadata, providing further potential for enriched subset selection and consequent analysis. Note that we are dealing here purely with the *metadata*, and leave the audio source files untouched (but provide links to the online resources).

2. ETREE

The Internet Archive Live Music Archive[1] (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[2] allowing comparisons across performances. Furthermore, in live situations artists will frequently play works by other artists ("covers"), providing source content for cover detection algorithms[8].

Managing collections or subsets of input data and results, here using metadata, is a key process for applying and combining computational and humanities analyses. An earlier prototype [9] 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 re-publishing 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 activities such as the production of sub-corpora for experiments or evaluation.

3. DATA PUBLICATION

Linked Data [2] publishing follows a number of basic principles: using dereferenceable `http` URIs for identification of entities; returning useful information when those URIs are

[1] `http://archive.org/details/etree`

[2] 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.

dereferenced; and including links to other resources in that information. This common approach facilitates the construction of applications, and linked data publishing is gaining traction in a number of domains.

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[3], GeoNames[4] or last.fm[5] is then added. 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, due to interpretation and alignment, is not simply "asserted truth", but has some subjectivity.

Data consumers then have the option of using the encoded raw source data or the additional layer of mapped relationships.

4. CONTENT AND MODELLING

The collection contains a number of basic entities including *Artist*, *Performance* (and entire concert), *Track* (individual song or piece performance) and *Venue* (location of performance)

Each *Artist*, *Performance*, *Track* and *Venue* is minted a URI in the collection namespace[6] with an appropriate path prefix. A number of ontologies are used for the description of entities including the Music Ontology[7] [10], Event Ontology[8], and the Similarity Ontology[9] [6] which 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. In addition, the W3C ontology PROV-O[10] [7] provides vocabulary for describing provenance, the W3C dataset metadata ontology[11] [1] VoID is used to assert overall collection metadata, and a bespoke etree ontology[12] defines subclasses of Music Ontology classes and specific properties used in the etree metadata.

The basic modelling pattern used within the data set are shown in Figure 1. In the figures, green, unlabelled links are `rdf:type`. Blue, unlabelled links are `rdfs:subClassOf`.

[3] `http://musicbrainz.org`
[4] `http://www.geonames.org`
[5] `http://www.lastfm.org`
[6] `http://etree.linkedmusic.org`
[7] `http://musicontology.com/`
[8] `http://motools.sourceforge.net/event/event.html`
[9] `http://purl.org/ontology/similarity/`
[10] `http://www.w3.org/TR/prov-o/`
[11] `http://www.w3.org/TR/void/`
[12] `http://etree.linkedmusic.org/vocab`

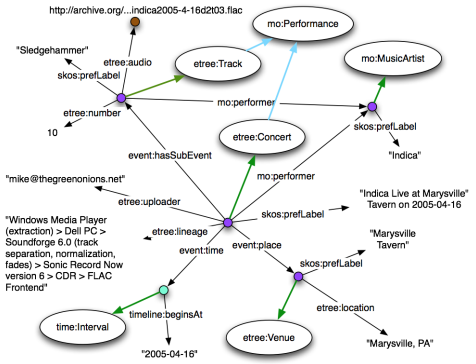

Fig. 1. Basic Model

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.

5. RECORD LINKAGE

A key aspect of a Linked Data publication approach is the provision of links or associations to external data sets. The LMA offers possibilities for record linkage with several 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[13] 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 3, the relationships between the artists and MusicBrainz are asserted using the Similarity Ontology.

Geographical Alignment Performances occur at a particular place[14] 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. Information about venues and general locations is given in the source metadata, with variable granularity and consistency, using the *venue* and *coverage* tags. Venue provides the name of the venue, e.g. concert hall, club, festival etc. where the performance was recorded and coverage describes 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 information may in some cases also be ambiguous, with only city or town name being

[13] `http://musicbrainz.org`
[14] To the best of our knowledge, the collection does not contain examples of performances recorded by artists collaborating virtually in geographically distributed locations.

given (e.g. *Amsterdam* or *Springfield*). As discussed in Section 3, 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.

Two external data sources provide additional information about venues and geographical locations which is of use here. GeoNames provides identifiers for over eight million place names, while last.fm provides a comprehensive list of music venues. Both sources provide lat/long information.

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 lat/long information). Geographical co-location (up to a threshold of 10 miles) then gives us further confidence in the potential alignment.

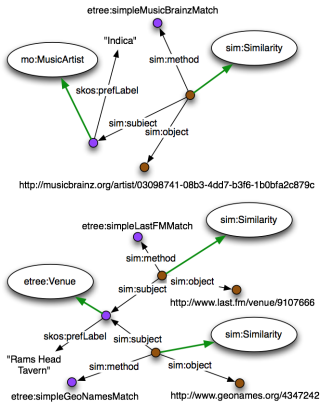

Fig. 2. Similarities

Alignments with external sources are represented as *similarities* using the vocabulary provided by the Similarity Ontology (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. Relationships from the W3C's PROV-O ontology [7] are used to assert additional information about the provenance of these mappings.

6. PROCESS

The pipeline for initial data transformation was as follows:

a) Query Internet Archive for performances.
b) Crawl and download XML metadata files.
c) Process XML files using bespoke scripts.
d) 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 the process of extracting and aligning further metadata: an approach we see as key to integrating multiple methods.

7. EXAMPLE BROWSING

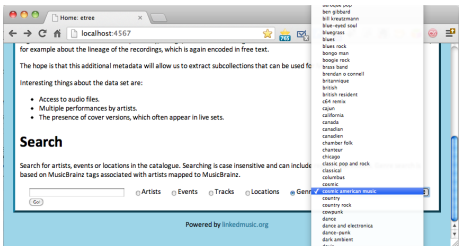

Fig. 3. Simple Browser Query using MusicBrainz keywords

As a demonstrator, a simple web application was developed that allows query and browsing of the collection via a collection of canned SPARQL queries. Figure 3 shows a query using genre keywords. The key thing to note here is that these keywords are not taken from LMA but are taken from MusicBrainz metadata—demonstrating the value of the linkage in terms of enhanced query.

The dataset[15] can be accessed via a SPARQL endpoint. Content-negotiated URIs (using a pubby[16] front end) are also available[17].

8. DISCUSSION

Figure 4 shows an overview of the approach. Original metadata describes the raw audio files held in the collection. Links to external resources provide additional information that can then be queried, enhancing access.

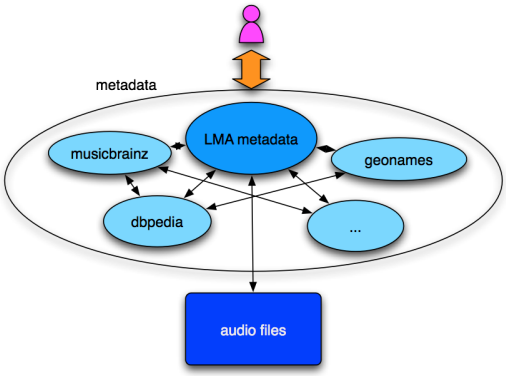

Fig. 4. Overview

We believe the dataset as published is a useful resource, providing access to underlying audio files through a standard-

[15]Data is published under the CC0 1.0 Universal licence: `http://creativecommons.org/publicdomain/zero/1.0/`. This license applies to the published metadata, not the source audio files served by the Internet Archive servers.

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

[17]`http://etree.linkedmusic.org`

ised query mechanism (SPARQL). 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. Linking of datasets using Semantic Web approaches is not new (e.g. see [11]), and we are not making claims of novelty here in terms of methods for linking—rather our novelty here is in the representation of those links and their uncertainty. 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[18]. The LMA 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? The profile of the collection—in particular the fact that multiple recordings of peformances are available—is again a potential point of interest.

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 3 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.

Improving alignments 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 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, due to *(i)* a 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. The current dataset uses simple string matching on names to align artists to Musicbrainz with 29% of the artists in the dataset being mapped. More sophisticated matching and the use of additional datasets (e.g. Discogs) would likely provide further linkage. Inclusion of manually curated mappings may also enhance linkage, albeit at an increased cost. We re-emphasis, however, that our focus is less on the *methods* of alignment and more on their *representation*.

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

Use in combination with computational analysis The linked data approach to metadata management (and in particular the layering used to separate raw data from annotations) will make it easier to combine the results of computational analysis in a unified framework.

Acknowledgements This work was undertaken during a visit to the Oxford e-Research Centre by Sean Bechhofer. He would like to thank the Centre 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 LMA.

[18] http://www.setlist.fm/

9. REFERENCES

[1] K.Alexander et al. Describing Linked Datasets with the VoID Vocabulary. W3C Interest Group Note, World Wide Web Consortium, 2012. http://www.w3.org/TR/void/.

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

[3] D. Byrd and T. Crawford. Problems of music information retrieval in the real world. *Inf. Process. Manage.*, 38:249–272, March 2002. 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. S. Downie. The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine*, 12(12), December 2006. http://www.dlib.org/dlib/december06/downie/12downie.html.

[6] K. Jacobson, Y. Raimond, and M. Sandler. An Ecosystem for Transparent Music Similarity in an Open World. In *ISMIR*, 2009.

[7] T. Lebo, S. Sahoo, and D. McGuinness. PROV-O: The PROV Ontology. W3C Recommendation, World Wide Web Consortium, 2013. http://www.w3.org/TR/prov-o/.

[8] C. C.S. Liem and A. Hanjalic. Cover Song Retrieval: A Comparative Study of System Component Choices. In *ISMIR*, 2009.

[9] K.R Page et al. 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.

[10] Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson. The music ontology. In *ISMIR*, 2007.

[11] Y. Raimond, C. Sutton, and M. Sandler. Automatic interlinking of music datasets on the semantic web. In *Linking Data on the Web (LDOW'08)*, 2008.