

# Linked Data Publication of Live Music Archives and Analyses

Sean Bechhofer<sup>1</sup>, Kevin R. Page<sup>2</sup>, David M. Weigl<sup>2</sup>, György Fazekas<sup>3</sup>, and Thomas Wilmering<sup>3</sup>

<sup>1</sup> School of Computer Science, University of Manchester  
sean.bechhofer@mancheseter.ac.uk

<sup>2</sup> Oxford e-Research Centre, University of Oxford  
{kevin.page, david.weigl}@oerc.ox.ac.uk

<sup>3</sup> Centre for Digital Music, Queen Mary University of London,  
{g.fazekas, t.wilmering}@qmul.ac.uk

**Abstract.** We describe the publication of a linked data set exposing metadata from the Internet Archive Live Music Archive along with detailed feature analysis data of the audio files contained in the archive. The collection is linked to existing musical and geographical resources allowing for the extraction of useful or interesting subsets of data using additional metadata.

The collection is published using a 'layered' approach, aggregating the original information with links and specialised analyses, and forms a valuable resource for those investigating or developing audio analysis tools and workflows.

## 1 Introduction and Context

The Internet Archive Live Music Archive<sup>4</sup> (further referred to here as LMA) is an online resource providing access to a large community-contributed collection of live recordings. Covering nearly 5,000 artists, chiefly in rock genres, the archive contains over 130,000 live recordings made openly available with the permission of the artists concerned. Audio files are available in a variety of formats (and with varying levels of quality), 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 as exemplified by the MIREX series of workshops [7]. It provides multiple recordings by individual artists<sup>5</sup> allowing comparisons across performances. It provides multiple recordings of single events, allowing for enhanced user experience through combinations of recordings [13]. Furthermore, in live situations artists will frequently play works by other artists ("covers"), providing source content for cover detection algorithms. The collection is not without challenges,

---

<sup>4</sup><http://archive.org/details/etree>

<sup>5</sup>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.

Layer	Details
Feature Metadata	Feature summary/provenance LD Service
Computational Analysis	Sonic Annotator feature analysis
Collection Metadata	etree LD Service
Collection	Internet Archive Live Music Archive

**Fig. 1.** Layers in the etree/CALMA dataset

however. Recordings in the LMA range in source from handheld tape recorders, through smart phones in the audience, to a feed from the mixing deck. A poorly tuned instrument or late entry constitute a ‘truth’ in live performances that would more likely trigger a re-take in the studio. The signals themselves can be noisy, with crowd chatter, on-stage banter and improvisation.

Semantic Web technologies have been previously applied in the context of digital music collections [8, 1, 6] and successfully applied to other projects under the auspices of *Transforming Musicology* [5, 11].

In CALMA<sup>6</sup> we have built a layered Music Digital Library using Semantic Web technologies to combine and interpret metadata and content-based analyses (see Figure 1). The data set builds on the source audio and (largely free-text) metadata, introducing consistent structure and links to external sources.

The original, community contributed, metadata has been converted to RDF and published as linked data [2] (the *etree* dataset). The data set has been enhanced with connections and links to a number of external data sources providing additional information about the entities in the data – sources such as MusicBrainz<sup>7</sup>, GeoNames<sup>8</sup> and last.fm<sup>9</sup>. In addition to the contributed metadata, audio analysis is being performed on the underlying audio files in the collection, resulting in a corpus of feature data, analysis results and provenance. This data has again been published [12] (the *CALMA* dataset), exposing the computational results as “blobs” with accompanying feature and provenance data in RDF. The original data sources are published “as is”, with collection metadata, computational analysis results and feature metadata layered on top of this substrate. This allows a clear separation between the source and enhancements, with corresponding provenance information. This is particularly important in a context such as musicology, where findings may be contingent or speculative. The results of analyses performed using different tools and algorithms may have varying results or quality given the context (e.g. noise in the recording), therefore the provenance data we publish can be crucial in establishing trust.

## 2 The Collection

The LMA is largely focused on recordings of live performances, concerts or events. Each of these will include a number of songs or tracks, with each song associated with

<sup>6</sup>Computational Analysis of the Live Music Archive

<sup>7</sup><http://musicbrainz.org>

<sup>8</sup><http://www.geonames.org>

<sup>9</sup><http://www.lastfm.org>

an uploaded audio file. These files are in a variety of formats, and have been produced using a variety of techniques, from hand held microphones in the crowd, through to high quality digital feeds taken directly from a mixing desk. Audio files have often been post-processed by the uploader before addition. As a result, the audio files vary not just in format, but in quality and may also contain artefacts due to the recording or processing techniques used. For example, many recordings contain crowd noise or on-stage banter. Metadata contributed by the uploader describes information about the event such as the location and date along with lineage or source information that describes, for example, the signal processing chain used for post-processing. Note that this information is not structured or controlled.

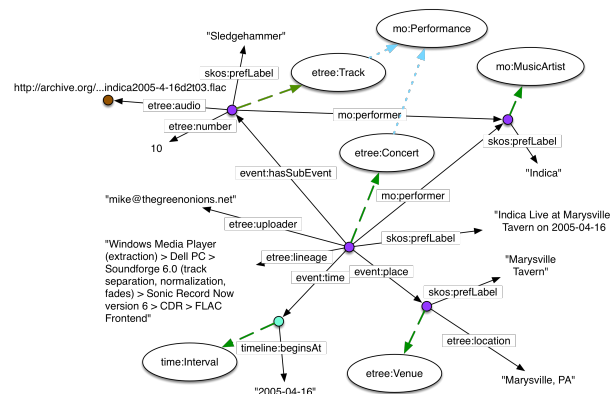
The LMA is typical of many collections in its metadata being the most comprehensive means for indexing and accessing what is clearly a valuable cultural resource. This metadata is, however, gathered using free text fields entered by the audio uploader, so is potentially prone to errors (e.g. in set list order) or typographic mistakes (e.g. misspelling of artists, track titles, or venues). Given one potential value of the LMA for study lies in comparing recordings of the same track, artist, or venue, correcting – or accepting – metadata imperfections within analyses must be addressed for scholarly adoption.

The initial motivation for this work was to enable investigations such as the following examples: 1/ Identify the same song performed by the same artist, but at multiple venues over multiple dates, analysing the audio for tempo. If an artist performs with a faster tempo at a venue, do other artists do the same at that venue? Is there a correlation between tempo differences and performance date, line up, weather etc.? 2/ Finding performances by artists in their home towns, does audience reaction (between songs) differ from other venues? 3/ Metadata from the LMA can be incomplete, with missing or erroneous labels. Can we cross-validate with audio analyses?

We note that investigations into the first two questions require additional information which is held outside of the LMA, for example geo-location data, membership or history of bands or meteorological data. The use of a Linked Data approach (See Section 3) offers promise in providing this additional contextual information.

We also recognise that an investigation of any one of these topics alone does not *require* this data publication; it is also plausible that software focussed on a single investigation might be achieved at lower cost in time and code. We argue that investment in our approach is returned when layers can be re-used, extended, and adapted; when one can re-use and extend the layers of others; and in their transparency for peer-review and validation.

The etree and CALMA datasets have differing, but complementary purposes. Considering them together illustrates benefits descended from our use of Linked Data: the consistent application of two distinct dataset motivations within each dataset (one bibliographic; the other audio analytic); the ability to build a new second dataset (CALMA) upon the foundation of the first (etree), where the former postdates the latter by several years and was achieved through the addition of new institutions and expertise into the collaboration; and the easy retrofitting of CALMA links back into etree/LMA referencing once the analyses were completed.



**Fig. 2.** Basic Data Model

### 3 Modelling, Ontologies and Vocabularies

Lynch [9] proposes that digital collections should be exposed as databases of raw cultural heritage materials along with layers of interpretation and presentation built upon these databases and making reference to the objects within them. This is the approach taken here. As discussed in [12], the original metadata is preserved and made available. Our publication process introduces a set of uniform identifiers for the entities represented in the data (artists, events, songs, venues etc) along with links, both within the dataset and to resources outside.

The basic modelling pattern used in the dataset is shown in Figure 2. In the figures, green, dashed, unlabelled links are `rdf:type`. Blue, dotted, 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. Where possible, existing ontologies and vocabularies have been used in the descriptions of entities. These include:

**Music** The Music Ontology or MO<sup>10</sup> provides terms that describe performances, artists and the relationships between them.

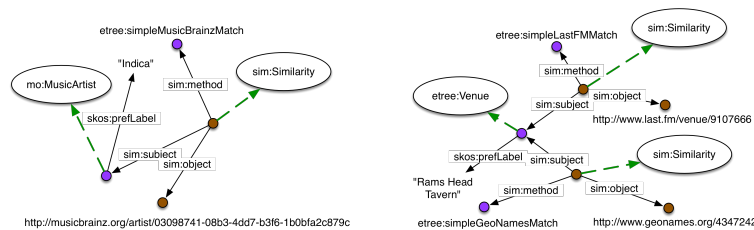
**Events** The Event Ontology or EO<sup>11</sup> provides terms for describing events.

**Similarity** The Similarity Ontology<sup>12</sup> provides terms for asserting associations between entities. This is used to associate entities in the collection such as artists or locations with external entities from e.g. MusicBrainz and GeoNames. A key design decision is to provide explicit resources modelling similarities in order to allow for the recording of provenance information on these similarities. The Similarity Ontology was chosen due to its existing usage within the computational musicology community.

<sup>10</sup><http://musicontology.com/>

<sup>11</sup><http://motools.sourceforge.net/event/event.html>

<sup>12</sup><http://purl.org/ontology/similarity/>



**Fig. 3.** Similarities with MusicBrainz (left) and Locations (right)

**SKOS** SKOS<sup>13</sup> labelling properties are used to label entities.

**PROV-O** The W3C provenance ontology<sup>14</sup>.

**VOID** The W3C dataset metadata ontology<sup>15</sup>.

**VAMP** A vocabulary describing VAMP<sup>16</sup> audio analysis plugins<sup>17</sup>.

In addition to the vocabularies listed above, bespoke ontologies<sup>18,19</sup> define subclasses of Music Ontology classes and specific properties used in the etree metadata.

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<sup>20</sup> is an “open music encyclopaedia” providing identifiers for a large number of music artists and is a clear candidate for linking from a collection like LMA. Alignments between etree artists and MusicBrainz use a combination of string matching on artist names and song titles. In keeping with the strategy outlined above, the relationships between the artists and MusicBrainz are asserted using the Similarity Ontology as shown in Figure 3.

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`). Artists are aligned in a similar way with last.fm<sup>21</sup>.

<sup>13</sup><http://www.w3.org/TR/skos-reference>

<sup>14</sup><http://www.w3.org/TR/prov-o/>

<sup>15</sup><http://www.w3.org/TR/void/>

<sup>16</sup><http://vamp-plugins.org/>

<sup>17</sup><http://purl.org/ontology/vamp/>

<sup>18</sup><http://etree.linkedmusic.org/vocab>

<sup>19</sup><http://calma.linkedmusic.org/vocab>

<sup>20</sup><http://musicbrainz.org>

<sup>21</sup><http://last.fm>

*Geographical Alignment* Performances occur at a particular place<sup>22</sup> 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, where *venue* describes the name of the venue where the performance was recorded e.g. *The 9:30 Club* and *coverage* gives the larger geographical area for the location, e.g. *Washington, DC*.

The raw location information suffers from inconsistencies in presentation (*Chicago, IL*; *Chicago, Il*; *Chicago, Illinois*; *Chicago* etc.). Location information may in some cases also be ambiguous, with only city or town name being given (e.g. *Amsterdam* or *Springfield*). As discussed above, 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 entity 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 (including latitude/longitude) 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. 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. Note that these geographical alignments are *not* necessarily asserting that the entities are the same. For example, a venue entity in LMA may be associated with a GeoNames district. As yet, no formal evaluation of the quality of the alignments has been done.

*Provenance* The use of the Similarity Ontology (see Figure 3) provides objects that represent associations between objects and thus allow 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 are used to assert additional information about the provenance of these mappings. PROV-O vocabulary is also used to record provenance of audio feature extraction, including information about the VAMP plugins used, parameters, etc.

The basic metadata collection (excluding audio feature metadata) contains over 12 million RDF triples concerning over 135,000 distinct performances and nearly 5,000

---

<sup>22</sup>To the best of our knowledge, the collection does not contain examples of performances recorded by artists collaborating virtually in geographically distributed locations.

artists with at least one performance. The analysis feature data currently covers 300,000 individual tracks and comprises approx 1.1TB of (compressed) data.

## 4 Usage & Access

Audio feature extraction (and analysis) is resource-intensive. The original harmonised metadata proved invaluable in supporting the selection of “interesting” events to initially target. Our initial question was to consider how performances of songs might possibly change over time, thus we focused on artists with multiple events where particular songs or pieces have multiple performances. A workset was established using a SPARQL query against the catalogue metadata layer, selecting those artists in etree with more than 200 but less than 1,000 performances, constrained to those who performed at least one song title in more than 100 distinct performance recordings<sup>23</sup>.

Although this *would* have been possibly via queries over the original metadata, the SPARQL endpoint made it simple. As discussed in [12], the workflow for analysis involves retrieving audio from the Internet Archive<sup>24</sup> and running Python scripts built around the Sonic Annotator platform for audio feature extraction [4]. The computational results are themselves stored as an RDF blob (although this may not always be an appropriate format hence our treatment of the results as a blob) with the feature metadata and provenance information regarding the execution of the workflow added to our data set. This metadata again uses existing published vocabularies as described earlier (MO and PROV-O).

An investigation into “typicality” summarising feature value distributions within multi-performance, same-song collections is reported in [12]. Metadata from LMA and CALMA feature analysis (accessed via the data set) has also been used to develop an immersive experience through alignment and clustering of recordings [13]. Discussions are ongoing with the Internet Archive, with the hope that information may be re-ingested into the IA collections. We also believe that the metadata could support musicological analyses and new music discovery tools (e.g. MusicWeb [10]).

The collection can be accessed via two persistent URLs <http://purl.org/etree> and <http://purl.org/calma>. The PURLs resolve to resources that provide VoID metadata about the respective collections. The LMA metadata is accessible via a SPARQL endpoint and a browsable (pubby) front end is also provided. Currently, the CALMA metadata and feature extraction results are available via direct download. Metadata and analysis results are made available under Creative Commons CC0.

The data in LMA is largely static, in that the audio and metadata are not usually edited once they have been deposited. Annotations may be added – for example alignments to other data sources – but as these use the layered approach described here, this is additive. Thus, to date, we have not been troubled with issues relating to versioning. Update of our translated resource is not, as yet, automated – the LMA is updated with additional recordings daily. The workflow supporting inclusion of additional performances into the collection is clear, however, and for the purposes of many investigations, historical recordings are sufficient.

<sup>23</sup>The SPARQL query can be found on the dataset descriptive web pages.

<sup>24</sup>Our collection contains only the metadata, not the audio files.

## 5 Conclusions and future work

We have presented a layered digital library providing multimedia access to audio, user-provided metadata, and audio-derived feature metadata of the Live Music Archive, in turn allowing novel exploratory analyses across and within its layers. The dataset provides access to a large open data collection in the Digital Humanities, supporting musicological scholarship at scale, and representing an augmentation and enrichment of a valuable public resource for fans and listeners. We also envisage applications that go beyond musicological analyses and support further services that are built on such “enhanced archives”.

*Acknowledgments* This work was supported through a subaward of the Semantic Media Network (EPSRC EP/J010375/1); the FAST IMPACT project (EPSRC EP/L019981/1); and the EU Commission H2020 grant AudioCommons (688382). We thank the Internet Archive and Brewster Kahle for encouragement, support and access to the dataset.

## References

1. D. Bainbridge, X. Hu, and J. S. Downie. A Musical Progression with Greenstone: How Music Content Analysis and Linked Data is Helping Redefine the Boundaries to a Music Digital Library. In *Proc. 1st International Workshop on Digital Libraries for Musicology*, pages 1–8. ACM, 2014.
2. S. Bechhofer, K. Page, and D. De Roure. Hello Cleveland! Linked Data Publication of Live Music Archives. In *14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services*. IEEE, 2013.
3. D. Byrd and T. Crawford. Problems of music information retrieval in the real world. *Inf. Process. Manage.*, 38:249–272, March 2002.
4. C. Cannam, M. Sandler, M.O. Jewell, C. Rhodes, and M. d’Inverno. Linked Data and You: Bringing Music Research Software into the Semantic Web. *Journal of New Music Research*, 39(4), 2010.
5. T. Crawford, B. Fields, D. Lewis, and K. Page. Explorations in Linked Data practice for early music corpora. In *Digital Libraries (JCDL), 2014*, pages 309–312. IEEE, 2014.
6. D. De Roure, G. Klyne, K. R. Page, J.P.N. Pybus, and D.M. Weigl. Music and Science: Parallels in Production. In *Proc. 2nd International Workshop on Digital Libraries for Musicology*, pages 17–20. ACM, 2015.
7. J. S. Downie. The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine*, 12(12), December 2006.
8. G. Fazekas, Y. Raimond, K. Jacobson, and M. Sandler. An Overview of Semantic Web Activities in the OMRAS2 Project. *Journal of New Music Research*, 39(4):295–311, 2010.
9. C. Lynch. Digital Collections, Digital Libraries and the Digitization of Cultural Heritage Information. *First Monday*, 7(5), 2002.
10. M. Mora-Mcginity, A. Allik, G. Fazekas, and M. Sandler. Musicweb: Music discovery with open linked semantic metadata. In *Metadata and Semantics Research: MTSR 2016*, 2016.
11. T. Nurmikko-Fuller, A. Dix, D. M. Weigl, and K. R. Page. In Collaboration with In Concert: Reflecting a Digital Library as Linked Data for Performance Ephemera. In *Proc. 3rd International workshop on Digital Libraries for Musicology*, pages 17–24. ACM, 2016.
12. K. R. Page, S. Bechhofer, G. Fazekas, D.M. Weigl, and T. Wilmering. Realising a Layered Digital Library: Exploration and Analysis of the Live Music Archive through Linked Data. In *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 2017.
13. T. Wilmering, F. Thalmann, and M. B. Sandler. Grateful Live: Mixing Multiple Recordings of a Dead Performance into an Immersive Experience. In *AES Convention 141*, 2016.