Realising a Layered Digital Library

Document Version
Accepted author manuscript

Citation for published version (APA):
Page, K., Bechhofer, S., Fazekas, G., Weigl, D., & Wilmering, T. (2017). *Realising a Layered Digital Library: Exploration and Analysis of the Live Music Archive through Linked Data*. Paper presented at ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) 2017, Toronto, Canada.

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

OPEN ACCESS

Realising a Layered Digital Library

Exploration and Analysis of the Live Music Archive through Linked Data

Kevin R Page
Oxford e-Research Centre
University of Oxford
kevin.page@oerc.ox.ac.uk

Sean Bechhofer
School of Computer Science
University of Manchester
sean.bechhofer@manchester.ac.uk

György Fazekas
Centre for Digital Music
Queen Mary University of London
g.fazekas@qmul.ac.uk

David M Weigl
Oxford e-Research Centre
University of Oxford
david.weigl@oerc.ox.ac.uk

Thomas Wilmering
Centre for Digital Music
Queen Mary University of London
t.wilmering@qmul.ac.uk

ABSTRACT

Building upon a collection with functionality for discovery and analysis has been described by Lynch as a 'layered' approach to digital libraries. Meanwhile, as digital corpora have grown in size, their analysis is necessarily supplemented by automated application of computational methods, which can create layers of information as intricate and complex as those within the content itself. This combination of layers – aggregating homogeneous collections, specialised analyses, and new observations – requires a flexible approach to systems implementation which enables pathways through the layers via common points of understanding, while simultaneously accommodating the emergence of previously unforeseen layers.

In this paper we follow a Linked Data approach to build a layered digital library based on content from the Internet Archive Live Music Archive. Starting from the recorded audio and basic information in the Archive, we first deploy a layer of catalogue metadata which allows an initial – if imperfect – consolidation of performer, song, and venue information. A processing layer extracts audio features from the original recordings, workflow provenance, and summary feature metadata. A further analysis layer provides tools for the user to combine audio and feature data, discovered and reconciled using interlinked catalogue and feature metadata from layers below.

Finally, we demonstrate the feasibility of the system through an investigation of 'key typicality' across performances. This highlights the need to incorporate robustness to inevitable 'imperfections' when undertaking scholarship within the digital library, be that from mislabelling, poor quality audio, or intrinsic limitations of computational methods. We do so not with the assumption that a 'perfect' version can be reached; but that a key benefit of a layered approach is to allow accurate representations of information to be discovered, combined, and investigated for informed interpretation.

CCS CONCEPTS

•**Applied computing** →**Digital libraries and archives;** *Arts and humanities;*

KEYWORDS

Linked Data, computational audio analysis, Music Digital Libraries, implementation, scholarly investigation using Digital Libraries.

1 INTRODUCTION

1.1 'Layered' Digital Libraries

In his far-sighted 2002 paper, Lynch [25] made numerous observations about the distinction between digital collections and libraries, and predictions regarding the academic study, interpretation and analysis of the contents therein, particularly in the context of what might now be viewed as the Digital Humanities. He posited that increasing quantities of digital (and digitised) materials, and the mechanisms we might use to investigate them, would lead to

> a world of digital collections – databases of relatively raw cultural heritage materials, for example – and then layers of interpretation and presentation built upon these databases and making reference to objects within them.

Lynch proposes that the purpose of digital library systems is to provide the layers which 'make digital collections come alive, make them usefully accessible, that make them useful for accomplishing work, and that connect them with communities'. Noting that 'there are going to be layers of mark-up' which might be provided by multiple different actors, with differing authority, 'we may need to be thinking about representations for things like contingent or speculative mark-up, mark-up with confidence levels and provenance'. Expanding upon the delightful notion of books 'talking to each other', he suggests

> one of the things they 'say' is what we code into them with mark-up. Really good deep mark-up that exposes intellectual and semantic structure, that exposes content for linkage and data mining, and computation

recognising the corresponding need for 'persistent identifier systems, which seem to me to be an absolute cornerstone of designing digital collections that are overlayable, reusable and repurposable'.

On computational processing, Lynch is minded it is 'useful to run these systems against digitized materials and put in preliminary or unevaluated tagging from automated analysis systems', at least 'until the humans come, until some intellectual analysis can be

done'. He highlights the importance of identifying results where different analytic approaches agree (and, by implication, disagree).

The vision laid out by Lynch in 2002 seems remarkably prescient in providing a framing which is equally applicable to contemporary challenges of discovery and analysis in large-scale collections; and yet few digital library systems have been developed with such a wide-ranging and distributed approach to layering in mind. In this paper we report our attempt to do so, not as a universal template for all potential layered systems, but as a realisation of layering atop a specific collection, using pragmatically chosen technologies, to show – as Lynch puts it – 'that the aggregation of materials in a digital library can be greater than the sum of its parts'.

1.2 The Internet Archive Live Music Archive

The foundational collection for our layered digital library is the Live Music Archive (LMA),[1] part of the Internet Archive, an online resource providing access to a large community contributed collection of live recordings. Covering over 5,000 artists, chiefly in rock genres, the archive contains a growing collection of over 150,000 live recordings of concerts made openly available with the permission of the artists concerned. Audio files are provided in a variety of formats, and each recording is accompanied by basic metadata describing information including dates, venues, set lists and the provenance of the audio files. The audio recordings are contributed from many sources, including artists, sound engineers, or members of the audience – sometimes there are multiple recordings of the same concert – with the expected variations in sound fidelity and performance quality (compared to the 'perfect' mix from a studio).

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[2]. While unvalidated crowd-sourced data entry is perhaps towards an extreme of a lacking metadata curation, LMA is not atypical in relying on imperfect or subjective metadata for the crucial process of discovery.

1.3 Layering the Live Music Archive

In the remainder of this paper we describe a digital library layered over the LMA called CALMA – Computational Analysis of the Live Music Archive. The raw LMA collection contains a rich multimedia mix of audio and performance ephemera (set lists etc.). In using layers, as per Lynch, we seek to enable investigation of the Archive content through the creation of new layers of findings, improvement of information within existing layers, and analysis across layers; with the latter contributing to the former two. We propose this as an effective means to perform scholarship over a

collection of this scale, iteratively and incrementally adding to our collective knowledge and understanding of it.

Lynch talks of information moving between databases [25]; in CALMA each layer takes the form of a Web service, or a tool interacting with a Web service; and rather than 'move' information, connections between the layers are achieved using Linked Data. While CALMA is described here with a clear boundary to facilitate clarity in reporting, the flexibility to add or adapt layers is integral to our use of Linked Data, and CALMA as described should *not* be taken as the ultimate or comprehensive system – we argue such an end is impossible, or at least undesirable. While layers in CALMA, as reported here, are hosted on common infrastructure, there is no technical reason they could not be distributed across the WWW. Indeed the layers were developed at different institutions, by different authors of this paper, incrementally over a number of years; the layers are compatible through an intersection of common vocabularies and data which are no more organisationally taxing than usual co-operation between academics sharing a common interest.

The *retrieval* of music is often realised as an optimisation towards some notion of 'correctness', or perhaps even perfection – at least from an informational or computational perspective. Yet as social artefacts, many music corpora can be considered inherently imperfect. This raises the question of how far, and by what means, we can undertake structured investigations of these valuable and sometimes large-scale collections while respecting the underlying, if messy, 'truth' of the dataset. So while there are challenges in the application of computational algorithms, as we later discuss, it is also clear that exhaustive close listening is impossible over a collection of this size. We must therefore mix methods in any in-depth investigation, recording accurate provenance descriptions of different approaches applied so these judgements can be correctly interpreted. This will necessarily add complexity, so a further challenge is to develop tooling and data structures to capture this information while allowing it to remain useful.

In conceiving CALMA, we were motivated to enable investigations such as the following examples:

– 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.?

– Finding performances by artists in their home towns, does audience reaction (between songs) differ from other venues?

– Metadata from the LMA can be incomplete, with missing or erroneous labels. Can we cross-validate with audio analyses?

We recognise an investigation of any one of these topics alone could be conducted *without* cause for a layered digital library; it is also plausible that software focussed on a single investigation might be achieved at lower cost in time and code. We argue the 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.

Figure 1 provides an overview of layers in the CALMA digital library, and the software and services implementing the layers (detailed in section 3): a *collection metadata* layer rationalising access and discovery in the collection; a *computational analysis* layer which extracts audio features from the collection sound content; and

---

[1]https://archive.org/details/etree
[2]See Angeles *et al.*[2] for an investigation of similar inconsistencies in the Codaich music research library.

Digital Library layer	CALMA services
Exploratory analysis	RCalma data analysis tool
Feature metadata	Feature summary & provenance Linked Data service
Computational analysis	Sonic Annotator audio feature extraction workflow
Collection metadata	etree Linked Data service
Collection	Internet Archive Live Music Archive

Figure 1: Digital library layers in CALMA

a *feature metadata* layer that enables discovery of feature-derived information and the corresponding collection items.

The final *exploratory analysis* layer provides tools for a researcher to combine all of the preceeding layers, iteratively and in any order, in pursuit of an investigation. In section 4 we demonstrate the feasibility of this particular layer and the wider CALMA digital library through two studies: the first to gain an understanding of 'typicality' of the musical keys detected by the computational analysis workflow; and the second to investigate the possible application of key typicality to cross-validation of LMA metadata. Once more, our intention is not to report these investigations as the full extent of our apparatus, but as illustrative and informative examples of the viability, flexibility and potential of a layered digital library.

2 RELATED WORK

In realising the CALMA layered digital library we build upon a broad range of earlier research, and draw from several disciplines, as might be expected in a system where each layer supports a distinct, if complementary, purpose. We use Linked Data both as a means for making connections between layers, and for structuring information within layers, particularly for collection metadata and feature metadata; approaches from Music Information Retrieval (MIR) within the computational analysis layer; and develop ideas from MIR and Music Digital Libraries to work in a layered, iterative, research environment for our exploratory analysis.

2.1 Linked Data for Libraries

The Semantic Web generally, and Linked Data [8] specifically, extend the linking structure of the World Wide Web by employing URIs to specify directed relationships between data instances, which may themselves be encoded by URIs or represented by literal values. A set of two such instances, linked by such a relationship, is referred to as a *triple*. Collections of triples may be stored in flat files expressed in a Resource Description Framework (*RDF*) format on a server, and accessed via HTTP; or within specialised RDF databases, known as *triplestores*. From the latter, data can be flexibly accessed using *SPARQL*, an RDF query language analogous to SQL, that enables retrieval of data by specifying patterns of interlinked triples. In employing a Linked Data approach, meanings of the relationships between the data are made explicit, allowing them to be understood by both humans and machine.

The use of Linked Data to supplement—or even replace—catalogue records with bibliographic ontologies remains a topic of active research and ongoing discussion both in libraries [23] and in the digital humanities [30]. Available ontologies include BIBFRAME,[3]

a conceptual bibliographic description model; RDF ontologies expressing the Metadata Object and Metadata Authority Description Standards (MODS/RDF[4] and MADS/RDF[5]); as well as the FRBR-aligned Bibliographic Ontology (FaBiO)[6], among others [22].

The benefits of Linked Data in the library context include metadata openness and sharing; facilitating serendipitous discovery of information; accommodating emergent metadata through dynamic processes of iterative enrichment; enabling the use of arbitrary dimensions and properties as navigation facets; and supporting the seamless linking to and from external data sources [1]. Worksets [20], specialized scholarly collections, have been proposed to support large-scale computational analyses over items in Digital Library collections in a way that capitalises on these benefits.

Existing approaches toward the application of Linked Data to Digital Libraries generally involve RDF migrations of legacy collections [17, 19, 29]. This approach identifies each entity and each relation in the dataset with an authoritative URI, enabling the precise, machine-readable specification of semantics, and supporting referencing from external data sources. However, existing topologies present in the legacy collection are persisted into the resulting RDF representation. As these topologies did not originate as Linked Data, they tend not to be distributed (rather, they are described as a set of triples generally residing within a triplestore hosted on a particular server); neither are they layered as per Lynch [25], nor explicitly designed for layering by others.

2.2 Music Information Retrieval

The field of Music Information Retrieval (MIR) has produced algorithms performing a large variety of audio analysis and musically-relevant feature extraction tasks. Such content-based MIR techniques could clearly offer algorithms for analysing audio from the LMA, but here the Archive constitutes a further challenge in both number and quality of the audio recordings within. While previous endeavours have undertaken computational analysis on a large-scale and created systems to do so [4, 7, 33, 36, 37] here we are less focussed on the – necessary – process of big data feature extraction, and more on the accompanying data structures and tools that might enable us to combine, cross-reference, and interpret the results of these analyses dynamically in the supplementary layers of the Digital Library. This investigative element is particularly relevant when considering the variable quality of recordings – and indeed performances – in the LMA, and the effect this can have on feature extractors (e.g. [44]). 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.

Analysis to identify patterns and alignments between audio recordings bears a resemblance to MIR tasks including audio fingerprinting [40] and cover song detection [16], which also seek to perform song classification using features extracted from audio content. Our situation, although apparently similar, differs profoundly

---

from these tasks for three reasons. First, while audio fingerprinters can filter noise from a signal, the nature of live audio means that the signal itself is noisy – due to crowd chatter, stage banter, improvisation, etc. Second, both fingerprinters and cover song detectors attempt to match input audio to a canonical target, often the high quality studio album recording. In our case, there is no such canonical exemplar; instead, the best we can do is to match to a representation of a 'typical' live performance, informed by the entire collection of renditions available in the corpus. Finally, it is worth noting that we are pursuing different goals; while fingerprinters and cover song detectors seek a specific result or definitive answer (typical for most MIR tasks which tend to be oriented around some notion of 'ground truth'), we instead seek to use features extracted from the musical content as informative measures within a wider analytical context, supporting a cyclical process of exploration, discovery, analysis, refinement, and metadata enrichment.

2.3 Music Digital Libraries

The use of Semantic Web technologies in the context of digital music collections has been previously implemented [3, 12, 13] and successfully applied to other projects under the auspices of *Transforming Musicology* [10, 29]. Projects such as *SALAMI: Structural Analysis of Large Amounts of Music Information* [5] and *RISM:Répertoire International des Sources Musicales*[7] are illustrative of recent projects with similar research agendas, whilst the *Répertoire International de Littérature Musicale*[8] exemplifies on-going work in the field of ontology design for musicological data.

Musical Digital Libraries containing musical encodings (be they audio or symbolic) afford means of access beyond traditional textual query interfaces. Descriptive feature data derived from the musical content using MIR techniques (section 2.2) offer non-textual interaction paradigms including acoustic inputting of queries, e.g. by singing or humming into a microphone [28]. Further, such content-derived features, as well as higher-level feature aggregations (e.g., mood classifications), may be accessed directly, both to inform on-going feature extractor development [15], and to provide answers to musicological queries [3].

For the most part, Music Digital Libraries have been collection-specific, primarily providing an efficient means of retrieval. Computationally-derived feature data has been used in access mechanisms for such collections. Less attention has been paid to iterative processes of musicological scholarship, driven by the consumption, enrichment, and reuse of layers characterised by Lynch [25].

In CALMA we are building a layered Music Digital Library using Semantic Web technologies to combine and interpret metadata and content-based analyses. This can be considered a 'Third Generation' e-Science approach [11] within a Digital Humanities context, characterised by the reuse of tools, data, and methods in support of increasingly data driven multidisciplinary scholarship. Applications and resources in the Music Information domain exhibiting complex interlinking of rich multi-modal content with social, bibliographic, and contextual metadata have previously informed thought in e-Science [14, 18]; and in earlier work we have demonstrated the potential of Linked Data for MIR [32]. Building upon these insights,

in CALMA we expand the range of both real-world data and real-world investigation: through the consistent use of globally unique, persistent identifiers (in the form of URIs), a common underlying model in RDF, shared ontologies for information exchange, and the alignment of distributed information we create a web of Linked Data of practical use for scholarship.

3 A LINKED DATA REALISATION OF A LAYERED DIGITAL LIBRARY

In this section we describe implementation of the layers forming the CALMA digital library (figure 1) which supplement the Live Music Archive collection (section 1.2). In general we can say that each provides a process that consumes information from other layers, and creates new information for its own (as later illustrated in figure 3). While graphs of Linked Data may exist within layers, and may indeed be more densely interconnected within, we are also creating links between layers – a graph of graphs (or web of webs). Since the layers are most effective in combination, an overarching concern are the requirements for each layer to consume or produce Linked Data that enables their combination; that is:

(1) an intersection of (related) entities being studied;
(2) common identifiers for those entities and schema or ontology terms which realise the intersection; and
(3) for the intersection not to preclude independent expansion and iterative modification of layers.

3.1 Collection metadata: 'etree'

While the LMA collection is at the heart of the digital library, we take the *etree*[6] Linked Data service[9] as our *collection metadata* layer. It provides a source of rationalised identifiers describing the LMA that we perpetuate through our tools and analyses.

etree assigns a URI for each *Artist*, *Performance*, *Track*, and *Venue* and asserts relationships between these entities using, as its core, the Music[34], Event, and Similarity[21] Ontologies. etree also associates external references for artists via MusicBrainz IDs, and last.fm venues and GeoNames for locations – these resources are, in effect, further external layers. A SPARQL endpoint is provided for semantic queries, along with RDF and graphical browsing.

etree contains over 12 million RDF triples concerning over 135,000 distinct performances and nearly 5,000 artists with at least one performance. It adopts a conservative method of exact string matching to align artist names, with more sophisticated proximity techniques for geographic alignment. As such, etree will not assert an alignment where the underlying LMA metadata has been entered erroneously. These alignments are published through the etree collection metadata layer, rather than modifying the underlying LMA data. It adds distinct similarity assertions between the candidates, retaining an explicit provenance record of the judgement made by the algorithm – a philosophy we maintain in our other layers.

3.2 Computational analysis: feature extraction

The *computational analysis* layer of CALMA is realised through an audio feature extraction workflow, operating upon music selected using collection metadata, then retrieved directly from the LMA.

Table 1: Vamp plugins (analysis modules) used in the CALMA feature extraction workflow

Vamp plugin	Description
nnls-chroma:chroma[26]	12-dimensional chromagram
nnls-chroma:chordino:simplechord	estimated chord times and labels
qm-mfcc:coefficients	mel-frequency cepstral coefficients
qm-tempotracker:beats	estimated positions of metrical beats
qm-tempotracker:tempo	song tempo
qm-similarity:beatspectrum	rhythmic autocorrelation profiles
qm-keydetector:key	estimated key for each key change
qm-segmenter:segmentation	estimated boundaries of structurally consistent segments
segmentino:segmentation[27]	(alternative) structural segmentation
bbc-speechmusic-segmenter:segmentation	estimated boundaries between speech and music
vamp-libxtract	low-level signal features

In an earlier survey of MIR systems and tools [31], we identified the utility of workflows for merging information and results from different methods and approaches, in particular the possibility of using RDF to re-use and re-combine results within investigation-centric aggregations (Research Objects), enabling 'reuse and exchange of related data beyond that produced and consumed by the MIR system alone'. Here we build upon this proposition, creating a layer for computational analysis in CALMA which algorithmically interrogates the collection audio, and generates output persisting the identifiers relating it to the LMA and collection metadata layers.

The layer takes the form of an extended workflow of Python scripts[10] built around Sonic Annotator [9], a tool for batch analysis of audio which performs feature estimation and extraction from audio files by the application of analysis modules known as Vamp plugins. The workflow *(i)* retrieves and queues the appropriate audio from the LMA using the *internetarchive* library[11]; *(ii)* converts the losslessly compressed audio files retrieved to standard WAV PCM; *(iii)* feeds the audio for parallel computational analysis using Sonic Annotator across a multi-core cluster, before *(iv)* re-processing the Sonic Annotator output into data '*blobs*', feature metadata, and provenance all linked back to the etree identifiers and each other (Section 3.3).

The CALMA feature workflow extracts 41 distinct high-level, intermediate, and low-level features through the application of several Vamp plugins (see Table 1). Examples include the high level, musicological task of structural segmentation; key detection, an intermediate-level task providing a continuous estimate of the key of the music by reference to chromagrams summarizing the spectral content attributable to each pitch class at a given moment in the signal; and low level features such as loudness, crest factor, and spectral centroid.

We also note that several of the classifiers were not necessarily developed with the expectation of being applied to such large and variable corpora, or indeed to live performances. For example, the qm-keydetector Vamp plugin[12] was conceived within a classical music context – the key profiles used by the plugin are drawn from an analysis of Book I of the Well Tempered Klavier by J. S. Bach, recorded at A = 440 Hz equal temperament. This potentially

imperfect fit of feature extractor to signal context (for the LMA, live popular music) is also considered in Section 4.

3.3 Feature metadata and provenance

The *feature metadata* layer is crucial in exposing the computational analysis to other layers in the CALMA system. Through our data model we promote consistent use of identifiers both within the computational analyses, and between them and our wider Linked Data; and to maintain a separation of concerns between functional elements without precluding further independent analysis efforts (or indeed corrections to our own results) in the future. Therefore a second workflow generates web-published Linked Data structures containing the features (from section 3.2), 'feature metadata', and provenance records for the workflow. This layer provides metadata for discovering and analysing the computational workflow results either independently of, or in conjunction with, the other layers.

The computational results themselves are stored unchanged in individual '*blobs*' (compressed tar files) in order to minimise space requirements and preserve the integrity of the computational output. We have designed the feature metadata and provenance structures to be agnostic to the extraction framework used; as such, each blob could theoretically be output from any (e.g. non-Vamp) workflow should suitable feature and provenance metadata be available.

Our RDF-encoded feature metadata and provenance structure (summarised in figure 2) re-uses several ontologies to express the relationships between the performance, audio, and the analysis of the audio within the workflow. In particular we apply the Music Ontology [34] to articulate the different audio versions within the music production and audio analysis processes, also linking back to etree; and the PROV-O onotology [24] to record specific software commands, parameters, and environments of the analysis execution. It is the nature of this descriptive data to be *precise*, and hence detailed, encoding relationships that might otherwise be ignored for reasons of efficiency. While this maximises the possibilities for future detailed investigation – for nothing is lost – it becomes crucial that environments manipulating the data are able to create appropriately reduced views (Section 3.4).

To publish the analysis and metadata through a standard HTTP server[13] and enable resource discovery, our scripts group RDF triples into resources:

analyses – overview of available features for a given track.
analysis_<hash> – provenance for an individual analysis.
analysis_blob_<hash>.tar.bz2 – compressed Sonic Annotator feature extraction output (blob data).
analysis_blob_side<hash> – relates global URIs to local URIs within the blob (e.g. signals, timelines)

The last '*sidecar*' file is required because, although in the specific case of our workflows Sonic Annotator provides its results in RDF, the identifiers used within are local to that execution – the sidecar provides a mapping between these local URIs and the globally unique URIs minted by the CALMA scripts. A further script is provided which, given an etree identifier, will download the analysis data and re-contextualise the analysis RDF using either the local or global identifiers as desired.

---

[10]https://code.soundsoftware.ac.uk/projects/feature-and-metadata-extractor
[11]https://pypi.python.org/pypi/internetarchive
[12]http://vamp-plugins.org/plugin-doc/qm-vamp-plugins.html#qm-keydetector

---

[13]An example can be found at http://calma.linkedmusic.org/data/00/track_00e2b986-5e29-4fb2-9911-f18c3a515d5b/
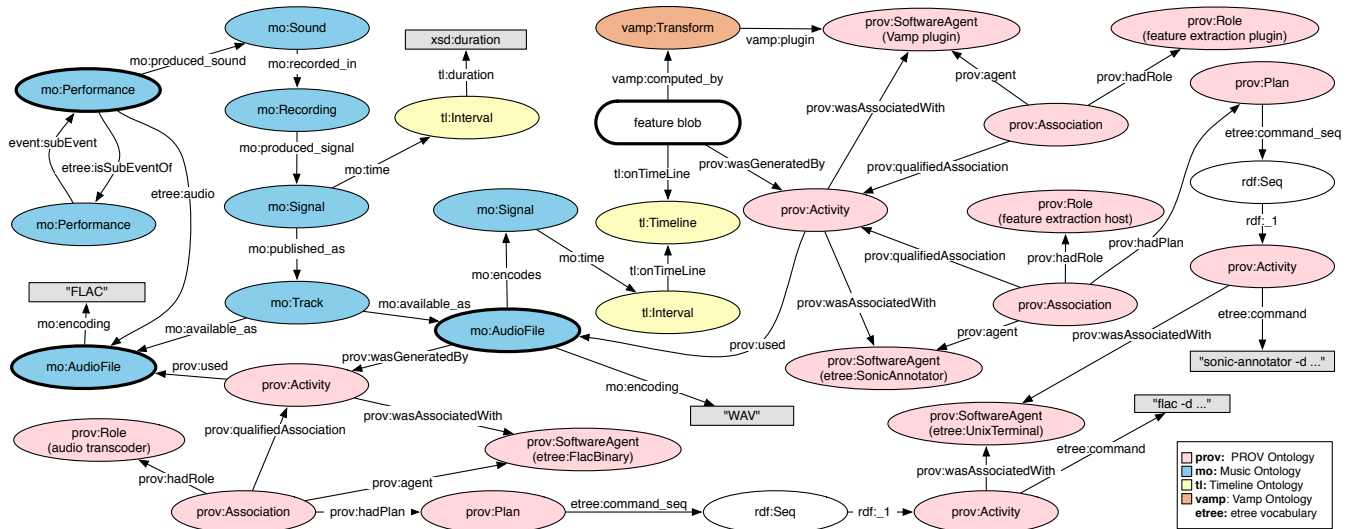
Figure 2: The CALMA data model and use of ontologies

3.4 Exploratory analysis tools

The previously described layers have exposed both the etree collection metadata describing LMA performances, and the CALMA feature data and metadata describing the computational analysis from audio recordings of those performances. Both are imperfect in different ways: etree being derived from hand-typed form data, where entry errors are inevitable; and the computational analysis from the use of feature algorithms on live audio recordings and genres for which they were not designed.

We have also intimated how, by publishing both as Linked Data, it is possible to re-combine the information in different ways through the links between the layers. Having presented a framework in which all the technical foundations for analysis of the data are present, we now introduce an exploratory analysis layer in which an individual can undertake a *meaningful* investigation of such large quantities of data – one in which the information can be gathered, processed, and visualised in a lightweight manner, iteratively working towards a fuller understanding and conclusions. To meet this need, we have developed a dataflow[14] pipeline from the layered digital library into the RStudio[35] environment for data science, in which information can be rapidly presented and represented, and hypothesis can be developed and tested.

At the input to this pipeline, RCalma automates the following: a collection of performed songs from the LMA are selected by user-specified SPARQL queries against the etree data,[15] returning the URIs through which the associated CALMA provenance and feature metadata is retrieved. The returned triples are ingested into a temporary RDF graph using the rrdf R package[43], which describes all the feature data available through CALMA for the specified etree tracks. Further details of features of interest are then accessed in order to retrieve references to the CALMA feature blob for each combination of track and feature. Finally, the feature

blob tarballs are accessed, and the component RDF is extracted and loaded into a further temporary RDF graph in order to retrieve the feature output and related information, e.g. key classifications and event durations. These outputs are then available as standard R data frames, which may be easily restructured and recombined (e.g. using dplyr [42]), visualised (e.g. using ggplot2 [41]), and analysed.

In the following section we describe an investigation undertaken with RCalma using information from across the CALMA layers.

4 AN EXPLORATORY INVESTIGATION: 'TYPICALITY' IN THE LMA

To test the feasibility of using CALMA, as an example of a Linked Data Layered Digital Library, we have undertaken an initial investigation into the LMA through RCalma. This exploratory data analysis was motivated by a desire to understand and cross-check the imperfections noted in the earlier sections: by reference to the potentially inconsistent user-provided concert metadata descriptions (etree), collections of audio tracks *probably* corresponding to hundreds of different performances of the same song by the same artist may be identified; and, by virtue of the potentially noisy feature extractor output, abstract notions of a 'typical' performance's feature profile may be established, and used as validity cues in the verification of user-provided metadata.

In this way it is also representative of the more general styles of investigation a scholarly user might wish to perform within CALMA, or other layered digital libraries. It is not, however, an exhaustive study of LMA content: quite to the contrary, we intend this demonstration to be a start rather than an end; and to show the ease with which enquiries can build upon existing layers, returning value from the initial costs of deploying a layered digital library.

An overview of the investigation that follows is shown in figure 3. Within each layer the processes described in the sections below draw Linked Data from the layers to their left; and enrich their own layer with new information.

---

[14]https://github.com/musicog/rcalma/
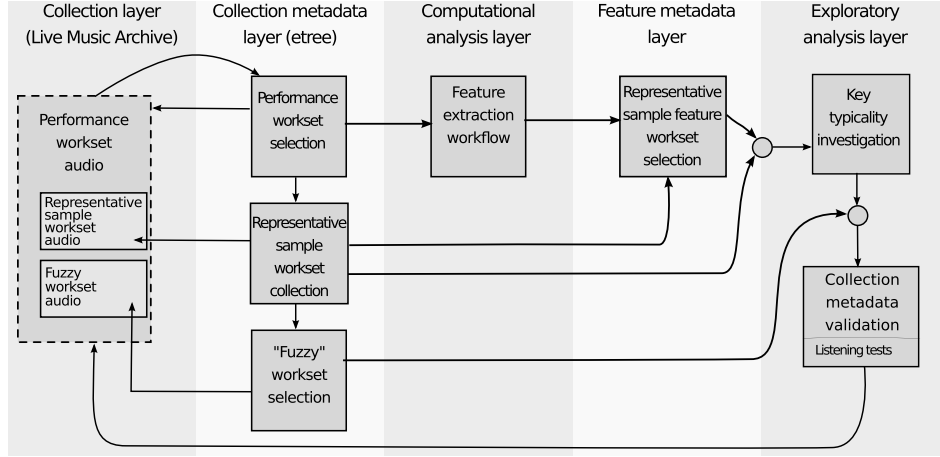[15]Using the SPARQL R package[39]

Figure 3: Digital library layers and processes during the exploratory analyses

4.1 Workset selection

An initial 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 (henceforth referred to as the *performance workset*). The performance workset contains over 16,000 concerts with 16-bit loss-less audio available from the LMA, with numerous artists having sub-collections of song titles that have been performed over 100 times. For example, there are recordings of 41 'Bob Walkenhorst' songs that each have over 100 performances across 445 concerts, 8 songs performed by 'The Brew' over 100 times across 253 concerts, 14 performed over 100 times by 'Cracker' over 322 concerts, 32 such 100-performance songs by 'Drive By Truckers', 22 by 'Guster', and so on. It should be noted that, since it is defined by a query against etree, the workset is only as 'reliable' in its constitution as that etree data, imperfections included.

A representative sample of songs was derived (again using SPARQL) from the *performance workset* (Section 3.2) by selecting artists who performed at least one song with associated CALMA feature data for more than 100 distinct recordings. The query was across both the catalogue metadata and feature metadata layers. For each of these artists (N = 54), 100 performances of one such song were sampled at random. The corresponding 54 songs cover a variety of annotated genres, from indie rock to funk, heavy metal, and country music. Most songs exhibit largely consistent performances in terms of musical structure and other general performance parameters, although several songs feature extensive jam sections with a large degree of variation across performances.

4.2 Key typicality investigation

For the purposes of this investigation, we have focussed on musical key detection obtained via the qm-keydetector Vamp plugin[16], one of the features used in the computational analysis layer. Although, as noted earlier, this was developed within a classical music context rather than for live performances, the feature output remains

decently consistent within our collection. This feature output is visualised for a sub-set of 18 renditions of *The Captain* by the alternative rock band Guster in Figure 4[17]. The key typicality value associated with each rendition quantifies the degree to which the key profile determined for the given rendition of the song is typical of the key profiles across the sample of 100 renditions (song performances). It is calculated as follows:

$$\text{Key typicality} = \sum_{K_i \in Keys} \frac{\Delta K_i}{\Delta R} \cdot \frac{\Delta C_{K_i}}{\Delta C_{K_{max}}}$$

where K_i corresponds to each musical key represented in the sampled set of 100 renditions, ΔK_i is the duration in seconds that the given rendition is in that key according to the feature extractor, ΔR is the duration in seconds of the given rendition of the song, ΔC_{K_i} represents the number of seconds spent in the given key across all of the song's renditions subject to analysis, and $\Delta C_{K_{max}}$ corresponds to the number of seconds spent in the most-commonly represented key across these renditions.

In other words, key typicality is determined by summing over the product of the local (per rendition) and global (all renditions) proportion of time spent in each key, normalised according to the global proportion assigned to the most-commonly represented key in order to enable comparisons across the renditions of different songs that may differ in terms of the number of keys employed. A typicality score of 1 indicates the track has the same distribution as all other tracks for this song; a typicality score of 0 a distribution that is entirely different from all other tracks for this song.

Figure 5 summarises the distribution of key typicality scores across our representative sample. Each density curve corresponds to the distribution of scores determined for the 100 distinct exemplars of each song; the dashed curve corresponds to the distribution of scores across the entire sample, providing a rough (imperfect!) measure of the 'typical' typicality distribution. Taken together, these analyses demonstrates that the key detector produces consistent if somewhat noisy output, providing useful and usable key

---

[16]http://vamp-plugins.org/plugin-doc/qm-vamp-plugins.html#qm-keydetector

[17]The full set of data for all renditions is available from the CALMA website.
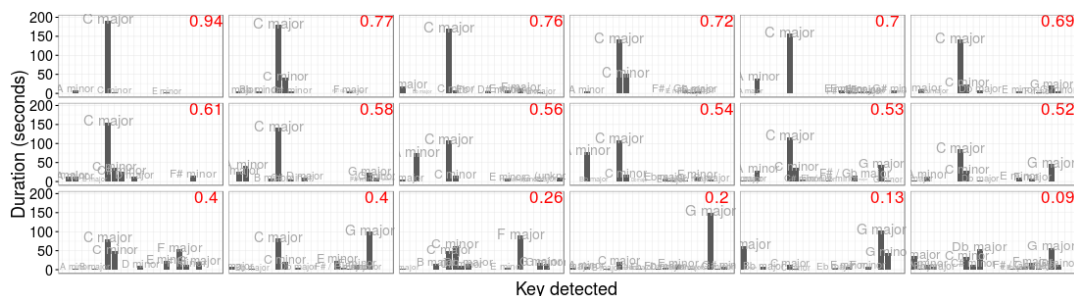
Figure 4: Key detector output for 18 (of 100 analysed) renditions of 'The Captain' by 'Guster'. Red numbers show key typicality scores. Label size corresponds to proportion of track classified as the respective key.
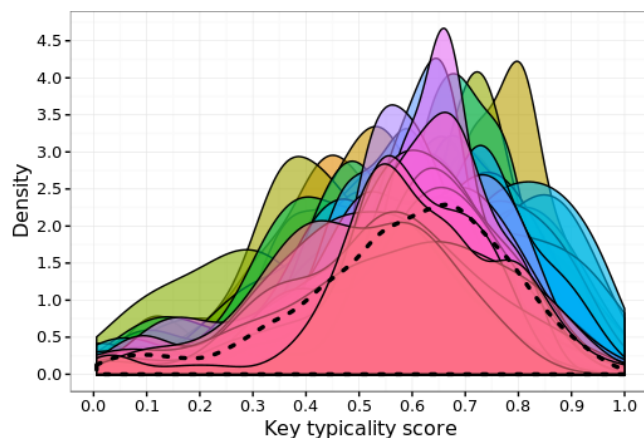
classifications in the majority of cases despite being far-removed from its 'native' habitat of high-quality classical studio recordings.

4.3 Collection metadata validation

We now return to the idea of validating user-provided LMA descriptions, as published in the etree layer, using feature data extracted via computational analysis (Section 3.2) and published in the corresponding metadata layer. The performance workset was constructed by grouping song renditions according to exact string matching of artist and song name metadata. While this technique affords a confidence that the resulting groupings indeed correspond to different renditions of the same target songs (although we cannot guarantee that the uploading user has not mislabelled an unrelated song with the target's metadata), the boundaries drawn around these groupings are highly conservative. Any variations from the target spelling are excluded, potentially neglecting some excellent renditions merely due to minor typographical errors. The key typicality distributions established for our sets of target song performances might offer a musical-content based cue for the cross-validation and potential acceptance of spelling variants.

To test this hypothesis we again used the exploratory analysis layer (Section 3.4) to access bibliographic metadata and key detector output for all song recordings attributed to the 54 artists forming our representative sample. We determined string distance measures (Jaro distance [38]) between each song's user-provided title, and the canonical title of our randomly sampled target songs (each of which had more than 100 instances within the etree data). Those recordings with a Jaro distance lower than a threshold of the mean minus 1.5 standard deviations[18] were selected to form the '*fuzzy workset*', which consisted of 3,485 renditions of songs exhibiting 618 distinct song name variations[19]. Key typicality scores were determined for each track relative to all instances of the target song in the performance workset.

Figure 6 illustrates the correspondence between string distance and key typicality, both that determined between the variants in our fuzzy workset and the corresponding target songs in our representative sample. Points in black correspond to individual fuzzy

---

[18]determined subjectively to capture variations of target strings while incorporating less clear exemplars and some strings likely representative of alternate songs.
[19]Fuzzy titles incorporating certain special characters were withheld from this analysis, as these are typically used to indicate relationships between multiple songs captured in the same recorded track; e.g. 'song1>song2' is conventionally used by etree uploaders to indicate a seamless transition from one song to the next.

Figure 5: Each curve is the distribution of key typicality across 100 performances of one of 54 songs in our representative sample; dashed curve aggregate for the sample.

variants; groupings sharing the same X-value are artefacts of fuzzy renditions sharing the same song-name variant. Points in red correspond to individual target tracks, and as such all share a string distance of 0, but are jittered along the X-axis to improve visibility. The distribution of red points on the Y-axis mirrors the dashed line in figure 5; this is echoed in the low-string-distance fuzzy variants, corroborating that these points likely correspond to renditions of the same target songs.

We performed listening tests[20] to verify our assumptions about the distribution of song-name variants, comparing the fuzzy rendition and against an exemplar – the rendition with the highest typicality score – of each corresponding target. We tested on all fuzzy variants with a string distance of 0.1 or less ($N = 196$), covering the low-distance 'echo'of the target distribution. Of these variants, which included song titles such as 'Breath', 'Abiliene', 'Honey B', and 'Carnvial Time' (with corresponding target songs 'Breathe', 'Abilene', 'Honey Bee', and 'Carnival Time', respectively), the overwhelming majority indeed represented renditions of the target. One variation – labelled 'Sun Dog' for the target 'Sundog' by artist 'Perpetual Groove' – turned out to be a short (35 second) clip of bass guitar improvisation between tracks, rather than the long,

---

[20]Playlists were generated within RCalma using the CALMA layered digital library.
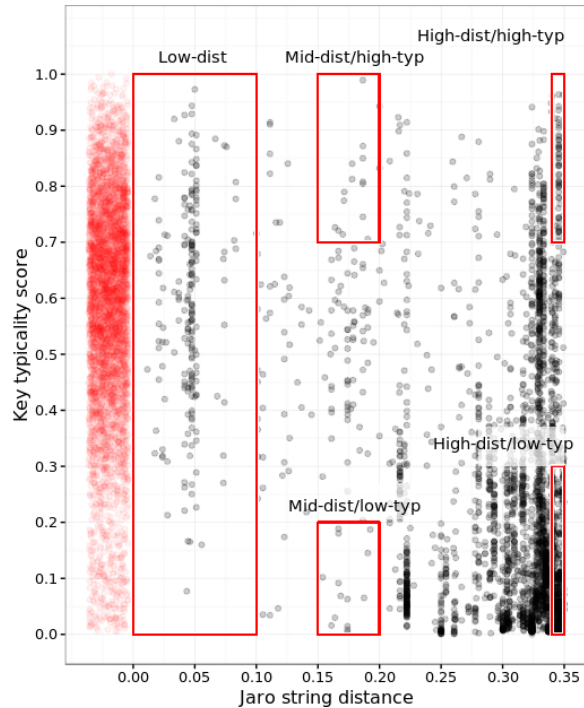
Figure 6: Key typicality and string distance. Targets, plotted as red points, share a string distance of 0. Boxed regions were manual verified with listening tests.

progressive art-rock jam more typical of the target; this variation was located near the bottom of the typicality distribution (0.27).

Further listening tests compared high- and low-typicality variants (greater than .7 or lower than .2 on the typicality scale), with a medium string distance measure of between .15 and .2 ($N = 14$ in each group; variant titles included 'Shanen' and 'Madeline' for targets 'Shaken' and 'Abilene'). Inspection of these groups was interesting: whereas all variations in the high-typicality group were true renditions of their corresponding target, 3 of the variants in the low-typicality group were different from their target. Each of these was labelled 'Sweethearts explanation' for the target song 'Sweethearts' by the artist 'Camper Van Beethoven', which turned out to be (three distinct instances of) introductory remarks by the band discussing the appropriateness of the song's lyrics, which had been penned in the 1980's, to the present political situation. Several other variations in this group, although ostensibly corresponding to their target songs (sharing lyrics and general song structure), differ significantly from the targets in terms of their arrangement.

A final comparison was made between variants exhibiting large string distance measures of more than .34 and their targets ($N = 56$ and $N = 592$, respectively; variant titles include 'Amazing Grace' for 'Mama Grady', and 'Madeline' for 'Abilene'). Further listening tests were applied in cases where mis-labelling seemed plausible (e.g. the title 'Madeline' might plausibly have been a misheard 'Abilene'). Inspection of these groups revealed the large-string-distance tracks to all be renditions of songs other than the target.

Overall, this clearly indicates that fuzzy song variations by target artists are likely to correspond to renditions of the target song where the string distance between fuzzy variation and target is very low, suggesting that the quality of the user-provided metadata is such that small title variations are tolerable while maintaining accuracy. Our findings regarding key typicality are more mixed: true representations of target songs exhibit a variety of scores. We are encouraged that the few discovered misclassified variations in the low- and medium-string-distance groups exhibited a low typicality, suggesting this measure may fulfil a useful filter function in this context.

Due to the large number of exemplars, listening tests lasted only long enough to confirm the presence or absence of a match. However, our anecdotal impression is that low typicality scores correlate with low quality tracks: factors including loud and constant crowd noise and chatter, large recorded portions of stage banter before and after songs, and poor recording quality. Based on this preliminary indication, typicality might serve as a proxy measure for audio quality. We intend to pursue this hypothesis, applying blinded listening tests masking the typicality score to avoid confirmation bias.

5 CONCLUSIONS AND FUTURE WORK

We have presented a layered digital library providing multimedia access to sound, user-provided metadata, and audio-derived feature metadata of the Live Music Archive, in turn allowing novel exploratory analyses across and within its layers.

CALMA provides analytical 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. As an illustrative investigation, we have presented a study of 'key typicality'. While the comparison of typicality with string distance has yielded some interesting and informative initial results, we anticipate that the utility of this approach will be improved by the addition of typicality measures for other features. Ongoing experiments with chord and tempo typicality show promise both individually and for the possibility of a multi-dimensional typicality measure, to be explored and reported in future work.

Our analyses have also highlighted weaknesses in the application of contemporary audio feature algorithms to live performances; our feature metadata can help create worksets to assist in their improvement. Tantalisingly, a cursory query of LMA data suggests some performances may contain multiple sources of recordings, which may enable comparative masking of background 'noise'.

The LMA itself has proven to be an extremely interesting – if imperfect – corpus, ripe for analysis and investigation. By continuing this work we hope to enable further insight into this valuable cultural resource. The layered approach, although under-explored since Lynch, has proved very valuable in supporting the presented analysis of interlinked collection and feature metadata.

More generally, CALMA provides an implemented example of how a *layered* digital library can be realised, and the benefits in capability and flexibility it can bring to scholarly users. Although under-explored since Lynch, the layered approach has proved very valuable in supporting our investigations. We believe the investment required to create and maintain layers is repaid through the

support they provide for iterative and incremental research, and easy adaptability to move from one investigation to another – for example, from key typicality to catalogue validation – through re-use of self-describing Linked Data layers.

ACKNOWLEDGMENTS

REFERENCES

[1] Getaneh Alemu, Brett Stevens, Penny Ross, and Jane Chandler. 2012. Linked Data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models. *New Library World* 113, 11/12 (2012), 549–570.

[2] Bruno Angeles, Cory McKay, and Ichiro Fujinaga. 2010. Discovering Metadata Inconsistencies.. In *ISMIR*. 195–200.

[3] David Bainbridge, Xiao Hu, and J. Stephen Downie. 2014. 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*. ACM, 1–8.

[4] Mathieu Barthet, Mark Plumbley, Alexander Kachkaev, Jason Dykes, Daniel Wolff, and Tillman Weyde. 2014. Big chord data extraction and mining. In *Proceedings of Conference on Interdisciplinary Musicology, Berlin*.

[5] Mert Bay, John Ashley Burgoyne, Tim Crawford, David De Roure, J Stephen Downie, Andreas Ehmann, Benjamin Fields, Ichiro Fujinaga, Kevin Page, and Jordan BL Smith. 2009. Structural Analysis of Large Amounts of Music Information. *White paper, available via http://www. diggingintodata. org/Home/AwardRecipientsRound12009/StructuralAnalys isofLargeAmountsofMusic/tabid/179/Default. aspx* (2009).

[6] Sean Bechhofer, Kevin Page, and David De Roure. 2013. Hello Cleveland! Linked Data publication of live music archives. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on*. IEEE, 1–4.

[7] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *ISMIR 2011: Proceedings of the 12th International Society for Music Information Retrieval Conference, October 24-28, 2011, Miami, Florida*. University of Miami, 591–596.

[8] C. Bizer, T. Heath, and T. Berners-Lee. 2009. Linked Data - the story so far. *International Journal on Semantic Web and Information Systems* 5, 3 (2009), 1–22.

[9] C. Cannam, M. Sandler, M. O. Jewell, C. Rhodes, and M. d'Inverno. 2010. Linked Data and You: Bringing Music Research Software into the Semantic Web. *Journal of New Music Research* 39, 4 (2010).

[10] T. Crawford, B. Fields, D Lewis, and K Page. 2014. Explorations in Linked Data practice for early music corpora. In *Digital Libraries (JCDL), 2014*. IEEE, 309–312.

[11] David De Roure. 2011. Machines, methods and music: On the evolution of e-research. In *High Performance Computing and Simulation (HPCS), 2011 International Conference on*. IEEE, 8–13.

[12] David De Roure. 2014. Executable Music Documents. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*. ACM, 1–3.

[13] David De Roure, Graham Klyne, Kevin R Page, John PN Pybus, and David M Weigl. 2015. Music and Science: Parallels in Production. In *Proc. 2nd International Workshop on Digital Libraries for Musicology*. ACM, 17–20.

[14] David De Roure, Graham Klyne, Kevin R Page, John PN Pybus, and David M Weigl. 2015. Music and Science: Parallels in Production. In *Proc. 2nd International Workshop on Digital Libraries for Musicology*. ACM, 17–20.

[15] J. Stephen Downie, Kris West, and Xiao Hu. 2008. Dynamic Classification Explorer for Music Digital Libraries. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '08)*. ACM, New York, NY, USA, 422–422.

[16] Daniel PW Ellis and Bertin-Mahieux Thierry. 2012. Large-scale cover song recognition using the 2D Fourier transform magnitude. In *Proceedings of the 13th international society for music information retrieval conference*. 241–246.

[17] María Hallo, Sergio Luján-Mora, Alejandro Maté, and Juan Trujillo. 2016. Current state of Linked Data in digital libraries. *Information Science* 42, 2 (2016), 117–127.

[18] Victor Henning and Jan Reichelt. 2008. Mendeley-A Last. fm for research?. In *eScience, 2008. eScience'08. IEEE Fourth International Conference on*. IEEE, 327–328.

[19] Ioannis Papadakis, Konstantinos Kyprianos, and Michalis Stefanidakis. 2015. Linked Data URIs and Libraries: The Story So Far. *D-Lib Magazine* 21, 5/6 (2015).

[20] J. Jett, T. W. Cole, C. Maden, and J. S. Downie. 2016. The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections. *Journal of Open Humanities Data* 2 (2016).

[21] Kurt Jacobson, Yves Raimond, and Mark B Sandler. 2009. An Ecosystem for Transparent Music Similarity in an Open World.. In *ISMIR*. 33–38.

[22] J. Jett, T. Nurmikko-Fuller, T. W. Cole, K. R. Page, and J S Downie. 2016. Enhancing Scholarly Use of Digital Libraries: A Comparative Survey and Review of Bibliographic Metadata Ontologies. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. ACM, 35–44.

[23] Angela Kroeger. 2013. The road to BIBFRAME: the evolution of the idea of bibliographic transition into a post-MARC Future. *Cataloging & classification quarterly* 51, 8 (2013), 873–890.

[24] Timothy Lebo, Satya Sahoo, Deborah McGuinness, K Belhajjame, J Cheney, and others. 2013. PROV-O: The PROV Ontology. W3C Recommendation, 30 April 2013. *World Wide Web Consortium* (2013).

[25] Clifford Lynch. 2002. Digital Collections, Digital Libraries and the Digitization of Cultural Heritage Information. *First Monday* 7, 5 (2002). http://firstmonday.org/ojs/index.php/fm/article/view/949

[26] M. Mauch and S. Dixon. 2010. Approximate Note Transcription for the Improved Identification of Difficult Chords. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*.

[27] M. Mauch, K. C. Noland, and S. Dixon. 2009. Using Musical Structure to Enhance Automatic Chord Transcription. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009)*.

[28] Rodger J. McNab, Lloyd A. Smith, Ian H. Witten, Clare L. Henderson, and Sally Jo Cunningham. 1996. Towards the Digital Music Library: Tune Retrieval from Acoustic Input. In *Proceedings of the First ACM International Conference on Digital Libraries (DL '96)*. ACM, New York, NY, USA, 11–18.

[29] Terhi Nurmikko-Fuller, Alan Dix, David M Weigl, and Kevin R Page. 2016. 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*. ACM, 17–24.

[30] T. Nurmikko-Fuller, J. Jett, T. Cole, C. Maden, K. R. Page, and J. S. Downie. 2016. A Comparative Analysis of Bibliographic Ontologies: Implications for Digital Humanities. In *Digital Humanities 2016: Conference Abstracts*. 639–642.

[31] Kevin R Page, Ben Fields, David De Roure, Tim Crawford, and J Stephen Downie. 2013. Capturing the workflows of music information retrieval for repeatability and reuse. *Journal of Intelligent Information Systems* 41, 3 (2013), 435–459.

[32] Kevin R Page, Benjamin Fields, Bart J Nagel, Gianni O'Neill, David C De Roure, and Tim Crawford. 2010. Semantics for music analysis through linked data: How country is my country?. In *e-Science (e-Science), 2010 IEEE Sixth International Conference on*. IEEE, 41–48.

[33] Alastair Porter, Dmitry Bogdanov, Robert Kaye, Roman Tsukanov, and Xavier Serra. 2015. Acousticbrainz: a community platform for gathering music information obtained from audio. In *International Society for Music Information Retrieval (ISMIRfi15) Conference*.

[34] Yves Raimond, Samer Abdallah, Mark Sandler, and Frederick Giasson. 2007. The Music Ontology. *Proceedings of the International Conference on Music Information Retrieval* (2007).

[35] RStudio Team. 2015. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA. http://www.rstudio.com/

[36] Alexander Schindler, Rudolf Mayer, and Andreas Rauber. 2012. Facilitating Comprehensive Benchmarking Experiments on the Million Song Dataset.. In *ISMIR*. 469–474.

[37] Jordan Bennett Louis Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J Stephen Downie. 2011. Design and creation of a large-scale database of structural annotations.. In *ISMIR*, Vol. 11. 555–560.

[38] M.P.J. van der Loo. 2014. The stringdist package for approximate string matching. *The R Journal* 6 (2014), 111–122. Issue 1. http://CRAN.R-project.org/package=stringdist

[39] Willem Robert van Hage, with contributions from: Tomi Kauppinen, Benedikt Graeler, Christopher Davis, Jesper Hoeksema, Alan Ruttenberg, and Daniel Bahls. 2013. *SPARQL: SPARQL client*. https://CRAN.R-project.org/package=SPARQL R package version 1.16.

[40] Avery Wang and others. 2003. An Industrial Strength Audio Search Algorithm.. In *Proceedings of the 4th international conference on music information retrieval*. 7–13.

[41] Hadley Wickham. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org

[42] Hadley Wickham and Romain Francois. 2015. *dplyr: A Grammar of Data Manipulation*. http://CRAN.R-project.org/package=dplyr R package version 0.4.3.

[43] Egon Willighagen. 2014. Accessing biological data in R with semantic web technologies. (2014). http://dx.doi.org/10.7287/peerj.preprints.185v3.

[44] Thomas Wilmering, György Fazekas, and Mark Sandler. 2010. The effects of reverberation on onset detection tasks. In *Audio Engineering Society Convention 128*. Audio Engineering Society.