Linked Data and Music: Current Projects and Opportunities

Elizabeth Joan Kelly

J. Edgar & Louise S. Monroe Library

Loyola University New Orleans

Abstract:

Music collections have special considerations when it comes to indexing and information retrieval. This article describes Linked Data and the Semantic Web and summarizes current projects which utilize Linked Data to enhance online music collections.

Keywords:

Linked data, music literature, information retrieval, music

Article Classification:

Article

This is a post-print [post-peer review] version of a manuscript originally published in *The Indexer* 33, no. 1 (March 2015): 2-7. http://www.theindexer.org/

Linked Data and Music: Current Projects and Opportunities

Introduction

The indexing and description of music involves considerations that do not affect textual media. Increases in online searching, both by music professionals and nonprofessionals, also create new opportunities for the information professional in how to best get users to resources. Many of the challenges particular to music information retrieval (MIR) stem from the problems of providing context and cross-referencing related resources. This article describes Linked Data and the Semantic Web and summarizes current projects which utilize Linked Data to enhance online music collections.

Major problems

The major issues involved in indexing and describing music collections have been previously discussed in this journal. The primary concerns can be summarized as follows:

- Search for music-related text tends to assume a familiarity with music theory and history terminology. Databases need to allow for search including melody, rhythm, chords, and contour so indexes are searchable by both trained and untrained musicians (Kelly, 2010: 164).
- Cross-referencing of notated music is necessary for musical fragments, especially for folk and hymn tunes where the same motif may appear under many different names (Kelly, 2010: 164).
- Search of notated music needs to differentiate between plain and embellished versions of melodies (Kelly, 2010: 164).
- Multiple manifestations may be found of one work--for example, scores published at different times and/or in different languages; or recordings, video, and scores all for the same piece (Kelly, 2010: 164). Search needs to both find the appropriate format of the piece, and link to other formats.

MIR behaviors have also recently changed in part because of the widespread use of the internet by both non-professional and academic researchers. Users are more frequently looking for songs instead of albums online due to online music stores (iTunes, Amazon) (Gracy, Zeng, & Skirvin, 2013: 2082). 43.8% of music information searches online are for help identifying musical works, 36.4% for identifying artists, and 16.7% for locating music recordings (Gracy, Zeng, & Skirvin, 2013: 2082). Known-item searches are assisted by query-by-humming tools and music recommendation services like Pandora and Last.fm (Gracy, Zeng, & Skirvin, 2013: 2082). Finally, the context in which a musical work is created is increasingly important to researchers; a 2004 study found that the top three categories of data that users seek are title, lyrics, and artist information (Gracy, Zeng, & Skirvin, 2013: 2082). As will be discussed later in this article, context and the full lifecycle of a work (including not only manifestations but additional scholarship) is rarely included in library bibliographic records.

Some attempts to solve these problems have emerged. Recent MIR has begun to describe the work instead of the document, allowing for better linking of different formats (Kelly, 2010: 164). Uniform titles can be applied to all manifestations of the same work to link them together (Kelly, 2010: 164). Functional Requirements for Bibliographic Records (FRBR), a holistic approach to the structure and relationships of bibliographic records that is supposed to better differentiate between different formats of a work, is helpful for issues such as multiple formats of a musical score and connecting works to their historical, analytical, critical, technical and performance background (Pietras & Robinson, 2012: 557). Information not included in published data, however, such as which edition of a piece is performed on a specific recording, is still lacking in classification (Pietras & Robinson, 2012: 557).

The idea of cross-referencing and linking between different manifestations of a music work is persistent throughout the literature. Where the traditional print index can assist in providing cross-references, an online search for a single song lacks this ability. A key solution may be found in structured data.

Structured data and indexes

As previously discussed in this journal, indexing for data formats (like the eBook standard EPUB) differs from print indexes which are created through the 'single intellectual exercise' of creating both content and print layout (Goodenough, 2013: 133). The process is different for data formats, where 'the intellectual structure of the index must be encoded in its own right,' and both content and relationships (hierarchies and cross-references) must be encoded and then turned over to software developers (Goodenough, 2013: 133). Writing indexes for data formats creates a target audience of systems, not humans (Goodenough, 2013: 133). Problems can occur when the data is shared out of its original context, so metadata is necessary to understand what the data is for (Goodenough, 2013: 134).

Linked Data

One method of publishing structured data in a way that links related data and therefore minimizes the problem of providing context is Linked Data. When Linked Data is made freely available for use and modification, it is called Linked Open Data (the data equivalent of Open Source software) (Rodriguez, 2009: 38). The network of Linked Data combines to form the Semantic Web. As a result, the Semantic Web includes both relationships of and access to data (as opposed to just collections of data) (W3, 2013).[1] The Principles of Linked Data were laid out by Tim Berners-Lee in 2006:

1. Use URIs (Uniform Resource Identifiers) as names for things
2. Use HTTP URIs so that people can look up those names

[1] The terminology involved here begins to get complicated, and exact definitions may be up for debate. Some prefer to use the term Web of Data over Semantic Web; others don't believe the two are the same thing. Tom Heath's 'Linked Data? Web of Data? Semantic Web? WTF?' attempts to clarify (2009), as does the Linked Data page at W3 (2013). For the purposes of this article, Semantic Web will be used exclusively to refer to the network of data created by publishing Linked Open Data (LOD).

3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs so that they can discover more things.

Some further explanation of these principles may be necessary. Resource Description Framework (RDF) is a standard for linking URIs using statements in order to create a network of URIs (Rodriguez, 2009: 38). Technologies like RDF, OWL, SKOS, SPARQL, and others retrieve information from the Semantic Web, drawing inferences about related data using ontologies (W3, 2013). Ontologies are used to define concepts and relationships and any possible restrictions on using them.[2]

Just as Open Source software has implications for software developers, Open Data has implications for application and data providers. Currently, web applications maintain their own data sources; with Open Data, the same data can be used by different developers for different users (Rodriguez, 2009: 40).

There are certainly opportunities for improving indexing and information retrieval using Linked Data. There are additional implications for the indexing of music resources which already have their own challenges. Following is a review of literature related to linked data and music collections.

Linked Data and Music: Current Projects

Publishing Linked Open Data

The Sheet Music Consortium (SMS) is a collaboration of several universities to digitally publish sheet music and metadata as inspired by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Davison, Sugiyama, McAulay, & Horning, 2013: 141). The consortium was begun in 2002 but revised their website services in 2013 for easier contributor submission. More than a dozen addition institutions began contributing to the SMS and, as a result, created a more diversified collection of sheet music metadata; however, the wealth of new information incorporated a wider variety of systems, schema, and standards, resulting in a non-normalized data set (Davison, Sugiyama, McAulay, & Horning, 2013: 142).

Following this second phase of the SMS project, part of the consortium from the University of California, Los Angeles (UCLA) decided to experiment with publishing normalized metadata in the SMS as Linked Open Data (LOD) (Davison, Sugiyama, McAulay, & Horning, 2013: 143). In order to add to existing LOD but also contribute rare or unique data, the group at UCLA chose to experiment with publishing linked data of publisher names from the consortium records (Davison, Sugiyama, McAulay, & Horning, 2013: 145). Publisher information provides its own unique challenges as it is usually uncontrolled in catalog

[2] The terms vocabulary and ontology are sometimes interchanged while other times ontology is used to refer to complex collections where hierarchies and relationships are clearly delineated. Again, W3C provides a definition (2013). For the purposes of this article, ontology will be used exclusively to refer to collections of terms used in Semantic Web information retrieval.

records and thus transcribed exactly as it appears on the resource itself, leading to vast inconsistencies in publisher names (Davison, Sugiyama, McAulay, & Horning, 2013: 145). In choosing to work with LOD, especially attractive to the UCLA group was the fact that RDF allows for adding links to an identifier instead of revising existing metadata (Davison, Sugiyama, McAulay, & Horning, 2013: 146). The group used the text analysis tools Google Refine (now Open Refine) and Voyeur to identify and normalize the names of the publishers across all the SMC collections (Davison, Sugiyama, McAulay, & Horning, 2013: 155). They then established permanent identifiers for distinct publisher entities using the University of California Curation Center's EZID service (Davison, Sugiyama, McAulay, & Horning, 2013: 157). Each ID was assigned a URI, and then the normalized publisher names were rewritten into the consortium data as 'user supplied data,' and RDF records for each publisher were published providing links to related publishers and other relevant information (Davison, Sugiyama, McAulay, & Horning, 2013: 157). In this particular instance, LOD provided the SMS with a partial solution to the problem of non-normalized metadata sets, contributing to overall discovery.

The Music Ontology

The Music Ontology was created to be a flexible ontology allowing for both the nonprofessional and the trained musicologist to accurately describe music. For the purposes of this project, the authors defined ontology as 'a formal description of concepts and relationships in a domain' (Raimond, Abdallah, Samer, & Sandler, 2007: 417). The Music Ontology is comprised of both Timeline and Event Ontologies (Raimond, Abdallah, Samer, & Sandler, 2007: 418). The Timeline Ontology is built on top of the concepts Interval and Instant (Raimond, Abdallah, Samer, & Sandler, 2007: 418). The Event Ontology is made up factors (like a musical instrument), agents (like performers), and products (like the actual sound produced in a performance) (Raimond, Abdallah, Samer, & Sandler, 2007: 418). Finally, FRBR's concepts of Work, Manifestation, and Item are used along with Friend-of-a-Friend's (FOAF) concepts of Person and Group (Raimond, Abdallah, Samer, & Sandler, 2007: 419). Since the goal of the ontology is applicability to a wide variety of users, the ontology is divided until several levels of expressiveness: musical metadata (editorial info), cultural metadata (music genres, social networking info), and content-based information (Raimond, Abdallah, Samer, & Sandler, 2007: 421). The Music Ontology has gone on to be used in a number of Web of Data projects:

1. Zitgist: used the Music Ontology to convert the MusicBrainz metadata repository into RDF
2. DBTune: publishes and interlinks several Creative Commons music repositories and links to DBpedia data source (see more about DBPedia under 'Cross-domain linking')
3. Foafing-the-music: reads ID3 tags from MP3s, queries other music web services, and then sends that information in RDF using the Music Ontology back to the user
4. EASAIER (Enabling Access to Sound Archives through Integration, Enrichment and Retrieval): provides an interface for producing instance data from an audio archive using the Music Ontology as a knowledge representation foundation
5. OMRAS2 (see more under '*Online Music Recognition and Search II (OMRAS2)')*

(Raimond, Abdallah, Samer, & Sandler, 2007: 421)

The proliferation of the Music Ontology through these other projects contributes to the goal of creating a semantic web, but web data isn't the only data that can be enhanced by LOD. The Metadata Vocabulary Junction project studies how libraries can benefit from LOD. A subproject of this research

explored what library bibliographic data could be linked to available music information sources by mapping MARC to the Music Ontology. Preliminary research explored different methodologies for aligning bibliographic data with Linked Data (Gracy, Zeng, & Skirvin, 2013). The researchers found that library data structures (like MARC) typically describe the object in one instance of its lifecycle (for example, a particular recording of an opera), and leave out context which is usually present in archival or museum description; however, interpretation or reuse of materials is rarely addressed in bibliographic, archival, or museum description (Gracy, Zeng, & Skirvin, 2013: 2078). The researchers chose to study a music ontology because of the extreme importance of context in musical works; while bibliographic records may describe scores and performances, web users now are looking for information about the performers and other 'agents' involved in the music-making process (Zeng, Gracy, & Skirvin, 2013: 257). In addition, the widespread use of mobile devices means that users are more interested in getting straight to the item they want (ie, the song) than in looking through entire albums (Zeng, Gracy, & Skirvin, 2013: 258). MARC records do not always list songs for entire albums or sheet music collections, and even if they do, sometimes the 'Notes' field with this data is not visible (Zeng, Gracy, & Skirvin, 2013: 258).

For this study, the researchers chose eleven music-related datasets to analyze including BBC Music, 3 servers from DBTune, Discogs, John Peel Sessions, Last.FM, Moseley Folk Festival, MusicBrainz, MusicNet, and Surge Radio (Zeng, Gracy, & Skirvin, 2013: 260). They created a crosswalk to the Music Ontology using both MARC records and non-MARC (typically Dublin Core) metadata to find common metadata elements that could be connected to linked data (Zeng, Gracy, & Skirvin, 2013: 262). In the process of mapping, the researchers found the Music Ontology to be much more specific than MARC and Dublin Core records (Zeng, Gracy, & Skirvin, 2013: 263). They finally developed a list of the most useful elements for mapping to linked data: Title Information, Responsible Body, Subject and Genre, Physical Characteristics, and Location (Zeng, Gracy, & Skirvin, 2013: 265).

Online Music Recognition and Search II (OMRAS2)

The Music Ontology was created within the Online Music Recognition and Search II project (OMRAS2). The Centre for Digital Music has worked on a number of projects related to the data extracted from audio and music, and some of these have involved Semantic Web technologies. One relevant project focused on the problem of sharing music data: copyright protections are frequently an impediment to the development of the field of music informatics, extending from commercial audio recordings to newer iterations (computer-notated scores, MIDI transcriptions, etc) of out-of-copyright works (Cannam, Sandler, Jewell, Rhodes, & d'Inverno, 2010: 313). OMRAS2 uses online collections of music to focus studying on annotation and search, and to enable researchers to share data (Cannam, Sandler, Jewell, Rhodes, & d'Inverno, 2010: 313). Semantic Web and Linked Data technologies were selected as tools for this project for the following reasons:

1. There is substantial previous research on describing music metadata for Semantic Web
2. There is already data about music in the Semantic Web (MusicBrainz, BBC broadcast data, Wikipedia entries)
3. Semantic Web languages can use existing code
4. RDF is suitable for publishing results because it is human-readable
5. The Semantic Web is good for one-off projects because it is data-oriented instead of service-oriented

(Cannam, Sandler, Jewell, Rhodes, & d'Inverno, 2010: 314)

Some of the OMRAS2 tools which use Semantic Web technologies include:

- Vamp audio analysis plugin API
- audio feature extractor Sonic Annotator
- audio analysis and annotation Sonic Visualise
- audio collection contextual search database audioDB

The ultimate goal of OMRAS2 is that the projects will give researchers confidence that data and descriptions of data can be understood independently of specific tools and therefore they won't have to learn new ways of searching (Cannam, Sandler, Jewell, Rhodes, & d'Inverno, 2010: 323).

Cross-domain linking

For content that has multiple disciplines, LOD is extremely useful but may be challenging to connect. Yet another use of the Music Ontology can be found in the BBC Music and BBC Programmes services. The BBC websites are vast but are divided into domains which are managed by different staff. This has led to a problem with cross-domain linking; for example, Madonna is a musician (as defined by MusicBrainz), as well as an actor and person (as defined in Wikipedia), but information about her on various BBC websites was not bridged (Scott et al., 2009: 724). These different data sources also use different vocabularies, leading to more difficulty in cross-domain linking. The BBC thus decided to attempt to use DBpedia to provide a common controlled vocabulary, and add 'topic badges' to existing webpages for better discovery and linking of related web pages (Scott et al., 2009: 725). Because the BBC is not the only data provider for music information, they chose to collect URIs for artists and tracks from the MusicBrainz metadata service (Scott et al., 2009: 726). The Music Ontology was used for MusicBrainz (Gracy, Zeng, & Skirvin, 2013: 2083). The BBC Programmes service also was developed to provide one URI per programme broadcast by the BBC (Scott et al., 2009: 726). The principles of Linked Data were implemented in the design of both the BBC Music and Programmes. Web identifiers are represented in XHTML for the user interface, RDF is used for structured data, and both the XHTML and RDF representations link to further web identifiers (Scott et al., 2009: 726-727). In order to link to even more BBC domains, the BBC's legacy auto-categorization system CIS was bridged to DBpedia, which brings data from Wikipedia to the Semantic Web (Scott et al., 2009: 727-728). DBpedia provides both Linked Data URIs and structured data about concepts and relationships, making it an ideal source to act as the controlled vocabulary connecting the various BBC domains (Scott et al., 2009: 728). Additionally, text documents (like web sites and news articles) were tied in using the named entity extraction system Muddy Boots (Scott et al., 2009: 731). Finally, aggregation (or topic) pages were used to pull together BBC Programmes data (CIS mapped to DBpedia) and BBC News articles using DBpedia vocabulary, interlinking both structured and unstructured data found across the BBC domains (Scott et al., 2009: 733).

Concerns and Challenges

While Linked Data provides some possible solutions to the problems facing MIR, it is not a cure-all. The Connected Music Experience (CME) is a consortium formed to promote technical standards for enhanced digital media packages, including music releases (Kellogg, 2011). CME initially used the Music

Ontology but later created their own ontology in order to describe non-music releases (Kellogg, 2011). Eventually, low participation caused by the difficulty of working with Semantic Web technologies led the CME to adopt more established web technologies and proprietary metadata formats (Kellogg, 2011). Author Gregg Kellogg, who served as Architect/Technical Work Group Chair for the CME from 2007-2011, wrote, 'In many ways, the music industry is not ready for many of the open aspects of an RDF format; the concept of using existing universal identifiers (such as DBPedia URIs) that they do not directly control can be a barrier, and they are not yet prepared to maintain their own publicly available repository of unique identifiers representing their artists, musical works and releases.' Using the CME as an example, a truly Open Web will never materialize while some industries are unable to maintain the control they desire over their own products. In addition, the learning curve still inherent in publishing Linked Data, and particularly in working with RDF, is largely influenced by academic interests and inhibits potential content contributors (Kellogg, 2011).

Conclusion

Recent research involving music and Linked Data includes publishing music metadata as LOD; creating music ontologies and linking them to bibliographic records; tools for music researcher data sharing; and cross domain linking using Linked Data. All of these projects contribute to a greater awareness of Linked Data and how it can specifically benefit music collections and, in particular, the problem of linking context and related materials in music metadata and other records. While the Semantic Web has not yet reached its full potential, continued research and publishing of Linked Data can continue to improve MIR.

Resources

Projects and tools:

BBC
http://www.bbc.com/
The BBC Programmes website (http://www.bbc.co.uk/programmes) and Music website (http://www.bbc.co.uk/music) were designed using the principles of Linked Data. Other domains may also integrate Linked Data.

Linked Data
http://linkeddata.org/
Website for the Linked Data community.

The Metadata Vocabulary Junction (MV-Junction) Project
http://lod-lam.slis.kent.edu/
Research project questioning how libraries can benefit from Linked Data. Includes studies and resources specifically related to music description.

MusicNet
http://musicnet.mspace.fm/
A suite of resources and tools for aligning musicology data to minimize redundancy in reference URIs of music composers.

Music Ontology
http://musicontology.com/
Website for the Music Ontology vocabulary with guidelines on publishing music metadata

OMRAS2
http://www.omras2.org/
Framework for annotating and searching collections of both recorded music and digital score representations, funded by the EPSRC (Engineering and Physical Sciences Research Council).

The Sheet Music Consortium:
http://digital2.library.ucla.edu/sheetmusic/index.html
Tools and services for discovering online sheet music

Databases and datasets:

DBPedia
http://wiki.dbpedia.org/

Crowd-sourced structured data extracted from Wikipedia.

DBTune
http://dbtune.org/
Servers of LOD datasets including MusicBrainz, Audio-Scrobbler, Magnatune, and more.

Discogs
http://www.discogs.com/
User-generated database of discography information. The dataset is available for download from http://www.discogs.com/data/.

freeDB
http://www.freedb.org/
General Public License (GPL), user-generated database of CD information.

Last.fm
http://www.last.fm/
Music recommendation site. Develops listener preference profiles based on data collected from songs listened to by users. Also allows for user-generated content. The dataset is available through the Million Song Dataset.

Linked Jazz
http://linkedjazz.org/
Tools and research about LOD and its application to digital archives of jazz history.

Live Music Archive Linked Data
http://etree.linkedmusic.org/
Server of RFD-converted Linked Data from the Internet Archive's Live Music Archive (available at https://archive.org/details/etree)

Million Song Dataset
http://labrosa.ee.columbia.edu/millionsong/
Collection of metadata made up of community-contributed datasets including the SecondHandSongs dataset, the musiXmatch dataset, the Echo Nest Taste Profile Subset, and the Last.fm dataset.

MusicBrainz
http://musicbrainz.org/
A database of publicly contributed and available music metadata made Linked Data-compliant with LinkedBrainz (http://linkedbrainz.org/) using the Music Ontology.

seevl.fm
https://developer.seevl.fm/

A turnkey API for music discovery and retrieval built on RDF data collected from the Open Web.

References

Berners-Lee, T. (2006). 'Linked data'. Available at http://www.w3.org/DesignIssues/LinkedData.html (accessed 30 April 2014).

Cannam, C., Sandler, M., Jewell, M. O., Rhodes, C., & d'Inverno, M. (2010). 'Linked Data and You: Bringing Music Research Software into the Semantic Web.' *Journal Of New Music Research*, 39(4), 313-325.

Davison, S., Sugiyama, Y., McAulay, E., & Horning, C. (2013). 'Enhancing an OAI-PMH Service Using Linked Data: A Report from the Sheet Music Consortium'. *Journal Of Library Metadata*, 13(2/3), 141-162.

Goodenough, S. (2013). 'Structured data, standards, and indexes'. *The Indexer* 31(4), 133-137.

Gracy, K. F., Zeng, M., & Skirvin, L. (2013). 'Exploring methods to improve access to Music resources by aligning library Data with Linked Data: A report of methodologies and preliminary findings.' *Journal Of The American Society For Information Science & Technology*, 64(10), 2078-2099.

Heath, T. (2009). 'Linked Data? Web of Data? Semantic Web? WTF?' *Tom Heath's Displacement Activities*. Available at http://tomheath.com/blog/2009/03/linked-data-web-of-data-semantic-web-wtf/ (accessed 14 May 2014).

Kellogg, G. (2011). 'CME and the Semantic Web.' Gregg Kellogg. Available at http://greggkellogg.net/cme-semweb (accessed 10 June 2014).

Kelly, E. (2010). 'Music indexing and retrieval: current problems'. *The Indexer* 28(4), 163-166.

Pietras, M., & Robinson, L. (2012). 'Three views of the "musical work": bibliographical control in the music domain'. *Library Review*, 61(8/9), 551-560.

Raimond, Y., Abdallah, Samer, & Sandler, Mark. (2007). 'The music ontology'. In *Proceedings of the International Symposium on Music Information Retrieval*. Vienna, Austria. Available at http://ismir2007.ismir.net/proceedings/ISMIR2007_p417_raimond.pdf (retrieved 30 April 2014).

Rodriguez, M. A. (2009). 'A Reflection on the Structure and Process of the Web of Data.' *Bulletin of the American Society for Information Science & Technology*, *35*(6), 38–43.

Scott, T., Kobilarov, G., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C. & Lee, R. (2009). 'Media Meets Semantic Web--How the BBC Uses DBpedia and Linked Data to Make Connections'. Presented at the 6th European Semantic Web Conference. Available at http://derivadow.files.wordpress.com/2009/06/eswc2009-bbc-dbpedia-2.pdf (retrieved 30 April 2014).

World Wide Web Consortium (W3) (2013). 'Linked data'. Available at http://www.w3.org/standards/semanticweb/data (accessed 30 April 2014).

Zeng, M. L., Gracy, K. F., & Skirvin, L. (2013). 'Navigating the Intersection of Library Bibliographic Data and Linked Music Information Sources: A Study of the Identification of Useful Metadata Elements for Interlinking'. *Journal of Library Metadata*, *13*(2/3), 254–278.