JazzCats: Navigating an RDF triplestore of integrated performance metadata

Daniel Bangert
University of Göttingen
bangert@sub.uni-goettingen.de

Terhi Nurmikko-Fuller
Australian National University
terhi.nurmikko-fuller@anu.edu.au

J. Stephen Downie
University of Illinois Urbana-Champaign
jdownie@illinois.edu

Yun Hao
University of Illinois Urbana-Champaign
yunhao2@illinois.edu

ABSTRACT

Applying Linked Data techniques to musical metadata can facilitate new paths of musicological inquiry. *JazzCats: Jazz Collection of Aggregated Triples* is a prototype project interlinking four discrete jazz performance datasets and external sources as references. Tabular, relational, and graph legacy datasets have necessitated different RDF production and ingestion workflows to support scholarly study of performance traditions. This paper highlights critical processes of data curation for digital libraries, including quality assessment of the ingested datasets. In addition, we describe research questions enabled by *JazzCats*, raise musicological implications, and offer suggestions to overcome current limitations.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**; **Resource Description Framework (RDF)**; *Web Ontology Language (OWL)*; *Ontologies*;

KEYWORDS

jazz, performance, metadata, ontologies, semantic web, SPARQL, digital musicology, Linked Data

ACM Reference Format:
Daniel Bangert, Terhi Nurmikko-Fuller, J. Stephen Downie, and Yun Hao. 2018. JazzCats: Navigating an RDF triplestore of integrated performance metadata. In *5th International Conference on Digital Libraries for Musicology (DLfM '18), September 28, 2018, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3273024.3273031

1 INTRODUCTION

The increasing availability of digital resources for musicology presents opportunities to integrate complementary information using Linked Data technologies. For musicologists, publication of Linked Data about performances, recordings, musicians, venues, and associated ephemera, enables navigation across datasets at scale and can facilitate the tracing of musical histories and traditions. While the release of Linked Open Data (LOD) continues to grow at a rapid rate,[1] much data of musicological interest is not yet available as LOD. In their survey of musical data on the Web, Daquino et al. [3] note the scarcity of resources that are Linked Data ready, especially with regard to music notation. Of the 327 resources analysed,[2] only 12 were categorised as 5 Star Linked Data according to the Open Data scheme (i.e. published in RDF and having a SPARQL endpoint).

This paper focuses on resources related to jazz, using the *JazzCats* prototype project to demonstrate how networks formed by real-time musical collaboration can be published as LOD and navigated as a scholarly resource.

1.1 Linked Jazz

A pioneering example of the use of Linked Data with jazz metadata is *Linked Jazz*.[3] This prosopography of musicians [10] is captured as RDF and openly available for query and download.[4] The ontological structure informing the knowledge graph is predominantly based on a cluster of properties linking specific instances for which no owl:Class has been defined. These properties include Vocab.org relationships[5] (e.g. :acquaintanceOf, :influencedBy), those from the Music Ontology (e.g. :collaborated_with) [13], and project-specific ones (e.g. :bandmember and :inBandTogether).

1.2 JazzCats: Jazz Collection of Aggregated Triples

A recent addition to Linked Data projects focusing on jazz is *JazzCats: Jazz Collection of Aggregated Triples*,[6] a prototype resource that brings together four discrete, but complementary datasets:

(1) A discography of the jazz standard *Body and Soul*;
(2) The Weimar Jazz Database (*WJazzD*), containing information about transcribed jazz solos [12];
(3) *Linked Jazz* [6, 9–11];
(4) *J-DISC*,[7] a specialised digital library of jazz recording sessions.

These datasets were transformed to RDF by applying two different methods for either tabular or relational data [8]. The first method consisted of the mapping of tabular data to a previously

DLfM '18, September 28, 2018, Paris, France

ACM ISBN 978-1-4503-6522-2/18/09…$15.00
https://doi.org/10.1145/3273024.3273031

[1] https://lod-cloud.net/
[2] musoW: Musical Data on the web (https://github.com/enridaga/musow)
[3] https://linkedjazz.org/
[4] https://linkedjazz.org/access/
[5] http://vocab.org/relationship/
[6] http://jazzcats.cdhr.anu.edu.au/ (migrated from http://jazzcats.oerc.ox.ac.uk/)
[7] http://jdisc.columbia.edu/

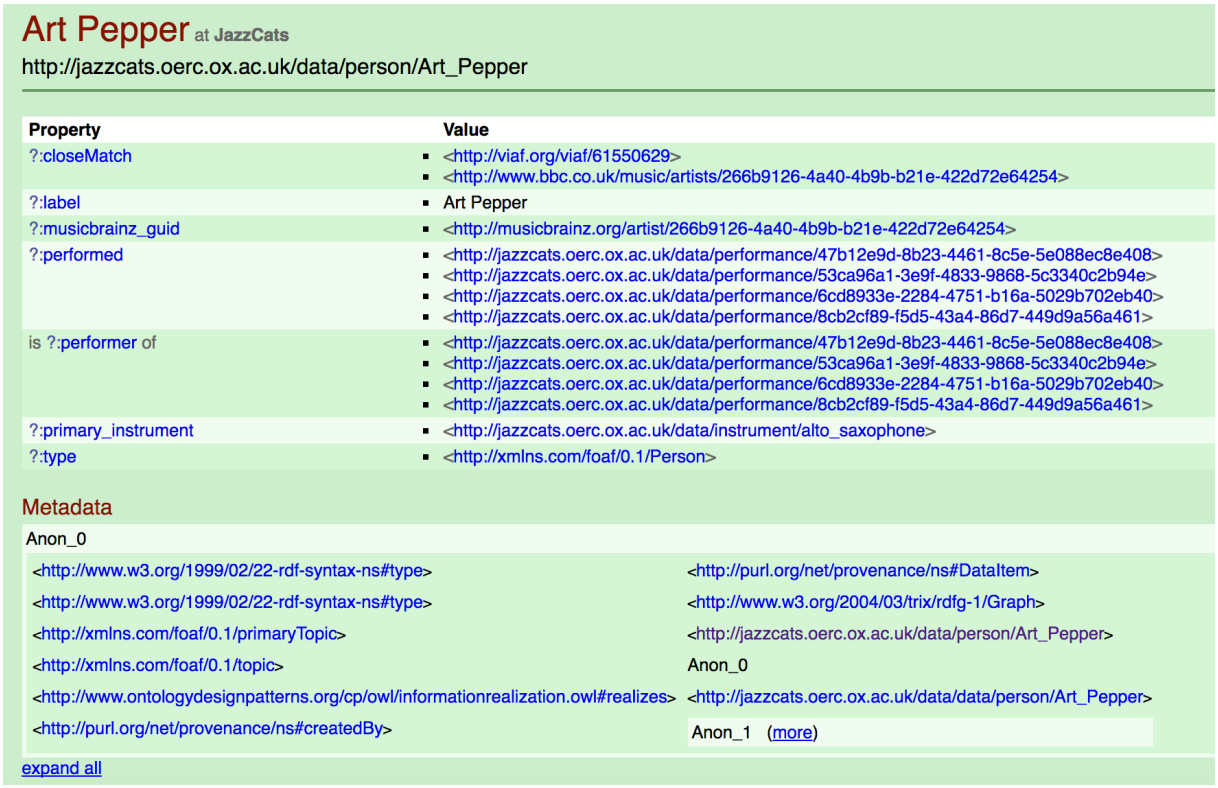

Figure 1: Pubby user-interface displaying data for Art Pepper.

created ontological structure using Open-Source software from the University of Southern California (Web-Karma).[8] The second method was a largely automated workflow using a relational database (SQLite3) together with an automated tool (D2RQ[9]) designed to publish relational data as RDF. Both methods are described in further detail in [7]. The project currently contains more than 90,500,000 RDF triples.

By providing an integrated access point for related jazz performance metadata, *JazzCats* enables navigation across its constituent projects and facilitates paths of inquiry for jazz scholars. Recognizing that not all potential users have the necessary prerequisite skills to compose and execute appropriate SPARQL queries, access to the underlying knowledge graph has been made available through a Pubby[10] interface.

This interface enables users to explore *JazzCats* data using the 'follow-your-nose' method[11] of information discovery. In the example illustrated by Figure 1, 14 triples (all of which have the *JazzCats* URI as the subject) capture information about Art Pepper, aligning this data with the external authorities VIAF[12] and BBC Music[13], providing a human-readable label, a cluster of performances (expressed bidirectionally), Pepper's primary instrument, and assert the subject URI as an instance of foaf:Person. By clicking on a URI, a user can discover further information, expressed similarly in RDF triples, regarding that new subject. Using this method, the entire knowledge graph can be navigated.

2 DATA CURATION AND QUALITY ASSESSMENT

At several points in the workflows used for RDF production, forms of quality assessment were conducted to verify, clean and enrich data. In particular, we sought to reconcile and align data with external sources and URIs to improve the interoperability and reusability of *JazzCats* LOD. To demonstrate, we give two examples.

(1) For Body&Soul, detailed data cleaning was required, drawing on domain and data curation expertise. The Body&Soul dataset was derived from supplementary data [2] posted in pdf format on the author's website.[14] For the *JazzCats* workflow, a .csv of these data was obtained on request and then analysed, cleaned and enriched. Discographic information was combined (where applicable) with key and tempo information from [2]. Names, instruments, and dates were then clustered and normalised within OpenRefine.[15] For disambiguation, *JazzCats* identifiers for performers were aligned

[8] http://usc-isi-i2.github.io/karma/
[9] http://d2rq.org/
[10] http://wifo5-03.informatik.uni-mannheim.de/pubby/
[11] https://www.w3.org/2001/sw/wiki/Linking_patterns
[12] https://viaf.org/
[13] https://www.bbc.co.uk/music/
[14] http://josebowen.com/body-and-soul/
[15] http://openrefine.org/

with VIAF URIs using a VIAF reconciliation service.[16] Further URIs for performers from MusicBrainz and BBC Music were assigned by querying the MusicBrainz API.[17] For musical works, MusicBrainz and Wikidata URIs were used, and the latter were also applied to places. The resulting .csv is openly available [1].

(2) For *Linked Jazz*, semantic integration issues were considered prior to ingesting the existing RDF. Specifically, to facilitate schema-level alignments between *Linked Jazz* and the other projects, equivalences between their instance data and *JazzCats* URIs (themselves instances of foaf:Person), were asserted using skos:closeMatch to allow for a degree of flexibility not provided by owl:sameAs [4]. Alignment was possible for 226 entities and the application of skos:closeMatch can be seen in Figure 1.

3 RESEARCH OPPORTUNITIES

The initial impetus for *JazzCats* started from the research opportunities of querying performance metadata at scale:

- How does a performance tradition become established? Once established, how does it continue to evolve?
- Which performance features are passed on? Which performances have these features? How are these features transmitted?

By linking the datasets inherent to *JazzCats*, these questions were refined to centre on the nature of shared characteristics between recordings, musicians, and performances:

- Which performance features occur frequently in a range of different recordings?
- Which recordings feature the same performers?
- What performance features are common between recordings by the same artist?

JazzCats enables specific queries to be constructed that navigate and identify performances and recordings with particular characteristics or connections according to musical features or artists. For example:

- Which performances were recorded in a specific place in a particular style? For example, swing recorded in London.
- Which performances feature a particular combination of instruments, in a specific key? For example, recordings with trumpet and piano, performed in the key of Db.
- Which performances were recorded in New York by artists that played with a particular artist? For example, artists who played with Roy Eldridge during their career. A SPARQL query for this example is given in Figure 2.

The results from these queries highlight the complementary nature of these datasets. Although seemingly disjoint as questions, the same instance-level entities are featured in the generated results.

4 WORKED EXAMPLE

In *Who Plays the Tune in "Body and Soul"? A Performance History Using Recorded Sources* [2], Bowen highlights a number of influential recordings and innovations in performances of *Body and Soul*. One influential change was the introduction of a "jump chorus" in which the rhythm section doubles the pulse, creating a double-time feel.

[16] http://refine.codefork.com/

[17] https://wiki.musicbrainz.org/Development/XML_Web_Service/Version_2

```
prefix mo: <http://purl.org/ontology/mo/>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
prefix skos: <http://www.w3.org/2004/02/skos/core#>
prefix foaf: <http://xmlns.com/foaf/0.1/>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix lj: <http://linkedjazz.org/ontology/>
prefix event: <http://purl.org/NET/c4dm/event.owl#>

SELECT DISTINCT ?artist WHERE {
?artist a foaf:Person ;
       mo:performed ?performance ;
       skos:closeMatch ?another_ID ;
       rdfs:label ?artist_name .

?work a mo:MusicalWork ;
       rdfs:label "Body and Soul" .

<http://dbpedia.org/resource/Roy_Eldridge>
       lj:playedTogether ?another_ID .

?performance a mo:Performance ;
       mo:performance_of ?work ;
       mo:produced_sound ?sound .

?sound a mo:Sound ;
       mo:recorded_in ?recording .

?recording a mo:Recording;
       event:place <http://www.wikidata.org/wiki/Q60> . }
```

Figure 2: Artists who played with Roy Eldridge and made a recording of *Body and Soul* in New York.

```
prefix mo: <http://purl.org/ontology/mo/>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
prefix skos: <http://www.w3.org/2004/02/skos/core#>
prefix foaf: <http://xmlns.com/foaf/0.1/>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix event: <http://purl.org/NET/c4dm/event.owl#>

SELECT DISTINCT ?artist ?connection_to_Eldridge ?performance
WHERE {
?artist a foaf:Person ;
      mo:performed ?performance ;
      skos:closeMatch ?another_ID .

?performance a mo:Performance ;
       mo:performance_of ?work .

?work a mo:MusicalWork ;
      rdfs:label "Body and Soul" .

<http://dbpedia.org/resource/Roy_Eldridge>
      ?connection_to_Eldridge ?another_ID . }
```

Figure 3: Recordings of *Body and Soul* by artists with a connection to Roy Eldridge.

The jump chorus was introduced in Chu Berry and Roy Eldridge's 1938 recording of the tune and this performance feature is found in several later recordings. Bowen discusses some similarities and differences between recordings with a jump chorus, but further context can be given through knowledge of the relationships between artists of the era. The query in Figure 3 demonstrates the types of connections between Roy Eldridge and other musicians that recorded *Body and Soul*, and several of these artists also employed a jump chorus in their recordings. For example, multiple types of connections can be shown to have existed between Roy Eldridge and the saxophonist Charlie Parker (namely rel:knowsOf, rel:hasMet, lj:playedTogether, and rel:friendOf). Parker made a number of recordings of *Body and Soul* between 1940-1950 and most of these contain a jump chorus.[18]

In addition, analysis of the *J-DISC* dataset (sub-set of *JazzCats* data) reveals the direct and indirect relationships between Roy Eldridge and Charlie Parker. Within *J-DISC*, there are 38 artists who performed with both Eldridge and Parker. This is shown in Figure 4, where the relative thickness of an edge represents the relative frequency of artists playing in the same session [5]. This sub-network of indirect connections includes further artists whose recordings of *Body and Soul* feature a jump chorus (e.g. Don Byas, Cozy Cole, Ben Webster, Slam Stewart).

These forms of analysis highlight the personal and professional connections between artists that can help to transmit innovations, transform expectations of how a tune is played, and contribute to the creation of a performance tradition.

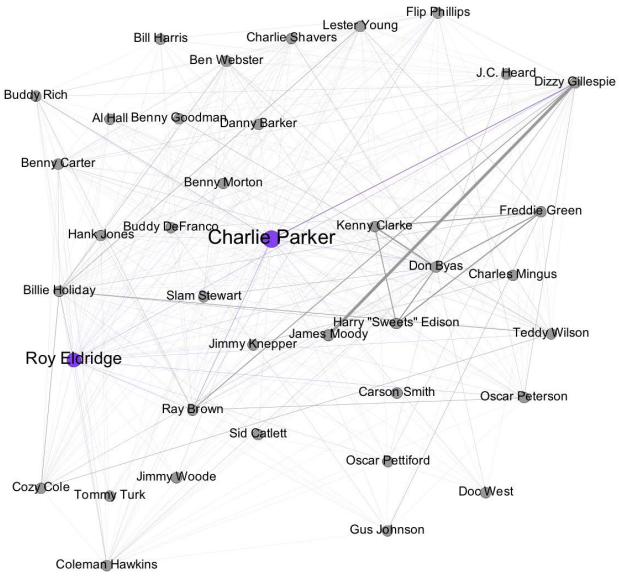

Figure 4: *J-DISC* sub-network featuring Roy Eldridge and Charlie Parker.

[18]Bowen notes Parker's direct quotation of Eldridge, indicating familiarity with the 1938 Berry/Eldridge recording [2].

5 CONCLUSION

This paper has described the application of Linked Data techniques to performance metadata and the implications in terms of research opportunities for musicology. As a prototype, *JazzCats* demonstrates how data quality assessment, semantic integration and publication of LOD can be achieved in ways that take into account variations in data content and format. Current limitations of the project include the limited accessibility for scholars unfamiliar with constructing SPARQL queries and the lack of integrated performance data in notated or audio form. Future work should address gaps in the data and develop more automated methods for both querying and expanding the *JazzCats* triplestore.

6 ACKNOWLEDGEMENTS

The authors would like to acknowledge the support and assistance of John Pybus and Dr David Weigl from the University of Oxford's e-Research Centre, and Dr Alfie Abdul-Rahman from King's College London. For the datasets aggregated by *JazzCats* we want to thank Professor José Antonio Bowen, President of Goucher College; Martin Pfleiderer, Hochschule für Musik Franz Liszt Weimar, and the Jazzomat Research Project team; M. Cristina Patuelli, Pratt Institute, and the *Linked Jazz* team; and the *J-DISC* team at Columbia University.

REFERENCES

[1] Daniel Bangert. 2016. JazzCats Body and Soul discography. https://doi.org/10.5281/zenodo.163886

[2] José Antonio Bowen. 2015. Who Plays the Tune in "Body and Soul"? A Performance History Using Recorded Sources. *Journal of the Society for American Music* 9, 3 (2015), 259–292.

[3] Marilena Daquino, Enrico Daga, Mathieu d'Aquin, Aldo Gangemi, Simon Holland, Robin Laney, Albert Merono Penuela, and Paul Mulholland. 2017. Characterizing the Landscape of Musical Data on the Web: State of the art and challenges. In *Second Workshop on Humanities in the Semantic Web - WHiSe II.* ACM.

[4] Harry Halpin, Patrick J. Hayes, James P. McCusker, Deborah L. McGuinness, and Henry S. Thompson. 2010. When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data. In *The Semantic Web – ISWC 2010*, Peter F. Patel-Schneider et al. (Eds.). Springer, Berlin, 305–320.

[5] Yun Hao, Kahyun Choi, and J. Stephen Downie. 2016. Exploring J-DISC: Some Preliminary Analyses. In *Proceedings of the 3rd International Workshop on Digital Libraries for Musicology (DLfM 2016).* ACM, New York, NY, USA, 41–44.

[6] Michael C. Heller. 2016. Review: Linked Jazz. *Journal of the American Musicological Society* 69, 3 (2016), 879–891. https://doi.org/10.1525/jams.2016.69.3.879

[7] Terhi Nurmikko-Fuller, Daniel Bangert, Alan Dix, David M. Weigl, and Kevin R. Page. 2018. Building Prototypes Aggregating Musicological Datasets on the Semantic Web. *Bibliothek Forschung und Praxis.* 42 (2018), 206–221. https://doi.org/10.1515/bfp-2018-0025

[8] 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 *Proceedings of the 3rd International Workshop on Digital Libraries for Musicology (DLfM 2016).* New York, USA, 17–24.

[9] M. Christina Pattuelli. 2012. Personal name vocabularies as linked open data: A case study of jazz artist names. *Journal of Information Science* 38, 6 (2012), 558–565. https://doi.org/10.1177/0165551512455989

[10] M. Cristina Pattuelli, Matt Miller, Leanora Lange, Sean Fitzell, and Carolyn Li-Madeo. 2013. Crafting linked open data for cultural heritage: Mapping and curation tools for the linked jazz project. *Code4Lib Journal* 21 (2013).

[11] M. Christina Pattuelli, A. Provo, and H. Thorsen. 2015. Ontology Building for Linked Open Data: A Pragmatic Perspective. *Journal of Library Metadata* 15, 3-4 (2015), 265–294.

[12] Martin Pfleiderer and Klaus Frieler. 2010. The Jazzomat project. Issues and methods for the automatic analysis of jazz improvisations. In *Concepts, Experiments, and Fieldwork: Studies in Systematic Musicology and Ethnomusicology.*, Rolf Bader, Christiane Neuhaus, and Ulrich Morgenstern (Eds.). Peter Lang, Frankfurt am Main, 279–295.

[13] Yves Raimond, Samer A. Abdallah, Mark B. Sandler, and Frederick Giasson. 2007. The Music Ontology. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007).* 417–422.