Swinging Triples: Bridging Jazz Performance Datasets using Linked Data

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

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

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

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

ABSTRACT

The jazz performance metadata prototype *JazzCats:Jazz Collection of Aggregated Triples* uses Linked Data to bridge four discrete jazz music datasets: *Linked Jazz*, with prosopographical and interpersonal information about musicians; the Weimar Jazz Database (*WJazzD*), containing musicological metadata; a discography of the jazz standard *Body&Soul*; and *J-DISC*, a fourth independent but complementary and extensive discographic project. Through the use of custom-built ontological structures the data, originally stored in various different information structures, has been converted to RDF and merged together in a single triplestore. The result is a new digital resource that can be used to support and enrich scholarship and research in musicology and performance studies.

CCS CONCEPTS

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

KEYWORDS

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

ACM Reference Format:
Terhi Nurmikko-Fuller, Daniel Bangert, Yun Hao, and J. Stephen Downie. 2018. Swinging Triples: Bridging Jazz Performance Datasets using Linked Data. In *1st International Workshop on Semantic Applications for Audio and Music (SAAM '18), October 9, 2018, Monterey, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3243907.3243914

1 INTRODUCTION

JazzCats: Jazz Collection of Aggregated Triples[1] is an interdisciplinary jazz performance metadata prototype, which makes use of the Linked Data information publication paradigm to enable the expansion of musicological study and the enrichment of data available for performance analysis through aggregated datasets.

As an aggregator, *JazzCats* links four discrete and heterogeneous but complementary datasets. Alignments between them have been possible at both schema-level (e.g. `foaf:Person`) and instance level (for example, Roy Eldridge as a specific instance of `foaf:Person`, who appears in more than one of the datasets). These similarities between the data categories are illustrated in Table 1.

The original iteration of the project [6] consisted of 90,493,082 RDF triples representing the information held in three datasets: *Linked Jazz*, with prosopographical and interpersonal information about musicians [9]; the Weimar Jazz Database (*WJazzD*), containing musicological metadata [10]; and a discography of the jazz standard *Body&Soul* [2]. The recent addition of *J-DISC* data[2], a specialised digital library containing information about jazz recording sessions, brings a fourth, independent but complementary, dataset to the mix - one containing information such as musicians' skills beyond the sphere of musicology ("engineer", "Graphic Design", etc. captured in 6,451 additional triples), and their gender (there are 5,735 triples capturing either "Female", "Male", or "NULL"). Enriching musicological data includes additional, rarer, instruments: of note are the nine triples for "cowbell".[3]

In this paper, we report on the practical processes necessary for the addition of the *J-DISC* dataset to *JazzCats*, recapping the heuristics of the development of ontological structures and workflows [6, 7] used in earlier iterations of the project.

Table 1: Complementary datasets bridged by *JazzCats*

Data category/ Subprojects	*Body&Soul*	*WJazzD*	*Linked Jazz*	*J-DISC*
Place	✓			✓
Title	✓	✓		✓
Performance type		✓		
Event metadata	✓			✓
Performance	✓			✓
Person	✓	✓	✓	✓
Musical work	✓	✓		✓
Instrument	✓	✓		✓
Digital signal metadata		✓		

[1] Available from https://jazzcats.oerc.ox.ac.uk

[2] Available from http://jdisc.columbia.edu/

[3] We hypothesise *JazzCats* needs more cowbell.

SAAM '18, October 9, 2018, Monterey, CA, USA

ACM ISBN 978-1-4503-6495-9/18/10…$15.00
https://doi.org/10.1145/3243907.3243914

2 THE ORIGINAL *JAZZCATS*: DESCRIPTION AND RDF CONVERSION

The first iteration of *JazzCats* brought together three different projects, each with their own discrete datasets, captured in three different types of data structures: the tabular data of *Body&Soul* was made available to us as a CSV, and was subsequently enriched and tidied in the process; the relational data of *WJazzD* was in a SQLite3 database; *Linked Jazz* was available for ingestion as RDF.

In order to facilitate the needs of this heterogeneous collection of data types, the tabular data for *Body&Soul* and the relational data of *WJazzD* were converted into RDF using bespoke ontological structures and two distinct workflows [6, 7] that matched the needs and requirements of each of the native data structures. All the resulting RDF was imported into a single instance of a Virtuoso triplestore, where the triples were separated into distinct named graphs based on the dataset of their origin (e.g. http://jazzomat.hfm-weimar-jazzcats.de/people and https://linkedjazz.org/people, both of which assert the `rdf:type` relationship between specific instances and `foaf:Person`) but could be queried simultaneously.

2.1 *Body&Soul*

The structured but non-relational data of *Body&Soul* was converted to RDF using a purpose-built ontological structure (Figure 1), which relied heavily on the Music Ontology (MO) [11]. The RDF production workflow was completed using Open Source software developed by the University of Southern California, Web-Karma[4], which provides a click-and-point user interface and allows uploaded data to be mapped to any ontology. The workflow is time-consuming but results in high quality RDF triples which require little post-production processing or tidying.

2.2 *WJazzD*

The RDF production workflow for *WJazzD*'s musicological metadata [6, 7] was a largely automated one, completed using the D2R server [3], and tied where possible to the properties and classes of MO. This having been said, the vast majority of the properties and many of the classes used for the underlying ontological structure (Figure 2) remain project specific ones (e.g. `jcv:solo_info`).

2.3 *Linked Jazz*

Uniquely amongst the datasets within *JazzCats*, the *Linked Jazz* data was already published as RDF [9]. The design decisions regarding the ontology (Figure 3) were thus not part of the workflow process for *JazzCats*, which focused on the ingestion of the data [6].

2.4 New additions

Recently, a fourth dataset has been added to the *JazzCats* project. The *J-DISC* dataset, the RDF production workflow and the design of the underlying ontological structure all tapped into existing heuristics and reusing earlier workflows and ensuring schema-level alignments between the datasets (Table 1).

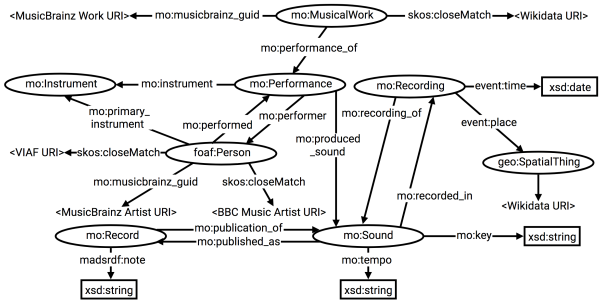

Figure 1: Bespoke ontological structure for *Body&Soul*

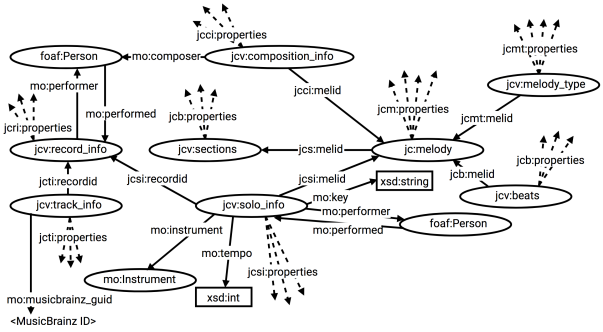

Figure 2: Bespoke ontological structure for *WJAzzD*

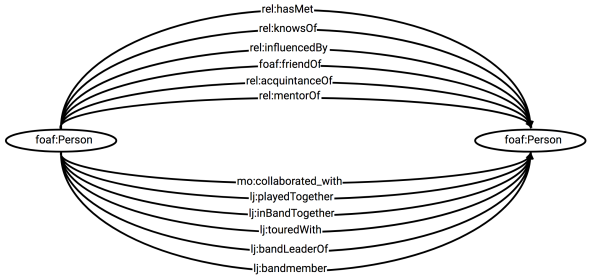

Figure 3: The ontological structure for *Linked Jazz*

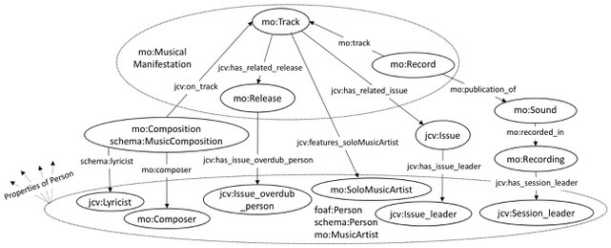

Figure 4: Bespoke ontological structure for *J-DISC*

[4]https://github.com/usc-isi-i2/Web-Karma/wiki

3 *J-DISC* DATA: DESCRIPTION AND RDF CONVERSION

J-DISC is a digital library (as defined in [1]) specialising in jazz recording sessions. The dataset of 19 relational tables was created as a snapshot [5] from the original data produced by the Center for Jazz Studies at Columbia University[5]. The *J-DISC* data can be downloaded in two ways: SQL dump[6], or CSV files[7]. The dataset contains over 2,700 unique recording sessions and their associated metadata, capturing details such as performer name (over 5,700 individuals), venue (predominantly locations in the US), and the number of musicians taking part in any one session (anywhere between 1 and 53). Other types of data contained within *J-DISC* includes information regarding the composition (captured as a `mo:Composition`); the individuals involved in its creation (`jcv:Lyricists` and `mo:Composers`); musicians and their skills (both musicological and otherwise), as well as personal characteristics (ethnicity, gender); information related to the performed sound (sessions, tracks, and issues); and composition titles.

3.1 Designing the ontology for *J-DISC*

Most of the classes and properties (Figure 4) for the *J-DISC* data repeat those utilised in other parts of *JazzCats*. Examples of the properties from the MO include `mo:published_as` (for connecting the Session to the Track); `mo:composer`; and `mo:instrument`. Classes include `mo:MusicalWork` (for the composition) as well as `mo:Recording` (for the session). Other ontologies were used minimally and only where they could be utilised unambiguously (e.g. `schema:MusicComposition` (class) and `schema:lyricist` (property)). For the most part, classes and properties for which no clear equivalent could be found in the MO, the default *JazzCats* vocabulary was used in order to preserve the original data structures and to make opportunities for future edits to the model easy to identify.

3.2 Linking to external authorities

Where possible, instance-level data was used to connect to external authority files, predominantly through the assignment of VIAF[8] URIs as unique identifiers for people, of Wikidata[9] URIs for locations, and of DBpedia URIs for a range of data instances. In this regard, the *J-DISC* RDF parallels the earlier, existing *JazzCats* RDF, which is similarly linked to these authorities (Figure 5).

These links to the external authority files and instance-level alignments internal to *JazzCats* were created using `skos:closeMatch` (Figure 6). A conscious decision was made to avoid `owl:sameAs`, as this was considered to be too strict logically - although we refer to the same entities, the contexts are different, and not all the properties of one URI are necessarily true of the other [4].

3.3 Workflow

Since the *J-DISC* data is available as both relational and tabular data, there were two possible and relevant workflows for the RDF production which already formed part of *JazzCats*. We opted to use the more time-consuming and labour-intensive method of utilizing Web-Karma with tabular data, as it allowed us to produce accurate triples with little post hoc editing (Figure 7). The workflow, developed for metadata of concert ephemera from the 19th and 20th centuries [7, 8], was repeated for each of the 19 CSV files (Figure 8).

Each of the separate clusters of RDF were designed to contain at least one entity type which it shared with at least one other, enabling instance-level alignment between them.

All the *JazzCats* RDF can be accessed in two ways from the project website: either through the Pubby user-interface, facilitating the 'follow-your-nose' method for information discovery, or through the project SPARQL endpoint.[10]

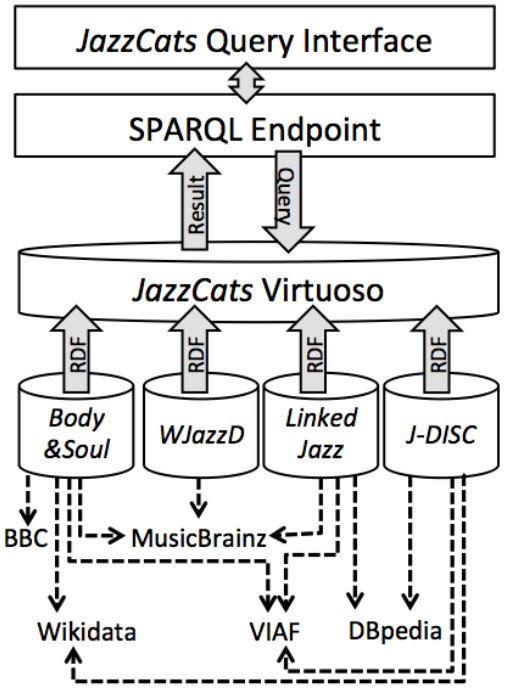

Figure 5: *JazzCats* system architecture illustrating links to external authority files

```
<http://jazzcats.oerc.ox.ac.uk/data/person/Roy_Eldridge>
skos:closeMatch <http://dbpedia.org/resource/Roy_Eldridge> ;
skos:closeMatch <http://viaf.org/viaf/94211928> .
```

Figure 6: Instance-level aligment between a *JazzCats* URI and the DBpedia and VIAF external authorities

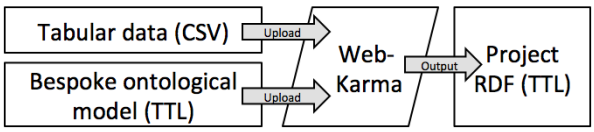

Figure 7: Web-Karma workflow

[5]http://www.music-ir.org/mirex/wiki/2016:GC16UK:JDISC
[6]http://www.musicir.org/mirex/gc15ux_jdisc/jdisc_data.sql.zip
[7]http://www.musicir.org/mirex/gc15ux_jdisc/jdisc_csv.zip
[8]https://viaf.org/
[9]https://www.wikidata.org/wiki/Wikidata:Main_Page
[10]https://jazzcats.oerc.ox.ac.uk/sparql

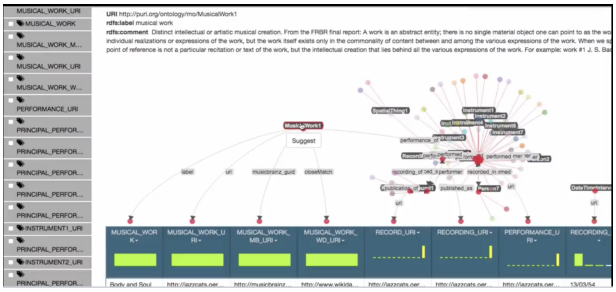

Figure 8: Web-Karma user-interface illustrating the process of mapping *JazzCats* data to bespoke ontological structures

```
SELECT DISTINCT ?vocalist ?bassist
WHERE {
?artist a foaf:Person ;
        skos:closeMatch ?another_ID ;
        rdfs:label ?vocalist ;
        mo:primary_instrument jci:vocals ;
        foafterms:gender "Female" ;
        mo:performed ?performance .

?performance a mo:Performance ;
        mo:performer ?performer ;
        mo:produced_sound ?sound.

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

?recording a mo:Recording;
       event:place <http://www.wikidata.org/wiki/Q1490> .

?performer a foaf:Person ;
        mo:primary_instrument jci:bass ;
        rdfs:label ?bassist .

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

Figure 9: SPARQL query identifying instances of a female vocalist and a bassist of unspecified gender who performed together in Tokyo, and have played with Roy Eldridge at some point in their careers. Answering it necessitates data from all four datasets (Ella Fitzgerald and Ray Brown).

Bringing in the additional *J-DISC* data (e.g. `foaf:gender`) has allowed for new and more specific SPARQL queries. An example of this is a query concentrating on female vocalists (Figure 9).

4 NEW DATA BRIDGES, NEW SCHOLARLY OPPORTUNITIES

Jazz metadata is a rich and multifaceted source of information. The capturing of the practice of jazz performance does present a number of challenges, due to notions of fluidity, the significance of improvisation, and the way in which various performers can and do contribute to different versions of a piece depending on their role in that specific session or performance. The notion of the role a person plays become particularly significant when the same musicians frequently collaborate in different contexts.

Linked Data can provide a solution for the complex dilemma of attaching various different, social and even spatio-temporal roles to an individual, representing their varied and dynamic contributions to any one given session or performance. It enables us to open up the data, and to more accurately and richly capture the diversity of skills rather than pigeonholing musicians into pre-determined information categories. Not only does this allow for a more truthful representation, but it supports a more diverse type of scholarship, empowering musicologists to pursue more diverse but also more specific research agendas.

JazzCats has successfully increased the openness and discoverability of information related to various different aspects of jazz music recordings, performances, and the musicians themselves by bridging four independent datasets through RDF. Connections which previously could only have been discovered through the consultation of each dataset separately can now be found through SPARQL queries in a single point of access. Bridging J-DISC data with the existing *JazzCats* triples has further enhanced the project, increasing its scope, and contributing to the creation of an online resource that has the potential to support and diversify future musicological research and investigation.

Future development for *JazzCats*. include enriching the currently exclusively metadata-based project with links to external digital resources such as audio samples and historical photography, such as the William P. Gottlieb Collection, Library of Congress. Another user-interface, which will allow users to access the aggregated information through means other than querying the SPARQL endpoint will also be developed.

REFERENCES

[1] 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 *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*. ACM, 1–8.

[2] Jose Bowen. 2013. Body and Soul discography. *http://josebowen.com/body-and-soul/* (2013).

[3] Richard Cyganiak and C Bizer. 2012. D2RQ-Accessing relational databases as virtual RDF graphs. *http://d2rq.org/d2r-server* (2012).

[4] Harry Halpin and Patrick J Hayes. 2010. When owl: sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web. *LDOW* 628 (2010).

[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*. ACM, 41–44.

[6] Terhi Nurmikko-Fuller, Daniel Bangert, and Alfie Abdul-Rahman. 2017. All the Things You Are:Accessing An Enriched Musicological Prosopography Through JazzCats *(Digital Humanities 2017)*. Montreal, Canada. https://dh2017.adho.org/abstracts/305/305.pdf

[7] Terhi Nurmikko-Fuller, Daniel Bangert, Alan Dix, David M. Weigl, and Kevin R. Page. 2018. Die Erstellung prototypischer Anwendungen von verknüpften musikwissenschaftlichen Datensätzen: Building Prototypes Aggregating Musicological Datasets on the Semantic Web. *Bibliothek Forschung und Praxis*. (2018).

[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*. ACM, 17–24.

[9] 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).

[10] Martin Pfleiderer and Klaus Frieler. 2010. The Jazzomat project. Issues and methods for the automatic analysis of jazz improvisations. *Concepts, experiments, and fieldwork: Studies in systematic musicology and ethnomusicology* (2010), 279–295.

[11] Yves Raimond, Samer A Abdallah, Mark B Sandler, and Frederick Giasson. 2007. The Music Ontology. In *ISMIR*, Vol. 2007. Citeseer, 8th.