All the Things You Are: Accessing An Enriched Musicological Prosopography Through *JazzCats*

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

Daniel Bangert
d.bangert@unsw.edu.au
UNSW Australia, Sydney, Australia

Alfie Abdul-Rahman
alfie.abdulrahman@oerc.ox.ac.uk
University of Oxford, United Kingdom

Introduction

JazzCats (Jazz Collection of Aggregated Triples) is a prototype project which uses Linked Open Data (LOD) to support musicological, historical, and prosopographical analyses. It has increased access to (and the openness of) data published online through a twofold process: firstly, information hitherto unavailable to users has been shared and incorporated into the project, and secondly, data previously locked in non-Open types (e.g. PDF) has been published in a machine-readable format, increasing discoverability in the context of the wider Web. Connections between datasets that could only be identified through a human user engaging separately with each existing project have now been made explicit, and the resulting aggregated data is queryable from a single user-interface (UI).

Three projects contribute to *JazzCats*: a social network connecting musicians through various types of relationships is provided by LinkedJazz (Pattuelli, 2016); details of solos within performances (including pitch, key, and chord changes) are available from WJazzD (Pfleiderer et al., 2016); and Body&Soul (Bowen, 2013) is a discography of over 200 recordings. These complementary data contain instance-level overlaps for recordings and musicians. Bringing these resources together has enabled a new type of research question, possible only through using criteria from one dataset to inform and hone results from another.

Limitations of existing data publication methods

The sub-projects at the heart of *JazzCats* engage with the 5 Star Standard of LOD (Berners-Lee, 2006) to different extents (see Table 1). The data for Body&Soul has been published online as a PDF, making it an ideal example of 1 Star categorization. WJazzD allows users to download both the database and the software (2 Star). LinkedJazz provides two separate data-dumps of RDF (with an additional, earlier set of triples also available), containing both dereferenced URIs and those which point to human-readable pages. We have categorized this project as 5 Star, because DBpedia resource URIs (related through **owl:sameAs** relationships to LinkedJazz resource URIs) are used (even if the retrieving of additional data from external sources is currently not possible via the LinkedJazz SPARQL endpoint). Publishing the information from the first two datasets with distinct HTTP URIs, connecting them to each other as well as to the RDF acquired from LinkedJazz, makes *JazzCats* 5 Star standard.

Conversion to Linked Data (LD) does not automatically ensure that information is more reusable or discoverable by data consumers on the Web (Bechhofer et al., 2013; Janowicz et al., 2014). Closed systems can benefit from LD, and whilst adherence to the LOD paradigm is an essential criterion for enabling reuse of any project's RDF by other data publishers, effective queries by a wider base of users can be restricted by idiosyncratic or project-unique vocabularies. To encourage good practices, Janowicz et al. (2014) have introduced a 5 Star rating, ranging from LD without vocabulary use (0 Star) to a vocabulary that is linked to by other vocabularies (5 Star). The term vocabulary is used in a broad sense to include all types such as schemata, and ontologies. We have categorized LinkedJazz as 5 Star because it links to other vocabularies and metadata about the vocabulary is available.

JazzCats makes extensive use of properties and classes from other vocabularies, including the Music Ontology (MO) (Raimond et al., 2007), the Event Ontology (Raimond and Abdallah, 2007), FOAF (Brickley and Miller, 2014), and SKOS (Miles et al., 2005). It is currently classified as a 5 Star since metadata about the *JazzCats* ontology is available in a

dereferenceable and machine-readable form, but other vocabularies do not yet link to *JazzCats.*

Project	Licence	Data Type	LOD Star	LD Vocab Star
Body&Soul	No licence	PDF	★	N/A
WJazzD	Open Database Licence	Structured data	★★	N/A
LinkedJazz	No licence	RDF	★★★★★	★★★★★
JazzCats	Open Database Licence	RDF	★★★★★	★★★★★

Table 1: Evaluation of the JazzCats composite projects

Increasing accessibility through *JazzCats*

A previously unpublished CSV containing Body&Soul data was cleaned and enriched with additional information held in PDF files using OpenRefine (2013) to create a new, open access dataset (Bangert, 2016). An existing workflow (Nurmikko-Fuller et al., 2016) was then reproduced to map this tabular data into RDF using an Open-Source data integration tool (Web-Karma). This workflow relied on domain expert user-input to complete the ontological modeling and instance mapping stages within Web-Karma (University of Southern California, 2016). To support the future alignment and enrichment of this data with other musicological datasets, the underlying ontological structure extensively utilizes the properties and classes of the MO. The data structure has been documented on the website of the *JazzCats* project (Bangert et al., 2016).

Both the data and the software for WJazzD are available for download from the Jazzomat Research Project website. Although structured data, and in adherence with the 2 Star LOD publication criteria, information in this form is not accessible for machine-inferencing, and the clustered tables can be difficult for human users to navigate. The data was converted to RDF by repeating a second workflow as described in Nurmikko-Fuller et al. (2016) using the D2R server (Cyganiak and Bizer, 2012). This automated process produced clusters of triples based on database information categories (e.g. melody, beats, sections), which are mostly expressed through **xsd:strings** and **xsd:integers**. Mappings were made where applicable to connect these elements together using MO properties and classes. An overview of the ontological structure, and a detailed subsection illustrating the different properties are documented and defined on the *JazzCats* website.

LinkedJazz provides two separate datasets for entities and the 12 different types of interpersonal (both professional and social) relationships between them. Adding these RDF-dumps to the *JazzCats* triplestore enables queries combining this prosopography with performance metadata derived from the other projects which make up the entirety of the *JazzCats* data.

Publishing these datasets as RDF using common vocabularies and ontologies known to have been utilized in other digital musicology projects increases their discoverability and the value. As data publishers, adhering to LOD standards allows us to further benefit from any additional future linkage. A conscious decision has been made at the onset of the *JazzCats* development process to publish RDF, ontologies, and raw data in Open and accessible formats, with appropriate licensing, to allow for the replication of our workflow, verification of our findings, and reuse of any or all of the composite parts of the project. *JazzCats* is made available under the Open Data Commons Open Database License (Miller et al., 2008). The aim throughout has been to produce and publish data that adheres to the FAIR data principles of being findable, accessible, interoperable and reusable (Wilkinson et al., 2016).

Evaluation of *JazzCats*

Although the data in *JazzCats* adheres to a 5 Star standard of LOD for accessibility and openness, the current UI presents a barrier to access for human users. At present, the project RDF (RDF Core Working Group, 2014) is contained within an instance of the Open-Source version of the Virtuoso (OpenLink Software, 2016) triplestore, and is only queryable using the default SPARQL endpoint, accessible from the *JazzCats* website. Example queries demonstrate how results can be generated by drawing from all three datasets simultaneously (Nurmikko-Fuller, 2016), but engagement with the data beyond the parameters set by these existing samples requires the user to have the necessary skills to construct new SPARQL queries (W3C RDF Data Access Working Group, 2008). Aware of the dichotomy of skills between music scholars and the ability to formulate such queries, the authors acknowledge that at present, the site UI presents notable barriers to access.

To address this, we have provided extensive documentation of the underlying ontological structures on the project website. Each of the sub-projects is illustrated with diagrams that include an interactive feature which provides the scope notes for a given class when hovered over with the cursor (see Figures 1-3). The diagrams also show the inherent connectivity of the graphs within *JazzCats*, and the directionality of properties (arrows running from domain to range). The combination of easy access to

appropriate documentation for the data, the underlying ontological structure, and examples of functioning queries enables specialists to access *JazzCats* data according to their research interests. In addition, a Pubby interface (Cyganiak and Bizer, 2011) serves as an alternative Linked Data front end that enables users to navigate through the triples without the need to use SPARQL. Influenced by ongoing work at ResearchSpace (Oldman et al., 2016), planned future developments of *JazzCats* include the development of a GUI that will allow users to generate ontologically valid queries using dropdown lists generated by available properties for each class. This step will further help open *JazzCats* for experts and scholars along the full length of the digital humanities spectrum.

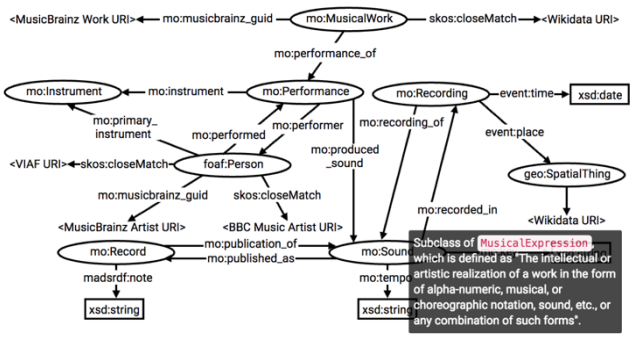

Figure 1. Body&Soul data structure illustrating pop-up scope notes for the class mo:Sound

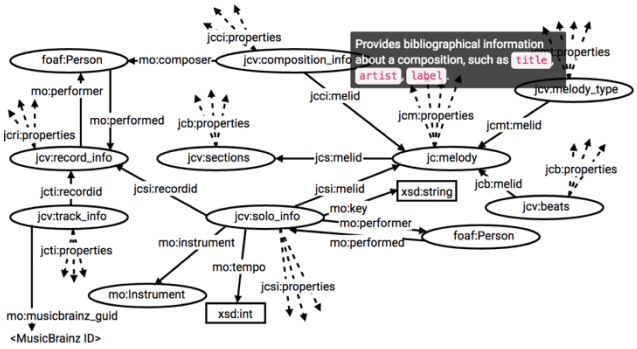

Figure 2. WJazzD data structure illustrating pop-up scope notes for the class jcv:composition_info

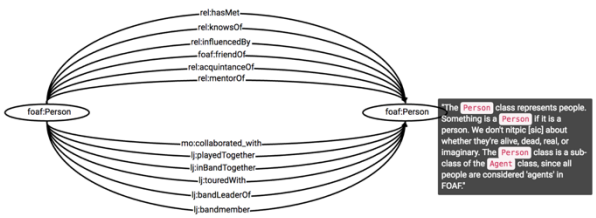

Figure 3. LinkedJazz data structure illustrating pop-up scope notes for the class foaf:Person

Bibliography

Bangert, D. (2016). JazzCats Body and Soul discography. Zenodo. Dataset. http://doi.org/10.5281/zenodo.163886

Bangert, D., Nurmikko-Fuller, T. and Abdul-Rahman, A. (2016). *JazzCats.* Available at https://jazzcats.oerc.ox.ac.uk. Date last accessed: 31 Oct 2016.

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., ... & Gamble, M. (2013). "Why linked data is not enough for scientists." *Future Generation Computer Systems*, 29(2): 599-611.

Berners-Lee, T. (2006). Linked Data. Available at https://www.w3.org/DesignIssues/LinkedData.html. Date last accessed: 31 Oct 2016.

Bowen, J. (2013). Body and Soul discography. Available at http://josebowen.com/body-and-soul/. Date last accessed: 31 Oct 2016.

Brickley, D. and Miller, L. (2014). FOAF Vocabulary Specification 0.99. Namespace Document 14 January 2014. Paddington Edition.

Cyganiak, R. and Bizer, C. (2011). Pubby: A Linked Data Frontend for SPARQL Endpoints. Available at http://www4.wiwiss.fu-berlin.de/pubby/. Date last accessed: 6 April 2017.

Cyganiak, R. and Bizer, C. (2012). D2RQ: Accessing Relational Databases as Virtual RDF Graphs. Available at http://d2rq.org/d2r-server. Date last accessed: 31 Oct 2016.

Janowicz, K., Hitzler, P., Adams, B., Kolas, D. and Vardeman, I. (2014). "Five stars of linked data vocabulary use." *Semantic Web,* 5(3): 173-76.

Miles, A., Matthews, B., Wilson, M. and Brickley, D. (2005). "SKOS Core: Simple Knowledge Organisation for the Web." *Proceedings of the International Conference on Dublin Core and Metadata Applications,* pp. 3-10.

Miller, P., Styles, R. and Heath, T. (2008). "Open Data Commons, a License for Open Data." *Linked Data on the Web*, 369.

Nurmikko-Fuller, T. (2016). SPARQL_queries_JazzCats. Zenodo. Dataset. http://doi.org/10.5281/zenodo.163879

Nurmikko-Fuller, T., Dix, A., Weigl, D.M. and Page, K.R. (2016). "In Collaboration with In Concert: Reflecting a Digital Library as Linked Data for Performance Ephemera." Proceedings of the 3rd International workshop on Digital Libraries for Musicology, pp. 17-24.

Oldman, D., Anagnostopoulou, M., Eales, G., Kelly, M. and Rychlik, A. (2016). ResearchSpace. British Museum. Available at http://www.researchspace.org. Date last accessed: 31 Oct 2016.

OpenLink Software, (2016). Virtuoso Universal Server. Available at https://virtuoso.openlinksw.com/. Date last accessed: 31 Oct 2016.

OpenRefine, (2013). OpenRefine 2.6 beta 1. Available at http://openrefine.org/download.html. Date last accessed: 31 Oct 2016.

Pattuelli, C. (2016). LinkedJazz Project. Available at https://linkedjazz.org/. Date last accessed: 31 Oct 2016.

Pfleiderer, M., Frieler, K., Abeßer, J., Zaddach, W-G., Burkhart, B. and Bartel, F. (2016). The Jazzomat Research Project, Doc v1.0. Available at http://jazzomat.hfm-weimar.de/dbformat/dboverview.html. Date last accessed: 31 Oct 2016.

Raimond, Y. and Abdallah, S. (2007). The Event Ontology. Technical report, Citeseer.

Raimond, Y., Abdallah, S. A., Sandler, M. B. and Giasson, F. (2007). "The Music Ontology." *Proceedings of the 8th International Conference on Music Information Retrieva*l, pp. 417-422.

RDF Core Working Group, (2014). RDF: Resource Description Framework. Available at https://www.w3.org/RDF/. Date last accessed: 31 Oct 2016.

University of Southern California, (2016). Karma: A Data Integraiton Tool. Available at http://usc-isi-i2.github.io/karma/. Date last accessed: 7 April 2017.

W3C RDF Data Access Working Group, (2008). SPARQL Query Language for RDF. Available at https://www.w3.org/TR/rdf-sparql-query/. Date last accessed: 31 Oct 2016.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... and Bouwman, J. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data,* 3.