**PATTUELLI — Accidental**

<1 figure>

**Accidental Discovery, Intentional Inquiry: Leveraging Linked Data to Uncover the Women of Jazz**

Pattuelli, M. C., Miller, M. and Hwang, K.

*We only see what we have interest in.*

—Paul Nougé, *Birth of the Object (The Subversion of Images)*, 1929–1930

In this paper we discuss the heuristic capabilities that the process of creating cultural heritage linked data may afford and the potential of integrating linked open datasets to facilitate and augment arts and humanities research. More specifically, we address the paths to a linked data-driven investigation in the history of women in jazz. The research context is provided by Linked Jazz (linkedjazz.org), an evolving project that experiments with applying linked open data principles and techniques to cultural heritage materials. The project currently uses oral histories from digital archives of jazz history as the main source of named entities to be represented as linked open data. Interview transcripts are processed through open-source applications developed in-house that employ automated and manual methods for name extraction, matching, and quality control. A linked open dataset has been created that includes over 9,000 proper names of musicians as well as the personal and social relationships connecting them. The dataset supports the creation and visualization of interactive social graphs for representing the highly interconnected community of jazz musicians (Pattuelli et al., 2013).

While the Linked Jazz tools and data services are intended to support data discovery and analysis, the development process itself has proven conducive to research inquiry by revealing unanticipated paths to discovery and engagement with heritage data. One of the initial steps in producing the dataset for Linked Jazz involved creating a directory of proper names to be leveraged in automated text analysis and extraction activities. To this end, a domain-specific vocabulary of personal names was built by extracting and filtering data from the US version of DBpedia. In line with linked open data best practices, existing URIs were reused whenever possible. As an RDF export of Wikipedia data, DBpedia is the largest and most popular hub of linked open data and thus encompasses an extensive array of jazz artist entries, including less prominent figures. For further refinement and consolidation, the names were mapped to VIAF, the international aggregation service of bibliographic name authorities including LC/NAF, the Library of Congress Name Authority File. While VIAF is widely used to disambiguate proper names and enrich the vocabulary with their variant forms, including nicknames and stage names, it contains only name instances derived from bibliographic records, a smaller subset of the pool that can be found in DBpedia. Through automated analysis of interview transcripts, names of musicians mentioned in the text were identified and matched to corresponding names in our directory.Instances occurred when names were recognized in the transcripts but did not have a corresponding match in our inventory. This was mostly due to the lack of an entry in Wikipedia, the source of DBpedia data. When needed, alternative linked open data sources were manually searched, including the music-specific encyclopedia MusicBrainz (https://musicbrainz.org).

When names were not found in any of the available sources, URIs had to be minted using the Linked Jazz namespace, e.g., http://linkedjazz.org/resource/Lynn\_Grissett.

Based on the random sample of 54 interview transcripts we have processed to date, a total of 219 name instances were newly coined. The tallying of minted URIs showed that 25 out of 219 minted URIs referred to women, and 18 of those women were involved in the jazz scene in various capacities, from less prominent jazz musicians to people otherwise involved in the music industry (producers, managers, educators, etc.). While these individuals were mentioned in our collection of oral histories, they were missing from major encyclopedic knowledge bases, such as Wikipedia and MusicBrainz, as well as from massive repositories of bibliographic name authorities such as LC/NAF and VIAF.

The bare data encoding, which is provided by the URI syntax, revealed aspects of the data worthy of further investigation. An obvious finding was that women make up only a tiny percentage of the entities with new identifiers (11%). Even a cursory glance at the entire Linked Jazz dataset shows a significant discrepancy in the ratio between female and male jazz musicians. The missing URIs confirm this trend, but also suggests that the morphology of linked data can serve as a fertile clue and provide threads to new lines of inquiry through the lens of linked open data.

These accidental findings have triggered a new stream of development strategies aimed at supporting research in gender representation, an understudied area of jazz (Placksin, 1982; Tucker, 2000). By looking for patterns of the missing data, performing cross-references against interview mentions and conducting data analysis through different data facets including gender, roles, or professional relevance, we expect to be able to expand the array of questions that can be asked of the data.

We are currently working to associate additional properties, starting with gender and instrument(s) played, to the person entities in the Linked Jazz dataset. This is achieved through data mashups with suitable sets of data from external sources. A key tenet of linked open data development, the integration of heterogeneous datasets is enabled by the design of RDF, a unifying data framework that relies on common modeling constructs such as URIs (Hendler, 2011). Data integration is the strategy that our project has adopted to add new layers of semantics to our set of triples and thus offer new opportunities for analysis and discovery.A dedicated RDF ontology is being developed to harmonize heterogeneous semantics and ease the process of integrating data. The ontology is created by expanding the set of predicates from the Linked Jazz dataset with predicates from selected data sources, resulting in a cohesive and expressive conceptual model. The data enrichment offered by mashing up datasets will bring together a wide range of information from temporal and spatial data (e.g., time periods, dates, events, geographic locations, etc.) to music-specific data (e.g., professional roles, instruments, recordings, music venues, etc.), opening up unanticipated strains of inquiry. A visual representation of a data integration use case is shown in Figure 1. It is based on a test of mashups centered on Ella Fitzgerald.

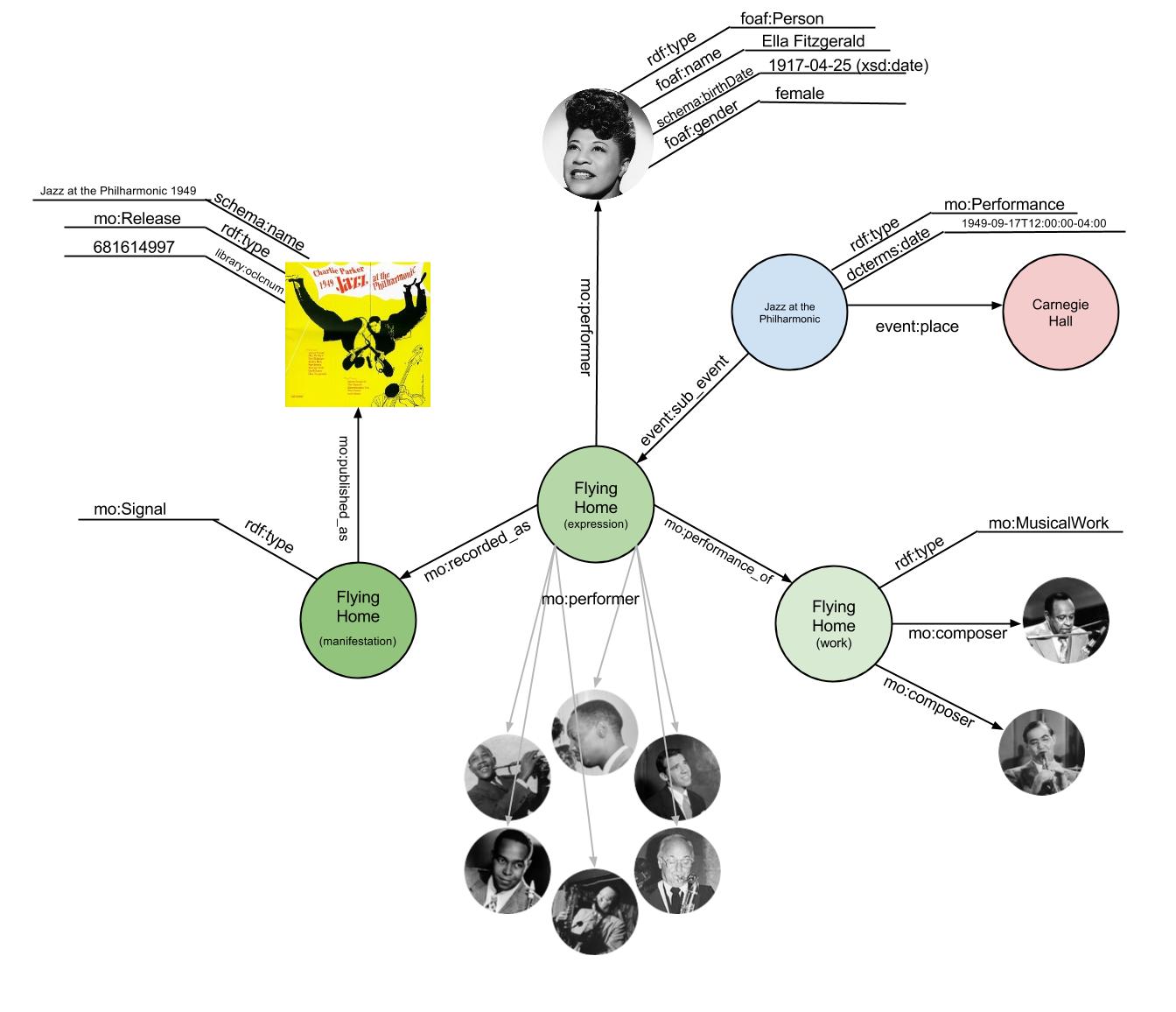


Figure 1. Data integration use case.

Three domain-specific databases will be leveraged to perform mashups: (1) Columbia University’s J-DISC,1 (2) Steve Albin’s BRIAN,2 and (3) Carnegie Hall’s performance archive database.3 The first two are open-source jazz discography databases4 that provide granular information including session details with songs, tracks, dates, locations, and artist contributions from the composer to the sideman. The third contains rich and detailed information including types of performer (e.g., singer, instrumentalist, conductor, etc.).

As a condicio sine qua non, however, we have to associate the attribute *gender* to the jazz musician names throughout the Linked Jazz dataset. Unfortunately, this won’t be a straightforward integration task as gender information, limited to binary *male* and *female*, is only recorded in J-DISC, which includes fewer than half of the artists present in BRIAN (and only 313 female jazz musicians). To supplement J-DISC data, we would need to make use of Wikipedia data via DBpedia. Specifically, we will leverage Wikipedia categories *women* and *female* as they are the only gender-related qualifiers applied to the jazz domain. The symmetric categories *men* and *male* are not specified, but assumed.

The heuristic capability inherent in the syntax of linked data and the data enrichment derived from knitting together linked open datasets have the potential to shed light on underrepresented contributors to the history of jazz and offer new perspectives to arts and humanities inquiry.

**Notes**

1. http://jdisc.columbia.edu.

2. http://www.jazzdiscography.com/Brian/.

3. http://www.carnegiehall.org/PerformanceHistorySearch/#!.

4. More specifically, Steve Albin’s BRIAN is an application that allows users to create and publish their own discographies.

**References**

**Hendler, J.** (2011). The Semantic Web from the Bottom Up. In Batcherer, T. and Coover, R. (eds), *Switching Codes: Thinking through Digital Technology in the Humanities and the Arts*. Chicago: University of Chicago Press, pp. 125–39.

**Pattuelli, M. C., Miller, M., Lange, L., Fitzell, S. and Li-Madeo, C.** (2013). Crafting Linked Open Data for Cultural Heritage: Mapping and Curation Tools for the Linked Jazz Project. *Code4Lib Journal,* **21**, http://journal.code4lib.org/articles/8670.

**Placksin, S.** (1982). *American Women in Jazz: 1900 to the Present—Their Words, Lives, and Music*. Wideview Books, New York.

**Tucker, S.** (2000). *Swing Shift: ‘All-Girl’ Bands of the 1940s*. Duke University Press, Durham, NC.