Accidental discovery, intentional inquiry: Leveraging linked data to uncover the women of jazz

M. Cristina Pattuelli and Karen Hwang
Pratt Institute, School of Information, USA

Matthew Miller
New York Public Library, USA

Abstract

In this article we discuss the heuristic capabilities that the process of generating, processing, and integrating cultural heritage linked data may afford, including its potential for enhancing arts and humanities research. More specifically, we report on our current work on detecting and assigning gender properties to person entities and semantically enriching a set of Linked Open Data (LOD) in the domain of history of jazz. Linked Jazz—an ongoing project that experiments with the application of LOD principles and techniques to cultural heritage materials—provided the context for this research. Linked Jazz aims to uncover meaningful connections between data and documents from digital archives of jazz history. It employs oral histories as the main source of named entities to be represented as linked data. The entities are then semantically connected and visualized as social graphs. Using the assignment of gender properties, this article describes how the data development process itself offers new and unanticipated paths of research inquiry and engagement with heritage data.

Correspondence:
M. Cristina Pattuelli
Pratt Institute School of Information,
144 West 14th Street
New York, NY 10011.
E-mail: mpattuel@pratt.edu

1 Introduction and Background

The process of deriving linked data from text is still a rather novel approach in the context of linked data development. The capability to identify and represent connections between entities derived from unstructured text has enormous potential for digital humanities research, as it not only allows deep analysis of text, but also, and most distinctively, it can reveal hidden connections between data and documents from diverse and heterogeneous sources. In other words, access to integrated views of otherwise dispersed information offers approaches to discovery that might otherwise be impossible, and new methods to engage with cultural heritage information.

Linked Jazz[1] is an ongoing project that experiments with applying linked open data to cultural heritage material, with a focus on creating network data from oral histories found in digital jazz archives. As such, the Linked Jazz project provides an ideal context in which to discuss challenges posed and opportunities offered by the process of creating, processing, and integrating cultural heritage linked data. Over 50 interview transcripts have been processed to date employing a combination of automated and manual methods for named entity extraction, data enrichment, and curation.[2] The resulting linked open dataset represents the proper names of artists as well as the personal and social relationships connecting them.[3] This dataset also serves as the source for an interactive visualization

Digital Scholarship in the Humanities, Vol. 32, No. 4, 2017.
doi:10.1093/llc/fqw047 Advance Access published on 10 October 2016

of social graphs[4] developed to represent the highly interconnected community of jazz musicians.

As a preliminary step to performing named entity recognition, a directory of proper names of jazz artists was created to support automated text analysis and extraction. Such a domain-specific vocabulary did not exist and had to be built from scratch. Name authorities, including Virtual International Authority File (VIAF)[5] and the Library of Congress Name Authority File (LC/NAF),[6] proved valuable sources of name instances available in Linked Open Data (LOD) format. While massive in size, they are nevertheless limited in their coverage, as they only include entities derived from bibliographic records. Our name dictionary of over 9,000 proper names of jazz artists was generated by extracting and filtering data from the US version of DBpedia.[7] As a Resource Description Framework (RDF) export of Wikipedia data, DBpedia is currently the largest and most widely used source of LOD. It encompasses a significant number of jazz artist entries, including less prominent figures who are not part of any library or archival catalog. Bibliographic authorities were later leveraged for name disambiguation and reconciliation. We relied on open-source tools developed in-house to perform identity management.[8]

During the automated analysis of interview transcripts, names of musicians mentioned in the text are identified and matched to corresponding names in our directory. Instances occur when names are recognized in the text, but do not have a corresponding match in our list. This is typically due to the absence of an entry for that name in Wikipedia, the source of DBpedia data. When needed, alternative LOD sources are manually searched, including the popular music-specific encyclopedia MusicBrainz.[9] If names are not found in any of the sources consulted, but a reliable citation can be found online, Linked Jazz mints the entity into the Linked Jazz namespace (e.g. http://linkedjazz.org/resource/Lynn_Grissett).

2 Empirical analysis

Based on the random sample of 54 interview transcripts that have been processed to date, a total of 219 Uniform Resource Identifiers (URIs) were newly minted for names that could not be located in external sources—10.9% of the entire pool of a total of 2,006 names. Technically, our coining of URIs is motivated by the need to make data about a person entity machine readable and linkable. This process, however, also carries the benefit of assigning a web identity to musicians who, for any number of reasons, did not previously have one in any major LOD source, including encyclopedic knowledge bases and repositories of bibliographic name authorities such as LC/NAF and VIAF.

Analysis of the minted URIs showed that only 25 out of the 219 minted URIs referred to women, of which only 18 were directly involved in the jazz community in various capacities, from less prominent jazz musicians to people otherwise active in the music industry (e.g. producers, managers, educators). Our finding that women active in the jazz community comprised only a tiny percentage of the newly minted artists (8.2%) was not unexpected given the lack of female representation in the entire Linked Jazz dataset. Even a cursory glance at the dataset revealed a significant discrepancy in the ratio between female and male jazz musicians. We took the findings that emerged from the encoding process as an opportunity to steer the development of our dataset in ways that would enable deeper data analysis from a gender perspective and possibly expose researchers to novel perspectives on gender representation—an under-studied area of jazz (Placksin, 1982; Tucker, 2000). By looking for patterns in the data, performing cross-references against interview mentions and conducting data analysis through different data facets including gender, professional roles, or artistic relevance, we expect to be able to expand the array of questions that can be asked of the data. Do female jazz artists mention other women in the context of their lives and careers more often than male jazz artists mention women? Do the interviews reveal that female instrumentalists experience the jazz world differently than female vocalists? Or are there noticeable differences in how women and men in the jazz industry speak of successful women in jazz? The broad narratives of the oral histories from which the data are derived serve as a rich wellspring where complex questions can begin to be answered (Fig. 1).

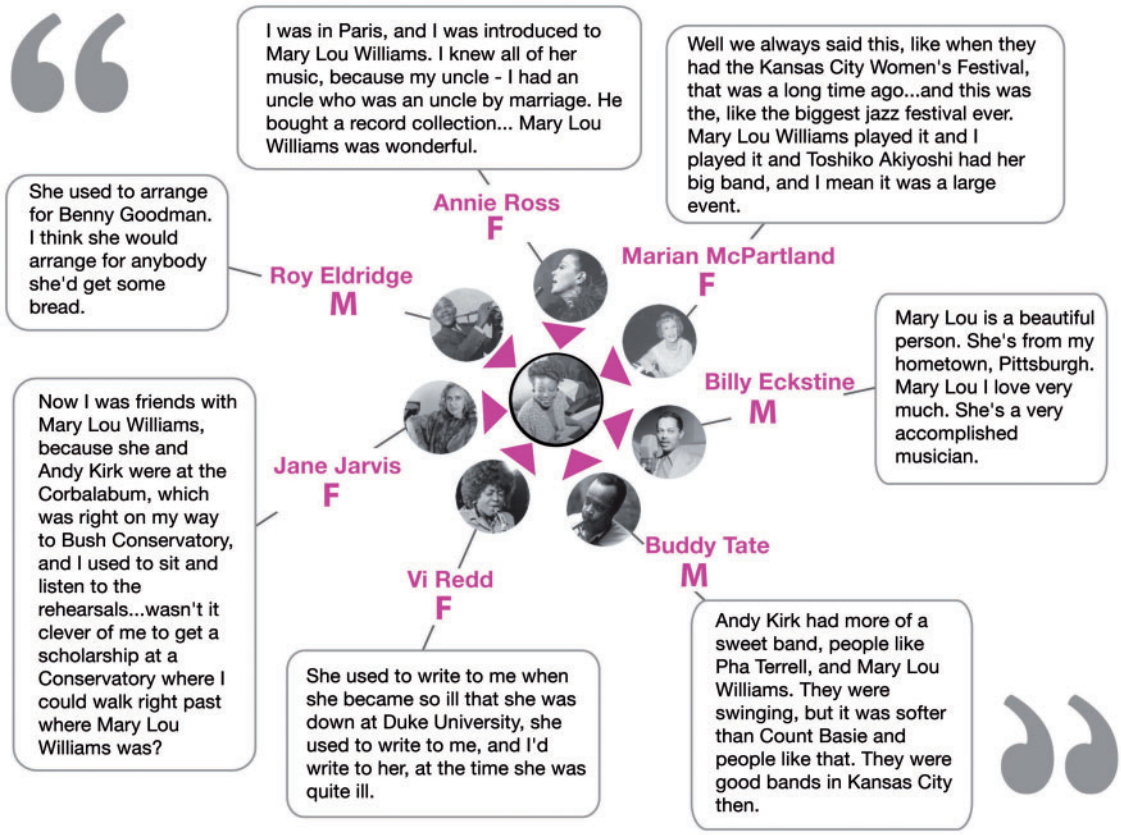

Fig. 1 Selected passages from processed oral histories mentioning pianist and composer Mary Lou Williams. Diagram illustrates possible ways to access the Linked Jazz dataset using attributes acquired

The networked structure of the linked dataset and its visualization via dynamic social graphs offer a means for deeper levels of analysis and discovery by placing data in context and enabling interaction with its multiple facets. But it is the power of the underlying LOD framework that makes it possible to ask more of our data. Because of its open and decentralized data representation model, LOD has the capacity to incorporate different perspectives on data and make them coexist through interlinking and unified discovery views.

3 Gender identification

Generally, the possibility to ask articulate and complex questions of a dataset is proportional to the expressivity of its data. In our context, the range of properties associated with a person would determine the type of analysis that the data can afford. It is axiomatic to say that the presence of the attribute gender is the condition sine qua for beginning to sort the data according to gender. This basic attribute is typically missing or inconsistently present in the data sources commonly used in linked data development.

A preliminary step to make our network data useful for analyzing sociocultural aspects of women in jazz consisted of identifying and assigning the gender attribute to each person entity in the Linked Jazz dataset. The importance of enriching bibliographic metadata with gender information—for example, for literary analysis of large-scale text corpora—has been recently discussed in the context


```
{
    "id": 197
    "name": "Alice Coltrane",
    "uri": "<http://dbpedia.org/resource/Alice_Coltrane>",
    "dbpedia_categories": [
        "http://dbpedia.org/resource/Category:American_jazz_organists",
        "http://dbpedia.org/resource/Category:Female_jazz_musicians",
        "http://dbpedia.org/resource/Category:Impulse!_Records_artists",
        "http://dbpedia.org/resource/Category:John_Coltrane",
        "http://dbpedia.org/resource/Category:Modal_jazz_pianists",
        "http://dbpedia.org/resource/Category:1937_births",
        "http://dbpedia.org/resource/Category:2007_deaths",
        "http://dbpedia.org/resource/Category:African-American_musicians",
        "http://dbpedia.org/resource/Category:American_jazz_pianists",
        "http://dbpedia.org/resource/Category:Avant-garde_jazz_musicians",
        "http://dbpedia.org/resource/Category:Cass_Technical_High_School_alumni",
        "http://dbpedia.org/resource/Category:Converts_to_Hinduism",
        "http://dbpedia.org/resource/Category:Deaths_from_respiratory_failure",
        "http://dbpedia.org/resource/Category:Disease-related_deaths_in_California",
        "http://dbpedia.org/resource/Category:Musicians_from_Detroit,_Michigan",
        "http://dbpedia.org/resource/Category:Warner_Bros._Records_artists",
        "http://dbpedia.org/resource/Category:American_jazz_harpists",
        "http://dbpedia.org/resource/Category:Followers_of_Sathya_Sai_Baba",
        "http://dbpedia.org/resource/Category:African-American_Hindus"
    ],
    "viaf_id_from_dbpedia": "5120047",
    "musicbrainz_id_from_dbpedia": "d24fc285-7b2f-47da-8ac5-4070b08570b1",
    "DBpedia_gender": "F",
    "musicbrainz_gender": "F",
    "viaf_gender": "Unknown",
},
```


Fig. 2 Output from the process of querying and parsing data for Alice Coltrane

of the HathiTrust Digital Library (Peng *et al.*, 2014). As the researchers point out, there are different methods for determining the gender of an author, including text analysis and mapping to external sources. The identification of gender through automated methods, including machine-learning techniques, is a popular approach when the names of the authors are not present in the source (Cheng *et al.*, 2011). However, when the author names are known, the practice of mapping names to authoritative name lists is typically adopted (Naldi *et al.*, 2004).

Our case is closer to the latter approach, as we are relying on a curated directory of proper names. The process we adopted consists of leveraging our name list to query resources via APIs and SPARQL endpoints to access structured data from various sources. Python scripts loop through our stored URIs to query a single resource and execute the dual task of (1) obtaining available gender data from the resource record for each entity and (2) gathering further references for the entity on other platforms, typically stored as URIs or unique identifiers in a record. This second step allows us to associate entities in our list with their exact references from other sources. Using the newly acquired identifiers, the process is iterated on selected platforms to exhaust all likely possibilities for acquiring gender data for a person. Insofar as the chain of URIs allowed, starting with our stored URIs, we have queried our list against DBpedia, MusicBrainz, and VIAF, the most suitable LOD sources in terms of scale and domain coverage at the time of this work (Fig. 2).

As discussed above, the processing of Linked Jazz oral history transcripts produced a list of 2,006 person entities, including the 219 newly minted URIs. Names minted with Linked Jazz URIs, however, were removed from the dataset before gender analysis, since the need to mint a name is predicated on the person's absence from major resources and therefore has no record to query. Another small test group of names was also removed from the list due to errors in automated processing, bringing the

target number to 1,772. With over 85% of our remaining target list stemming from DBpedia, this knowledge base was an obvious first resource for querying gender data. DBpedia records, however, contain no explicit gender property. Instead, gender qualifications are sometimes found in the subjects associated with a person. DBpedia subjects for each person were parsed to match gender-defining words, like women, female, men, and male, but less than 20% of our entire list could be enriched with a gender property using this method.

A point of note was the distribution of these properties throughout the name list. About half of the entities with an associated gender attribute were women, which was far from representative of our list of names. A more in-depth examination revealed that gender-qualifying terms are more often found for women than men in the DBpedia subject field. In other words, jazz musicians who are women are more often qualified by gender, such as *Female jazz musicians*, whereas men are merely described as *Jazz musicians* making the gender qualifier implicit and assumed by default. We also observed a recent shift in the provision of gender data in DBpedia. While not existing at the time of our first round of queries to DBpedia in late 2014, these data are now present and appear to be applied rather consistently. Systematic analysis of such changes in activity on decentralized platforms like Wikipedia, where the revision history can be openly accessed, may open up new lines of inquiry. Comparison of results from the same script run at different times, for example, might reveal patterns in revision activity and offer cues useful for analyzing the evolution of gender representation.

The other two resources queried—VIAF and MusicBrainz—include semantically defined fields for gender. In addition to designators for *male* and *female*, the first expands this option to also include the attribute *unknown*, while the latter allows the field to be left blank. The shortfall in representing the complex and fluid reality underlying the notion of identity is evident in the datasets we worked with and is a broader issue that the Digital Humanities and the Library and Information Science communities have begun to address (Billey *et al.*, 2014; Posner, 2015). Despite the limitations, having gender explicitly represented in the record makes our querying process more precise and direct compared to DBpedia. The success rate of obtaining gender information was much higher for both of these resources. Also interesting was the fact that the distribution of men to women in gender data acquisition, based on the number possible for each resource, closely reflected the gender distribution for new Linked Jazz entity mints: 10.6% women for those positively identified on VIAF and 12.4% women on MusicBrainz. Through these three resources, we were able to successfully acquire gender information for 75% of our target list.

4 Data enrichment

As the next logical step for enriching the Linked Jazz dataset, we are now working on associating additional descriptive attributes to the person entities that would enable deep and multifaceted investigation of the jazz community. This goal is achieved through data mash-ups and interlinking with suitable sets of data from external sources. The integration of heterogeneous datasets is a central tenet of LOD development. A traditionally arduous computational task, data integration is facilitated by the simple and open design of the LOD infrastructure. Key to making data easy to combine is RDF,[10] 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 RDF triples and thus offer opportunities for data analysis and discovery. A domain-specific RDF ontology is being developed to harmonize heterogeneous or inconsistent semantics, bridge data from disparate sets, and ease the overall process of integrating data (Pattuelli *et al.*, 2015). The semantic enrichment offered by mashing up datasets and interlinking the data into a unified view will bring together a wide range of information, from temporal and spatial data (e.g. time periods, dates, events, geographic locations) to music-specific data (e.g. professional roles, instruments, recordings, music venues), opening up unanticipated

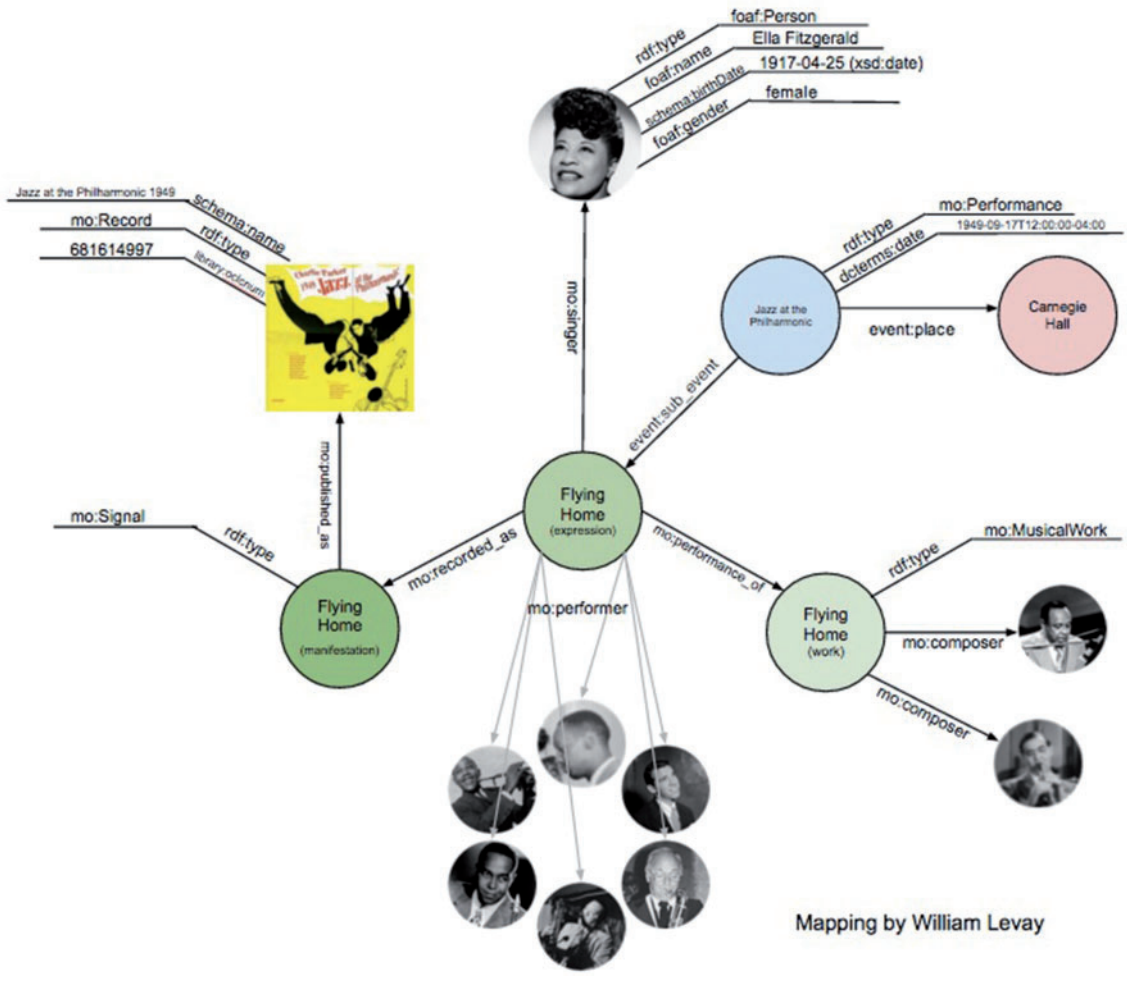

Fig. 3 Data integration use case

opportunities for scholarly inquiry. Several music-specific resources are under consideration to support data enrichment and integration including Columbia University's J-DISC,[11] Steve Albin's BRIAN,[12] and Carnegie Hall's performance archive database.[13] Because of their rich and diversified content, these data sources would enhance the description of our artists with a wide range of attributes, from types of performer and instrument played to types of contributions in recording sessions. A visual representation of a use case based on a test of mash-ups centered on Ella Fitzgerald is shown in Fig. 3.

5 Conclusion and future work

In this article we describe the process of generating and processing linked data from oral histories in the domain of jazz. At the core of the Linked Jazz project, the very creation of a linked dataset has proven heuristically fruitful to determine the next phase of project development. More specifically, the morphology and distribution of linked data, either originally coined or programmatically derived, have shown the possibilities that adding a gender perspective has to open up new lines of inquiry on

women in jazz, a long neglected area of jazz studies. We are currently working to enhance our dataset further through integration with external domain-specific sources to support deeper and multifaceted analysis. Semantic enrichment derived from knitting together these diverse yet complementary linked open datasets has the potential not only to offer new entryways for researchers to investigate the history of jazz, but to also shed light on its under-represented contributors through the unique lens of linked data.

REFERENCES

Billey, A., Drabinski, E., and Roberto, K. R. (2014). What's gender got to do with it? A critique of RDA 9.7. *Cataloging & Classification Quarterly*, 52(4): 412–21.

Cheng, N., Chandramouli, R., and Subbalakshmi, K. P. (2011). Author gender identification from text. *Digital Investigation*, 8(1): 78–88.

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 and London: University of Chicago Press, pp. 125–39.

Naldi, F., Luzi, D., Valente, A., and Parenti, V. I. (2004). Scientific and technological performance by gender. In Moed, H. F., Glänzel, W., and Schmoch, U. (eds), *Handbook of Quantitative Science and Technology Research – The Use of Publication and Patent Statistics in Studies of S&T Systems*. Dordrecht, Boston, London: Kluger Academic Publishers, pp. 299–314.

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 (accessed 24 September 2015).

Pattuelli, M. C., Provo, A., and Thorson, H. (2015). Ontology building for linked open data: a pragmatic perspective. *Journal of Library Metadata*, 15(3–4): 265–94.

Peng, Z., Chen, M., Kowalczyk, S., and Plate, B. (2014). Author Gender Metadata Augmentation of HathiTrust Digital Library. In *Proceedings of the American Society for Information Science and Technology*, Seattle, WA, November 2014.

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

Posner, M. (2015). What's next: the radical, unrealized potential of digital humanities, Miriam Posner's blog, 27 July 2015. http://miriamposner.com/blog/ (accessed 5 February 2016).

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

Notes

1 https://linkedjazz.org/
2 A detailed description of the tools and techniques that support the tasks can be found in Pattuelli, Miller, Lange, Fitzell and Li-Madeo (2013).
3 The data can be queried via a SPARQL endpoint (https://linkedjazz.org/sparql/) and also accessed via an API (https://linkedjazz.org/api/).
4 https://linkedjazz.org/network/
5 https://viaf.org
6 http://id.loc.gov/authorities/names.html
7 http://wiki.dbpedia.org/
8 A list of Linked Jazz tools, including the Name Mapping Tool and Curator Tool and Ecco! for named entity resolution, is available at https://linkedjazz.org/tools/
9 https://musicbrainz.org
10 http://www.w3.org/RDF/
11 http://jdisc.columbia.edu
12 http://www.jazzdiscography.com/Brian/
13 http://www.carnegiehall.org/PerformanceHistorySearch/#!