DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2011

Linked Jazz: An Exploratory Pilot

M. Cristina Pattuelli
Pratt Institute, School of Information and Library Science, USA
mpattuel@pratt.edu

Chris Weller
Pratt Institute, School of Information and Library Science, USA
chris@chrisweller.com

Genevieve Szablya
Pratt Institute, School of Information and Library Science, USA
gszablya@pratt.edu

Abstract

This paper reports on a pilot conducted within Linked Jazz, a project that investigates the potential of Linked Open Data (LOD) technology to enhance discovery and visibility of digital cultural heritage materials. The project explores the applicability of the Friend-Of-A-Friend (FOAF) ontology to digital archives of jazz history to expose relationships among musicians and reveal their community's network. Finding innovative ways of connecting cultural data and making them searchable in an open discovery environment generates unprecedented opportunities to create new meaning and elicit new streams of interpretation. The project consists of multiple phases and is intended to progress in an iterative and experimental way. The first step was to pilot a method to create a dataset of RDF triples representing jazz artists and their social connections.

Keywords: linked data; cultural heritage; digital archives; social network; relationships.

1. Introduction

Given the massive amount of cultural heritage data that is already available in digital repositories, and the rate at which memory institutions continue to digitize their holdings, the need for systems that enhance discovery and facilitate analysis of digital cultural objects is greater than ever. One of the challenges facing scholars is to identify the network of relationships that exist among the individuals described in resources often hidden behind repositories' invisible walls.

The application of LOD technology to cultural heritage data and metadata is a promising strategy to address these problems**.** Linked data is a recommended best practice for connecting distributed data across the web to facilitate interoperability (Heath & Bizer, 2011). Memory institutions including archives are natural candidates for contributing to the open-world paradigm of linked data. By linking their cultural data in meaningful ways they have the potential to extend the reach of their own collections beyond existing controlled environments, facilitating further discovery and interpretation. In the broader context of the web, the linked data environment would provide a unifying publishing framework for discovery, integration and reuse of cultural heritage data.

While the library community has been involved in the linked data movement since its inception, research focused on the cultural heritage domain, archives in particular, is still scarce. The Linked Jazz project investigates the applicability of one of the most popular linked data technologies, the Friend-Of-A-Friend (FOAF) ontology[1], to digital archives. FOAF is an RDF vocabulary for describing people and resources on the web using personal profile information and social relationships. Originally developed to create online communities, FOAF has been applied to various contexts and has been enhanced to address domain-specific representational requirements (Graves et al., 2007). Archival content is inherently permeated by the human

[1] FOAF Vocabulary Specification 0.98, D. Brickley, L. Miller, 9 August 2010. http://xmlns.com/foaf/spec/.

https://doi.org/10.23106/dcmi.952135792

element and as such, it offers an ideal environment for a semantic technology such as FOAF to interlink data centered on people.

2. Project Description

The Linked Jazz project intends to provide a case study that explores a new perspective on the interpretation of archival content. FOAF is proposed as a possible solution to help identify interconnections among jazz musicians referred from digital archival data. The jazz community is characterized by a high level of interaction and connectivity (Heckathorn & Jeffri, 2003). The FOAF ontology describes links among people-centered data in order to create social networks and has the potential to make visible the rich and diverse social networks within which jazz musicians often practice. Specifically, FOAF addresses personal aspects of information that include social relationships, represented by the property "knows". By leveraging the property `foaf:knows`, we can start answering questions such as, "How many people does this artist know?" or, "Is this artist connected to another specific artist?" FOAF, as any RDF-based vocabulary, can also be extended with properties from other vocabularies enabling users to describe an even broader range of personal, social and professional relationships. The Linked Jazz project intends to use FOAF in its basic form and later in an extended version to create a LOD dataset of descriptions of jazz musicians from archival collections. The goal is to develop a method for revealing the network of relationships among jazz musicians, ultimately providing a tool useful to scholars for analyzing the history of jazz.

The linked data approach is still highly experimental. LOD technology is likely to present unanticipated challenges, especially with its application to a new area such as cultural archives. The project is and will continue to progress through iterative steps, so that assessment and subsequent adjustments can be performed at each stage as needed.

This paper describes the pilot study conducted on a series of interview transcripts of jazz artists in order to test the methods needed to identify basic connections among jazz artists. We began with the assumption that if a musician talks about another musician in an interview it is likely that the two musicians have some type of relationship, be it friend or acquaintance of, knowledge of, or familiarity with. While this degree of "knowledge" is entirely implicit, defining the network of jazz artists based on citations found in the interview transcripts was deemed a necessary step in developing the first layer of linkages. To determine who talked about whom, personal names were extracted from interview transcripts and matched against a directory of jazz artist names stored as N-triples. A baseline social network of jazz artists was created based on the only social relation that the FOAF core vocabulary contains: `foaf:knows`. This property would provide the initial means to connect jazz artists.

3. Methodology

3.1 Sample Selection

Fifteen institutions with holdings of jazz history documents were chosen for the richness of their collections. Jazz archives often hold materials that are unique and, thus, extremely valuable. We learned first-hand, however, that these materials can be very hard to find. Only a small portion of content is digitized, which limited our choice of documents eligible for the sample dataset. Twelve documents were selected to serve as a sample for the pilot. The documents were all transcripts of taped interviews with jazz musicians acquired from Hamilton College Jazz Archive[2], Rutgers Institute of Jazz Studies Archive[3], and Smithsonian Jazz Oral Histories. [4] The documents were in PDF and text format, ranging from 12 to 187 pages in length. This type of

[2] http://www.hamilton.edu/jazzarchive.
[3] http://newarkwww.rutgers.edu/IJS/index1.html.
[4] http://www.smithsonianjazz.org/index.php?option=com_content&view=article&id=22&Itemid=28.

https://doi.org/10.23106/dcmi.952135792

resource was chosen because its content often presents a high density of interconnections among people.

3.2 Directory Creation

The next step was to create a directory of jazz artist names paired with dereferenceable URIs. When considering potential sources for name data, we deemed it important to choose one that was stable, used widely and one which provided useful, human-readable URIs to facilitate manual curation of the dataset. We investigated the Library of Congress Authorities & Vocabularies website[5] and the Virtual Internet Authority File (VIAF)[6] as potential sources of authority-controlled names. At this time, however, the Library of Congress Authorities do not include an RDF-based name authority. Library of Congress Subject Headings (LCSH) have been made available as linked data and were considered as a potential source of personal names. Nevertheless, the syndetic nature of the headings, with pre-coordinated term strings, would have required significant parsing to make them useful for our purpose. As for VIAF, its URIs dereference to bibliographic record data, which may be useful for future developments such as linking to discographies and bibliographic data, but were not appropriate for our immediate goal.

DBpedia was chosen as the source for URIs because it is a major and well-established LOD dataset, easily accessible via SPARQL endpoint. Its URIs are human readable and provide musical genre information, making it easier to retrieve data about specific categories, such as jazz musicians. DBpedia was initially queried for literal triples with a `foaf:name` predicate that satisfied the following criteria: the subject resource must have an `rdf:type` of `dbpedia:MusicalArtist` as defined in the DBpedia ontology and the resource must have a `dbpedia:genre` property of `dbpedia:Jazz`. Subsequent queries needed to be performed, and the results re-combined, to overcome the limit of 2,000 answers per query imposed by DBPedia's public SPARQL endpoint (Passant, 2010). Queries returned a total of 2,676 URIs paired with literals. Results were then normalized in a text editor where duplicate triples and stray errors, such as unexpected quotation marks, were removed. The resulting RDF triples were saved to an N-triples file (see Listing 1).

```
<http://dbpedia.org/resource/Artie_Shaw>
  <http://xmlns.com/foaf/0.1/name> "Artie Shaw"
```

LISTING 1. Example of a literal triple from the directory.

3.3. Test of the Directory

An initial assessment of the name directory was conducted using a single document as a test case. The transcript of an interview with the musician Mary Lou Williams was chosen from the sample because of the high number of personal names mentioned in the text. To search for and record the instance of these names, a Python script was written that parsed the PDF transcripts and searched for each name literal in the jazz directory. Of the 121 artist names mentioned in the interview, only 54 (44.6%) were matched against the names in the directory. The analysis of the results revealed three main issues: 1) ten names, although present in the directory, were not found by the script; 2) 39 names were not in the directory, but they had a corresponding DBpedia record; and 3) 18 names were not in the directory and did not have a corresponding DBpedia record.

The first issue concerned the presence of numerous variations of the form of personal names such as abbreviations or diminutives that were not included in the directory. For example, the name value "Miles Davis" in the interview did not match the name value "Miles Dewey Davis" in the directory. To optimize recall, the directory was expanded by modifying the SPARQL query to include the `rdf:label` property in addition to `foaf:name`. In subsequent searches of the

[5] http://id.loc.gov/.
[6] http://viaf.org/.

https://doi.org/10.23106/dcmi.952135792

test document, the ten missing matches were then found.

The second issue had to do with inconsistent categorization of musicians in Wikipedia, the original source of the DBpedia data. Prominent musicians who we expected to find by querying `dbpedia:Jazz` were not returned. This was the case with "Count Basie," which fell under `dbpedia:Swing_music`, `dbpedia:Big_band_music` and `dbpedia:Piano_blues`, but not under `dbpedia:Jazz`. This required us to revise our query method by expanding it to include additional relevant music genres found using the DBpedia ontology (e.g., `dbpedia-owl:influencedby`). Of the 39 names missing, 27 were found as the result of the query revision. The remaining nine were either not relevant as they referred to artists outside the domain of jazz (e.g., vaudeville performers or dancers) or were not found due to misspellings or other typos found in the source document. These were corrected manually.

The third issue was the absence of a corresponding Wikipedia entry for some of the jazz artists. This issue could not be resolved immediately. Additional datasets are under consideration as sources of name values to be integrated in the current directory.

After the revisions following the test, the directory increased from 2,676 triples describing 2,367 individuals to 17,559 triples (+557%) describing 6,444 individuals (+172%). At this time, we are keeping all the variants in the data set to maximize matching.

This test helped to identify immediate problems with creating a directory of names and to better understand the benefits and limitations of using DBpedia as a source of name data. Expanding the directory was beneficial to improve recall and did not affect in any significant way the processing speed of the script. It did, however, lead to a few false matches with non-jazz artists that were removed manually. As the project progresses, the directory will be refined to reduce redundancy and improve consistency and completeness. Effective methods to assist with the problem of matching name variants are being investigated. Also, as name authorities for linked data become available, we hope to be able to use them to reconcile and cross-reference name variants.

3.4. Encoding Social Connections

The final step of the pilot consisted of searching for jazz artist names within the transcripts and recording the resulting matches as an RDF link. For each name found in the document, a triple was created that indicated the interviewee "knows" this artist. In the example shown in Listing 2, the name "Art Blakey" was found in the text of an interview with "Mary Lou Williams."

```
<http://dbpedia.org/resource/Mary_Lou_Williams>
<http://xmlns.com/foaf/0.1/knows>
<http://dpedia.org/resource/Art_Blakey>
```

LISTING 2. Example of an RDF triple generated by the script.

4. Results and Future Work

This procedure was applied to the whole interview sample, generating 952 connections among the twelve interviewees and 529 of the jazz artists listed in the directory (see Table 1).

TABLE 1. Connections among interviewees and other jazz artists in the directory.

Artist	Total Connections
Artie Shaw	87
Billy Taylor	122
Chico Hamilton	108
Clark Terry	40
Danny Barker	112
Doc Cheatham	96
Frank Foster	183

https://doi.org/10.23106/dcmi.952135792

Jane Jarvis	15
Lionel Hampton	26
Marian McPartland	34
Mary Lou Williams	85
Slide Hampton	44
TOTAL	**952**

Basic statistics indicating the number of connections each interviewee shares with one another were also calculated (see Table 2). For example, Frank Foster shares 215 of his connections with one or more of the other interviewees, while Slide Hampton shares only 92.

TABLE 2. Interviewees' shared connections.

	Artie Shaw	Billy Taylor	Chico Hamilton	Clark Terry	Danny Barker	Doc Cheatham	Frank Foster	Jane Jarvis	Lionel Hampton	Marian McPartland	Mary Lou Williams	Slide Hampton
Artie Shaw		25	9	7	24	14	21	4	8	11	19	11
Billy Taylor			33	14	37	33	42	8	12	16	39	15
Chico Hamilton				12	19	12	33	3	10	10	14	11
Clark Terry					11	10	23	5	8	4	15	7
Danny Barker						29	28	6	8	10	23	8
Doc Cheatham							23	7	7	7	33	6
Frank Foster								5	17	12	34	19
Jane Jarvis									2	6	7	3
Lionel Hampton										7	9	8
Marian McPartland											11	7
Mary Lou Williams												11
Slide Hampton												

While a full analysis of the network of interactions resulting from this set of data is out of the scope for this paper, a visualization in the form of a force-directed graph was created to provide an overview of the network. As Figure 1 shows, jazz artist names in the directory are represented as circular nodes whose size conveys the frequency of mentions by the interviewees, who are represented by triangular nodes. Nodes are placed and clustered together based on their shared connections to the jazz artists from the directory. For example, Slide Hampton is placed between the two larger clusters, as he possesses a relatively even distribution of shared connections with the members of each cluster.

https://doi.org/10.23106/dcmi.952135792

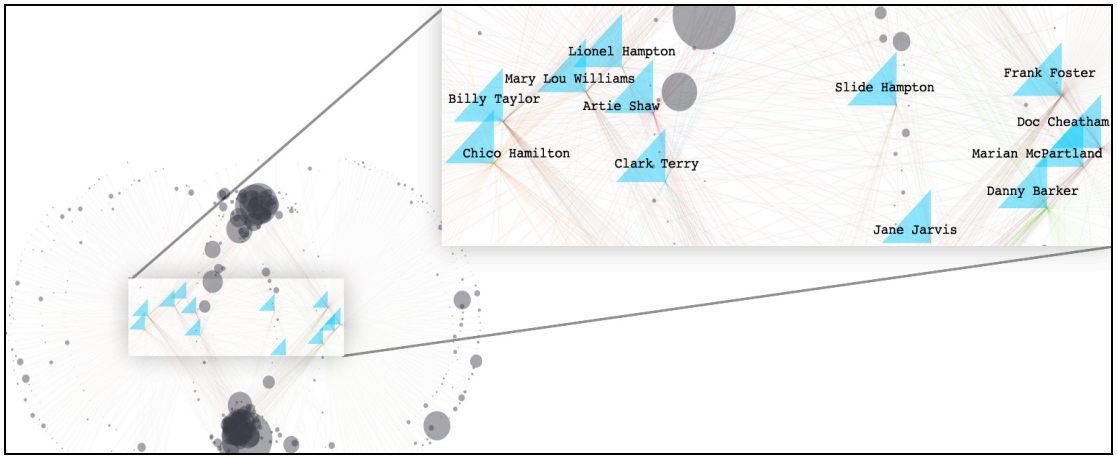

FIG. 1. Visualization of the social network of jazz artists as based on connections found in interview transcripts.

Because the nature of the connections identified in the pilot remain entirely implicit, we can only assume that jazz artists citing other jazz artists in their interviews are likely to have some kind of social connection. That connection could be anything from close friendship and collaboration to a brief encounter or simply familiarity. The pilot used the basic `foaf:knows` relationship as a sort of placeholder at this stage of development. Finding matches among jazz artists' names is only the initial step in a process of discovery that will map the relationships in the community of jazz musicians. Additional steps need to be taken to capture and interpret the nature and the degree of the interpersonal connections emerging from this social graph. Methods to refine the jazz artist directory are being investigated. As the project progresses, contextual data, including places and time periods, will be considered for data interlinking. A number of LOD datasets, including Geonames[7] and LODE, an ontology for Linking Open Descriptions of Events,[8] are likely to be used. In addition, a weighted scoring model that enables us to identify different degrees of social relationship is under construction. The accuracy of these relationships will be then evaluated by domain experts.

5. CONCLUSION

The pilot discussed in this paper describes a method to create a dataset of RDF triples representing jazz artist names and their social connections. This method, although experimental, offers a first step towards the construction of a well-defined linked data dataset connecting jazz artists.

Because this area of research is still in its early stages, there is a need for case studies and prototypes to be tested so that sound principles, methods, and best practices can be developed. In the spirit of learning by doing, this project aims to provide a case study that shows the opportunities and challenges of interlinking fragmented cultural data and making them searchable as a whole to create new kinds of meaning and elicit new streams of interpretation.

Acknowledgements

We would like to thank Ben Fino-Radin for his assistance with data analysis and visualization. This project was supported by the OCLC/ALISE Library and Information Science Research Grant, 2011.

[7] http://www.geonames.org/.
[8] http://linkedevents.org/ontology/.

https://doi.org/10.23106/dcmi.952135792

References

Graves, Mike, Adam Constabaris, and Dan Brickley. (2007). FOAF: Connecting people on the semantic web. Cataloging & Classification Quarterly, 43(3/4), 191-202.

Heath, Tom, and Christian Bizer. (2011). Linked data: Evolving the web into a global data space. San Rafael, CA: Morgan & Claypool.

Heckathorn, Douglas D., and Joan Jeffri. (2003). Social networks of Jazz musicians. In Changing the beat: A study of the worklife of Jazz musicians, (Vol. 3, pp. 48-61). Washington, D.C.: National Endowment for the Arts Research Division Report #43.

Passant, Alexandre. (2010). Dbrec — music recommendations using DBpedia. In Peter Patel-Schneider, et al. (Eds.), The semantic web – ISWC 2010 (pp. 209-224). Berlin: Springer.

https://doi.org/10.23106/dcmi.952135792