FOAF in the Archive: Linking Networks of Information with Networks of People: Final Report to OCLC

M. Cristina Pattuelli
School of Information and Library Science
Pratt Institute, New York
mcpattuelli@gmail.com

Final Report
2011 OCLC/ALISE Library and Information Science Research Grant Project
March 13, 2012

Suggested citation:
Pattuelli, M. Cristina. 2012. "FOAF in the Archive: Linking Networks of Information with Networks of People: Final Report to OCLC." 2011 OCLC/ALISE research grant report published electronically by OCLC Research. Available online at:
http://www.oclc.org/research/grants/reports/2012/pattuelli2012.pdf

Table of Contents

Statement of Purpose

Aim and Scope

The amount of cultural heritage data available in digital repositories is growing at an exponential rate and the need for systems that enhance discovery and analysis of digital cultural objects is greater than ever. The application of Linked Data technology to cultural heritage data is a promising strategy to address these needs. Linked Data is both a community effort and a W3C project, and is a recommended best practice for connecting distributed data across the web (Heath & Bizer, 2011). Linked Data technology has emerged in the context of semantic web development as a means to facilitate the exposure of machine-readable data across the web to enable data interlinking and to promote interoperability. As an extension of the traditional web, Linked Data intends to provide a unifying and open publishing framework for discovery, integration and reuse of data.

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, these institutions have the potential to extend the reach of their own collections beyond existing controlled environments to enable discovery and to generate new ways of interpretation of their content.

The Linked Jazz[1] project investigates the application of Linked Data technology to digital primary sources from jazz history archives. The central goal of the project is to develop methods for creating Linked Data semantics that help reveal to the network of social relationships among jazz artists.

Overview

We initially focused on the use of Friend of a Friend (FOAF),[2] one of the most popular Linked Data vocabularies, to describe social relationships among jazz artists. Based on the property `foaf:knows`, we created a set of RDF assertions representing a basic layer of

[1] http://linkedjazz.prattsils.org/
[2] http://www.foaf-project.org/

connections among jazz artists as described in our primary sources. The data source was a collection of transcripts of taped interviews from jazz history archives. A visualization was created representing the social network of jazz artists as represented by our dataset. We then extended FOAF to include properties from other Linked Data ontologies to enhance its expressive capability and to better represent different types of connections among jazz musicians. Both a machine- and a human-driven approach were identified to enrich the initial dataset with RDF triples in order to describe more granular connections among jazz artists. The machine-driven strategy was applied to determine the extent of musical collaboration among artists. The human-driven approach was deemed appropriate for identifying more nuanced social relationships expressing degrees of personal closeness among artists (e.g., friendship). The two complementary approaches required the design and development of a new array of methods that are currently being tested.

While Linked Data relies on a relatively simple technology framework based on a small set of open standards, its implementation remains largely experimental, especially in the field of cultural heritage. The Linked Jazz project consisted of multiple phases, sub-phases and iterative steps in a prototype development style. This progression allowed for assessments and subsequent adjustments to be performed at various stages to address unanticipated challenges.

The Linked Jazz project lies at the intersection of several fields of study. Its findings and research practices could be applied to Linked Open Data (LOD), digital humanities and archival research. LOD, the open license version of Linked Data, has recently emerged as a vital area of development in cultural heritage, with a growing and active community (e.g., LOD-LAM[3]). In this context, Linked Jazz contributes to the development and assessment of strategies for creating cultural LOD semantics for the purpose of creating services and web applications.

The Linked Jazz project is especially relevant to digital humanities. It offers an innovative approach to the representation of digital content centered on the identification of social networks. This perspective has the potential to foster new learning and scholarship by facilitating innovative forms of intellectual discovery and analysis of jazz history content.

[3] International Linked Open Data in Libraries, Archives, and Museums. http://lod-lam.net/summit/

Linked Jazz is one of the few projects that brings Linked Data research to the context of historical archives. Typically, access to archival materials relies on descriptive practices centered on the notion of "document" and employs collection-level finding aids, often confined to siloed databases. The Linked Jazz project demonstrates the potential of expanding traditional access to archival content by relying on web open standards to make archival resources part of a global and unified information environment.

Conceived as an exploratory study, this project also serves as a case study that illustrates both the potential and the challenges associated with implementing Linked Data technology in the cultural heritage domain. Because research exploring Linked Data in the cultural heritage domain is still relatively new, there is a need to develop a knowledgebase of sound principles, methods, and best practices for the field.

Literature Review and Related Works

In practice, the Linked Data initiative is centered on a growing collection of datasets that can be easily cross-referenced by machines, with millions of links already available among datasets. The Linked Data paradigm has been rapidly establishing itself as an influential framework for research in a wide range of fields. Most Linked Data initiatives have left the research lab to become full-fledged implementations. Linked Data and LOD have gained momentum in several contexts including government (e.g., data.gov.uk in Britain and data.gov in the United States), online business (e.g. Google and Yahoo), and news organizations (e.g. *The New York Times*, Thompson Reuters, and the BBC).

Libraries and cultural organizations are also actively participating in LOD research. LOD is seen as a promising solution to some of the more compelling problems digital libraries face in making their content findable beyond the boundaries of the repository and in providing new ways to create greater context and meaning. Libraries have been active participants in LOD research initiatives with the intent to expose their bibliographic data to the web. There is tremendous potential to transform legacy bibliographic metadata into Linked Data and make that bibliographic information cross-linkable as part of a global network (Malmsten,

2008). Based on their experience building a LOD service for the German National Library, Hannemann & Kett (2010) describe the benefits as well as the challenges of Linked Data technology as it pertains to libraries. The authors recognize the potential Linked Data has to complement bibliographic data with data from other domains and external sources. Exposing this information to a global interlinked discovery environment has the potential to greatly enhance a library user's search experience. In addition, by sharing their extensive collections of high quality bibliographic metadata and authority data, libraries could contribute significantly to the Linked Data movement by establishing a "backbone of trust[ed] data" (Hannemann & Kett, 2010, p. 2). Among the challenges the authors identify are the technical issues related to building a large-scale Linked Data service. Currently, available tools for implementing such a service often lack stability and supporting documentation. Hannemann & Kett (2010) also discuss data modeling issues and argue for better definitions of individual properties, as well as for the need to establish best practices for modeling library data. Recently, a few national library catalogs have been published as Linked Data, including the National Library of Sweden, the German National Library, and the National Széchényi Library (NSZL) of Hungary and, more recently, the British Library.

Research specifically focused on the cultural heritage domain is scarce. The W3C Library Linked Data Incubator Group recently collected an inventory of use cases and case studies from the library community that demonstrate the vitality of the field (Vila Suero, D., 2011). Few projects, however, are coming from the field of cultural heritage. One example is Civil War Data 150,[4] a prominent archive-related Linked Data project that is partnering with state and local archives to use LOD to link Civil War data and objects for greater discovery.

Compared to other cultural memories institutions, archives face unique challenges when it comes to participating in LOD. Archival metadata standards typically describe content at the collection level. Unlike libraries and museums, archives do not have a history of cooperatively creating and sharing machine-readable records. As such, archives lack an infrastructure of structured metadata that can be converted, published, and consumed as Linked Data.

Jazz archives present an even more challenging set of problems. As Fitzgerald (2008) points

[4] http://www.civilwardata150.net/.

out, jazz archives are a recent phenomenon since jazz is, in itself, a relatively young art form. The majority of efforts have been devoted to building collections rather than creating description and facilitating access. The lack of effective discovery systems tends to leave these collections hidden and under-used.

Methods and Findings

The central goal of the Linked Jazz project was to explore strategies and develop methods for creating RDF triples to represent relationships among jazz artists as described by primary sources. Such semantics could help create services and applications for supporting scholarly research and learning, including the facilitation of social network analysis for the jazz community. As a reference framework, we adhered to Tim Berners-Lee's (2006) five star LOD principles for methodology and future data dissemination: 1) the data created in the context of the Linked Jazz project are publicly available; 2) the data are available in machine-readable structured format; 3) we rely on non-proprietary format to publish; 4) we employ W3C open standards including RDF, SPARQL, and LOD datasets (i.e., DBpedia and MusicBrainz); 5) all of our URIs are URLs resolved to existing web pages to provide context and interlink with other people's data.

The project includes two main phases and proceeded in the form of prototype development. To this end, both the assessment and review of methods performance were intrinsic components of the process.

Phase one involved creating: 1) a dataset of connections among jazz musicians based on the property `foaf:knows`, 2) a visualization of the dataset as a social network.

Two subsequent pilots were conducted during phase one. This portion of the project spanned January 2011 to August 2011.

Phase two focused on: 1) the creation of a vocabulary extension to enhance representation capabilities beyond the basic FOAF ontology, 2) the identification of methods for representing more granular connections based on the extended vocabulary, 3) the generation of relationships describing collaboration between artists, 4) the design of a crowdsourcing tool for further analysis of relationships among artists.

The second phase occurred from September 2011 to December 2011. The OCLC/ALISE grant was pivotal for supporting both phases of the project and for laying the foundation for the further developments that are currently being explored.

Phase 1

The first phase of the project focused on the development of methods for identifying the connections among jazz artists to be recorded as RDF triples. We used transcripts of taped interviews with jazz musicians as the data source. Name citations found in the transcripts provided the basic units for creating a first layer of linkages among musicians. The linkages were expressed by the property `foaf:knows`. The assumption underlying this strategy was that if a musician mentions another musician in an interview it is likely that the two musicians have some type of relationship, be it friend of, acquaintance of, knowledge of, familiarity with, etc. By leveraging the property `foaf:knows`, we could start answering questions such as, "How many people does this artist know?" or, "Is this artist connected to another specific artist?"

A pilot was first conducted on twelve transcripts of taped interviews to test the methodology. A second iteration was later performed on an extended sample of fifty transcripts. Visualizations of the resulting network of relationships were created for both datasets.

Data Sources

Fifteen institutions holding collections of jazz history primary sources were identified as potential sources of data. Jazz archives often hold rich collections of oral histories along with other unique materials. Oral histories were deemed an ideal source for identifying social connections since their content often presents a high number of citations of people from the jazz community. While oral history collections are a common presence in several institutions, few of the oral histories have been transcribed from the original audio recordings. The limited availability of digitized transcripts reduced the number of data sources eligible for the project's testbed. Ultimately, transcripts were selected from four

institutions: Hamilton College Jazz Archive,[5] Rutgers Institute of Jazz Studies Archive,[6] Smithsonian Jazz Oral Histories,[7] and The University of Michigan's Nathaniel C. Standifer Archive of Oral History.[8] A jazz studies researcher was involved in the selection of transcripts to ensure diversity in terms of geographical location and time period. The documents, ranging from 12 to 187 pages in length, were originally in, or were later converted to, PDF format (see Appendix I).

Technical Framework

LOD principles and technology are in line with W3C open standards and technologies. Specifically, LOD rely on HTTP URIs that make reference to entities of any type that can then be de-referenced (e.g., looked up by humans or software agents) and uses RDF as its data model (Berners-Lee, 2006). The LOD infrastructure facilitates data interoperability and integration by offering a common format for heterogeneous data. The creation of a unifying data space has the potential to generate new opportunities for information discovery, visualization, and interpretation.

HTTP URIs and RDF were part of the Linked Jazz project platform, which also included SPARQL,[9] the query language for RDF. The FOAF ontology and other RDF vocabularies provided the semantics to describe connections among jazz artists.

FOAF was the first and one of the most successful attempts to develop a tool to represent online profiles of people and describe their interpersonal relationships. Originally conceived as an RDF project for online presence in a pre-Facebook era, FOAF has become part of the Linked Data initiative and is today a trusted and widely adopted ontology. According to its most recent specifications, released in August 2010, FOAF uses the web as an environment where factual information can be integrated with information from human-oriented documents (e.g., videos, books, spreadsheets, 3D models) for the purpose of supporting a wide range of information-linking efforts. The design of FOAF is intentionally simple and flexible to facilitate the inclusion of sets of terms that are useful to the web

[5] http://www.hamilton.edu/jazzarchive
[6] http://newarkwww.rutgers.edu/IJS/index1.html
[7] http://www.smithsonianjazz.org/index.php?option=com_content&view=article&id=22&Itemid=28
[8] http://www.umich.edu/~afroammu/standifer.html
[9] http://www.w3.org/TR/rdf-sparql-query/

community without losing focus of its key purpose, to link "networks of information with networks of people" (Brickley & Miller, 2010). Another advantage of FOAF is its decentralized nature, making it useful as an exchange format between existing networks as well as a standalone ontology tool (Dumbill, 2002).

From a knowledge representation perspective, FOAF is a simple RDF-based model with a high level of versatility to accommodate descriptive needs of specific communities and domains of interest. FOAF includes basic classes and properties that describe characteristics of people and social groups independently of time and technology, but virtually any RDF vocabulary could be embedded in a FOAF core vocabulary to enhance its representation capabilities. In the context of the Linked Jazz project, only the property `foaf:knows` was deemed suitable for expressing social relationship information. This property provided the means to create the first level of connections in phase one of the project. A range of Linked Data vocabularies, including the RELATIONSHIP[10] ontology and Music Ontology,[11] were employed as sources of additional properties to describe social relationships.

Name Directory

The first step in the development of the Linked Jazz project was to create a directory of jazz artist names paired with de-referenceable URIs. DBpedia was chosen as the primary source of URIs because it is a well-established linked open dataset that provides easy query access to large amounts of structured content. Its URIs are human readable and categorized according to a main ontology and other organization structures that facilitate query construction using domain specific predicates, (e.g., *Swing_music* or *Jazz*).

The English language version of DBpedia 3.6, based on Wikipedia dumps from October through November 2010 was queried using SPARQL 1.0. Several types of SPARQL queries had to be performed to create a directory that included the entire pool of URIs representing the personal names of the jazz artists we had identified from our data sources. An example of a SPARQL query utilized in this context is shown in Listing 1. For example, the directory was expanded by modifying the SPARQL query to include the `rdf:label` property in

[10] http://vocab.org/relationship/.html
[11] http://musicontology.com/

addition to `foaf:name`. Further revisions to the query were necessary to overcome issues related to inconsistent categorization of musicians in Wikipedia, the original source of the DBpedia data. This was the case with "Count Basie," which was returned with a query that included `dbpedia:Swing_music`, `dbpedia:Big_band_music` and `dbpedia:Piano_blues` predicates, but not, surprisingly, when we used the predicate `dbpedia:Jazz`. After initial revisions, the directory increased from 2,676 triples describing 2,367 individuals to 17,559 triples (+557%) describing 6,444 individuals (+172%). More details on the assessment of the name directory can be found in Pattuelli, Weller & Szablya (2011).

```
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
SELECT * WHERE {
?subject rdf:type dbpedia-owl:MusicalArtist.
?subject dbpedia-owl:genre dbpedia:Jazz.
}
```

Listing 1. Example of SPARQL query.

Expanding the directory was beneficial to improving recall and did not significantly impact the processing speed of the script. The directory was then refined. Results were normalized in a text editor where duplicate triples and stray errors, such as unexpected quotation marks, were removed. More work is needed to improve the quality of the directory in terms of parsimony, completeness, and consistency. To this end, testing is currently underway that involves mapping to authority names (e.g., Library of Congress Authorities, VIAF). The quality of the directory, however, was not an issue for our immediate goals. In fact, we kept all the name variants and tolerated a high level of redundancy for the sake of maximizing matches. The resulting triples were saved to an N-triples file. An example of an instance URI from the name directory is shown in Listing 2.

```
<http://dbpedia.org/resource/Miles_Davis>
<http://xmlns.com/foaf/0.1/name> "Miles Davis" .
```

Listing 2. Example of a literal triple.

Jazz Engine

The next step was to develop a method for identifying connections and recording them in RDF format. A set of Python programs, which we called Jazz Engine, was written for named

entity recognition. First, a Python script was written that searched for and extracted personal names from a sample of interview transcripts. The script parsed the PDF transcripts and searched for matching name literals in the jazz directory. Every instance of a match generated a record in the form of an RDF statement that included the URI of the interviewee as its subject, the property `foaf:knows` as its predicate, and the name value of the artist cited in the interview as its object. In the example shown in Listing 3, 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 3. Example of an RDF triple generated by the script.

Figure 1 shows the workflow of the process of creating RDF triples representing connections among jazz artists. This procedure was applied to the interview sample as a small-scale trial run to identify any practical issue. This pilot generated 952 connections among the twelve interviewees and 529 of the jazz artists listed in the directory. Basic statistics, indicating the number of connections each interviewee shares with one another, were also calculated (see Appendix II).

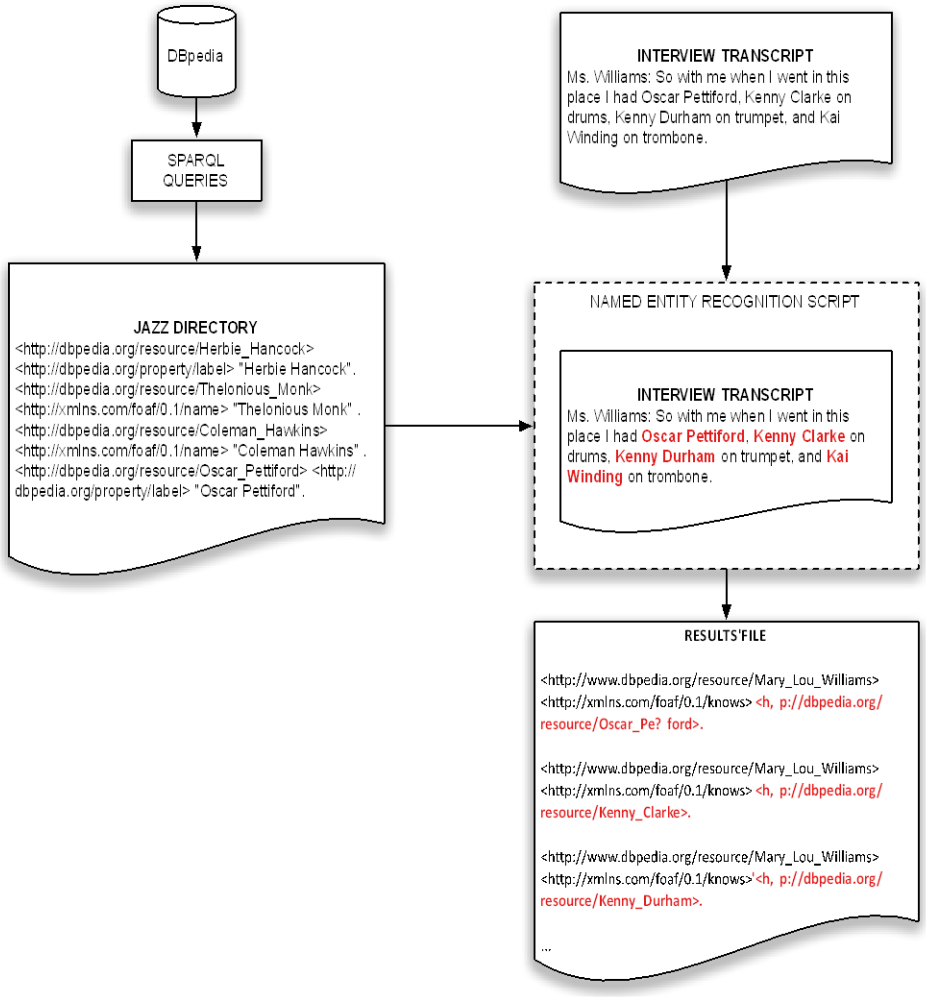


Figure 1. Workflow of the process of creating RDF triples.

While an analysis of the social network based on our results was not the focus of the project, we created various visualizations to provide an overview of the network and help to better understand its potential for data analysis. A social graph was generated using Javascript InfoVis Toolkit's[12] force-directed algorithm (Figure 2). Jazz artist names in the directory are represented as circular nodes, the size of which conveys the frequency of citations. The interviewees are each represented by triangular nodes. Both the circular and triangular nodes are placed and clustered according to 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. Additional visualizations of the dataset are shown in Appendix III.

Figure 2. Visualization of the social network of jazz artists.

To consolidate the method and ensure high reliability and validity of the findings, the procedure was replicated using a sample of fifty transcripts (see Appendix I). The process did not encounter any major problems and confirmed the effectiveness of the methods used. This run generated 8,307 connections among the fifty interviewees and 946 of the jazz artists listed in the directory. The results provided a sizable dataset of triples. The possibility to identify connections on a broader scale allows for more accurate analysis of the social network.

[12] http://thejit.org/

The creation of a network of jazz artists based on citations found in the interview transcripts was a key step for identifying the network of connections among the jazz artists described in the sample of interview transcripts. The `foaf:knows` relationship served as the connector to develop the first layer of linkages. The nature of the connections, however, remains entirely implicit and we can only assume that jazz artists citing other jazz artists in their interviews are likely to have some kind of social connection. This relationship could be anything from close friendship and collaboration to a brief encounter or familiarity. This method made it possible to take the initial step in a process of discovery that will map the relationships in the community of jazz musicians. New strategies need to be carried out to capture and interpret the nature and the degree of the interpersonal connections that emerged from the this phase of the project.

Phase 2

The second phase of the project focused on devising and implementing a suitable approach for defining more granular relationships among the jazz musicians mentioned in the interview transcripts. As discussed earlier, the FOAF ontology property `foaf:knows` has served as the semantic glue for connecting jazz musicians cited in the interview transcripts. No other properties, however, are provided by the FOAF vocabulary to describe social relationships. Therefore, the next step in the development of the Linked Jazz project consisted of creating an extension of FOAF that would include properties from other vocabularies suitable to refine our social network. From a knowledge representation perspective, FOAF is a simple RDF-based model and, as with any RDF vocabulary, can be extended to enhance its descriptive capabilities. Properties from different RDF vocabularies can be mixed and matched to accommodate specific representative needs. Indeed, re-use of properties, whenever possible, to integrate existing vocabularies, is recommended in the Linked Data environment as a best practice to maximize interlinking.

Relationship Refinement

To identify suitable properties with which to refine the connections from the pilot, RDF-based vocabularies with the potential to serve as proper sources of semantics were

analyzed and an inventory of candidate properties was compiled.[13] A sub-set of properties was then selected that was deemed suitable for describing the existing social network in more detail. The sub-set was derived from RELATIONSHIP vocabulary, Music Ontology and FOAF ontology. Figure 3 shows the relationships listed in a spectrum representing degrees of personal and professional connections. The properties are prefixed with the namespace of their corresponding ontology or vocabulary.

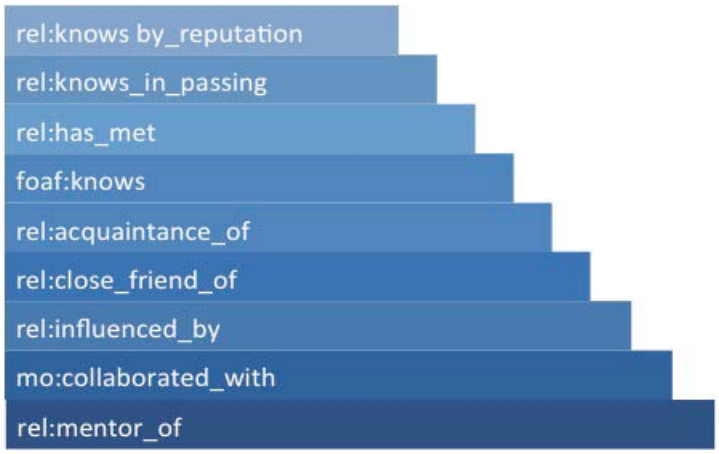


Figure 3. Spectrum of properties.

Specifying connections in a granular way is a challenging endeavor in many areas of application. The high degree of instability and semantic noise intrinsic to the LOD environment adds significant challenges to this effort. Initially, traditional approaches that employ Natural Language Processing (NLP) methods were considered. These were soon discarded due to the complexity anticipated given our specific context and the researcher's lack of experience in the NLP field. We decided instead on a dual approach to refining the connection that included both human- and machine-driven processes. This hybrid strategy seemed appropriate for pursuing the project's goals, and it was also realistic in terms of implementation. Social relationships present various levels of complexity when it comes to knowledge representation. Within our sub-set, there was only one property, `mo:collaborated_with`, which did not appear open to subjective interpretation and could therefore be derived automatically. A machine-driven method would be extremely hard to apply to the rest of the relationships in the spectrum due to the ambiguity of the

[13] http://linkedjazz.prattsils.org/?page_id=10

properties expressing nuanced degrees of personal connections (e.g., `rel:knows_in_passing` or `rel:close_friend_of`).

The two approaches focused on investigating: 1) a programmatic method to leverage linked open discography data for defining the relationship of collaboration (`mo:collaborated_with`), 2) a crowdsource-driven analysis for defining the rest of the properties.

Machine-driven Approach

As discussed earlier, the property `mo:collaborated_with` carries less semantic ambiguity then the rest of the properties from our spectrum. In the context of this project, `mo:collaborated_with` was considered a "low hanging fruit," unequivocal enough to be detected automatically. Collaboration could be claimed with a high degree of confidence when, for example, two or more musicians are listed in the same recording session or musical track. This is the type of information that can be found in discography databases, but it is also disseminated in various LOD datasets. MusicBrainz,[14] one of the richest and most established public datasets for music information, and part of the LOD environment, was selected as our source of recording data.

To create statements that represent relationships of collaboration among jazz musicians, a method was devised that included the following steps. First, we collected data from MusicBrainz that we could use to identify collaborations among the musicians. After analyzing the MusicBrainz data set, we determined that we could leverage the entity "recording" for our purposes. According to MusicBrainz terminology, the entity "recording" represents attributes associated with unique audio data. These attributes include both a musical track title and an artist credit.[15] Multiple artists can be associated with one recording, therefore making it possible to identify collaborations.

The strategy for acquiring MusicBrainz data was not, however, straightforward. It required a number of intermediary steps. First, we needed to find MusicBrainz identifiers

[14] http://musicbrainz.org/
[15] http://musicbrainz.org/doc/Recording

corresponding to the URIs in our name directory. We chose to adopt a method that made use of <sameAs>,[16] a public web service that helps find URIs referring to the same entity, should any be available.

We began by querying <sameAs> using URIs from the jazz name directory in order to retrieve corresponding MusicBrainz artist names. Although MusicBrainz URIs are not directly available through <sameAs>, MusicBrainz identifiers (MBIDs) are part of URLs on other websites, (e.g., DBtune,[17] Zeitgeist,[18] and the BBC), and the URIs of artist pages on these websites are often included in <sameAs> results. MBIDs were retrieved and matched with DBpedia URIs. In order to perform the matching, we took advantage of a crosswalk between MBIDs and DBpedia URIs, which was available as a direct download from DBpedia.[19] This crosswalk was created manually in 2009 to provide "links between artists, albums and songs in DBpedia and data about them from MusicBrainz."[20]

Results were then deduplicated and a dictionary was created with 14,917 entries. Each entry consisted of a URI resulting from the mapping between a DBpedia URI and its corresponding MBID (see Listing 4).

```
DBpedia URI:  http://dbpedia.org/resource/Ella_Fitzgerald
MBID:         54799c0e-eb45-4eea-996d-c4d71a63c499
```

Listing 4. Example of a dictionary entry.

The dictionary was stored as a tab-delimited text file and served as the basis for any subsequent matching. In the next step, an application looked up the results from our dataset of triples and generated a list of all artists for which we had an MBID. For each of these artists, the application submitted a query to the Musicbrainz XML Web Service, which used the artist MBIDs to retrieve a list of all the recordings each artist had created and, concurrently, a list of all other artists associated with each recording. Many recordings lacked attributes indicating associated relationships, but the web service API did not have the capability to filter out records with no relationships.

[16] http://sameas.org/
[17] http://dbtune.org/
[18] https://launchpad.net/zeitgeist-project
[19] http://downloads.dbpedia.org/3.6/links/musicbrainz_links.nt.bz2
[20] http://wiki.dbpedia.org/Downloads36#linkstomusicbrainz

Results were returned as XML files containing attributes about the recording (e.g., title, length) and, when available, a list of associated artist names as well as their artists' MBID and relation-type (e.g., instrument played, guest, etc.). These results were saved locally as 2,186 separate XML files, each containing approximately hundred recording records.

These files were parsed by a Python application written for the purpose of identifying the records that contained relationship data. The records containing relationship data were then written to separate XML files for each recording. The XML files were subsequently parsed by another Python application that extracted the relationship data and created RDF statements representing the collaboration between each artist listed in the data for each recording.

A fourth element was added to the RDF statements. This element turned the triples into quads and pointed to provenance of the data in the form of a link to the MusicBrainz page about that recording. An example of such an RDF quad is shown in Listing 5. The quad represents the relationship of collaboration between Dexter Gordon and Billy Higgins. The fourth URL points to the specific recording, *I Guess I'll Hang My Tears Out to Dry*, as the source of evidence for that collaboration.

```
<http://dbpedia.org/resource/Dexter_Gordon>
<http://musicontology.com/#term_collaborated_with>
<http://dbpedia.org/resource/Billy_Higgins>
<http://musicbrainz.org/recording/590c3e6a-7316-4c94-b336-
27f0af719f8e>.
```

Listing 5. Example of a quad.

The issue of provenance is often debated in the LOD community. Various mechanisms that would indicate provenance and thus ensure trustfulness and the ability to trace the origin of the data have been discussed (Orlandi & Passant, 2011). Provenance is a key issue in the context of archival information, and extending RDF statements with provenance information is a valuable addition to archival Linked Data for future traceability.

The method adopted for identifying and recording relationships of collaboration among jazz musicians proved successful in enriching our original dataset with more detailed

relationships. Additional data curation is still needed to improve the accuracy and comprehensive quality of our data set.

Human-driven Approach

For defining more granular properties, we chose crowdsourcing over the machine-based approach. As Oomen and Aroyo (2011) point out, crowdsourcing gained interest in the domain of cultural heritage as a means to support an array of labor-intensive and error-prone tasks including correction, transcription and co-curation of digital materials. Because of the need to improve discovery of massive amounts of cultural heritage material, crowdsourcing has been seen as a way to create "a more open, connected, and smart cultural heritage" by involving both data users and consumers (Oomen & Aroyo, 2011, p. 147).

In the context of this project, crowdsourcing was deployed to assist with the analysis of relationships from interview transcripts. This approach has the potential to yield data with a high degree of accuracy because it is based on human interpretation. Moreover, one of the advantages of crowdsourcing is in the large scale of data that can be collected in relatively short spans of time, especially when relying on a broad and engaged community of contributors.

A web application is currently under development that will leverage the knowledge of jazz scholars and researchers from academic centers for jazz studies as well as jazz aficionados from dedicated online forums. The tool provides an interface to our set of interview transcripts and makes it possible for users to analyze excerpts of text and discern relationships. Upon registering for an account, the user is able to browse the set of interview transcripts. Once a transcript is selected, the user is presented with a list of all of the individuals mentioned by the interviewee, including the number of times an individual was cited. As shown in Figure 4, transcript excerpts are displayed on the screen one at a time. Once an interview is chosen, small excerpts are displayed on the screen. User contributors are provided with a menu of options from which they can select the appropriate relationship represented in the excerpt. The user's input is stored in a relational database associated along with the specific mention. This data can be then

exported as RDF statements and will be integrated into our dataset. The web application is currently in the prototype stage of development, with usability testing planned for late spring 2012.

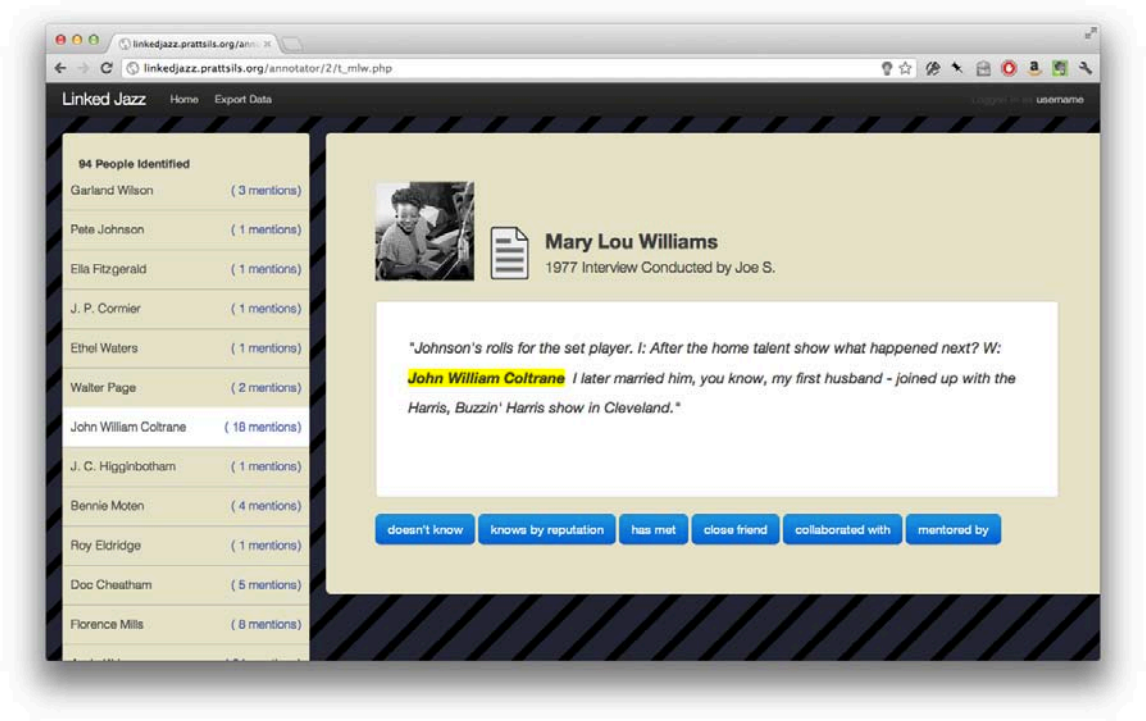


Figure 4. A screenshot of the web application interface.

Concluding Remarks and Future Work

The Linked Jazz project intended to explore innovative ways to enhance the discovery and interpretation of cultural heritage through the application of LOD. Its overriding goal was to help make visible the personal and professional connections among jazz artists. The focus on digital archives of jazz history enabled us to propose a novel stream of research that has the potential to foster new learning and scholarship by facilitating innovative forms of intellectual discovery and analysis of jazz history content. Furthermore, since both LOD and the application of LOD to digital cultural heritage are still relatively new and developing practices, the Linked Jazz project presents several new contributions to these emerging fields.

The project has developed methods for generating a dataset of RDF statements that represent personal and professional relationships among jazz musicians, as they are described in transcript interviews from jazz history collections. Such a dataset would provide the basis for developing web applications and services including visualizations for social network analysis.

Because this is a relatively recent area of research, there is a need for case studies to be shared and prototypes to be tested so that sound principles and best practices can be established. The Linked Jazz project succeeded in showing the potential of real-world LOD applications to enhance the visibility and interpretation of primary source materials.

Due to the experimental nature of LOD research, the project progressed through iterative stages with periodic revisions to the design. We devised and tested various methods to create LOD representations of the rich interpersonal relations among jazz musicians derived from digital archival materials. The Linked Data semantics from our initial pilot was visualized in the form of a social network. These methods can be transferred to other contexts for creating LOD cultural heritage datasets.

The iterative process yielded a number of lessons learned. We realized at the initial stage of the project that the amount of available structured cultural heritage data was too limited to support the complex interlinking among resources we had originally envisioned. This modified our original plans and led the project to an even more experimental direction. For example, we had to generate the semantics we needed on our own in order to create linkages. This proved a fertile step and an original contribution to LOD research.

We also faced several technical challenges that are instructive for the development of an independent LOD project of this nature and size. Some of the challenges have to do with the programming skills required to generate a LOD dataset. More significant challenges are posed by the varying quality of LOD datasets we relied on as data sources, including DBpedia and MusicBrainz. As LOD technology continues to mature and more stable tools become available, it will be possible to streamline methods and continue to explore the

unprecedented opportunities that LOD opens up for cultural heritage data discovery and interpretation.

The OCLC/ALISE grant provided invaluable support for the full process of development and testing and has laid the foundation for further developments. The Linked Jazz project is currently entering a new phase that entails the implementation of a crowdsourcing tool to refine the current LOD dataset. Data curation is also continuing with our existing dataset as well as on the name directory to improve data accuracy, completeness and consistency. All the data will be made available to the LOD community and accessible through the project website at http://linkedjazz.prattsils.org/ by April 1, 2012.

The Linked Jazz project has been presented at the following conferences and venues:

DC-2011, Eleventh International Conference on Dublin Core and Metadata Applications, The Hague, The Netherlands, September 23, 2011. [PPT].

Pattuelli, M.C. (2012). FOAF in the Archive: The Linked Jazz Project. *ALISE 2012*. Dallas, TX, January 19, 2012. [PPT]

Harnessing the Semantic Web for Scholarship (November 2, 2011). *Research Without Borders* event series, The Scholarly Communication Program at Columbia University. [PPT] [Video]

A paper focused on the pilot phase of the project has been published in the Dublin Core conference proceedings:

Pattuelli, M.C., Weller, C. and Szablya, G. (2011). Linked Jazz: An exploratory pilot. In *DC-2011: Proceedings of the International Conference on Dublin Core and Metadata Applications* (pp. 158-164). The Hague, The Netherlands, September 21-23, 2011. [PDF]

Exploratory work to inform the project in its preparatory phase has been presented and published in the follow venues:

Pattuelli, M.C. (2011). Mapping people-centered properties for Linked Open Data. *Knowledge Organization, 38*(4), 352-359. [Preprint]

Pattuelli, M.C. (2011). Mapping subjectivity: Performing people-centered vocabulary alignment. Presented at the *Third North American Symposium on Knowledge Organization (NASKO 2011)*, Toronto, Canada, June 16-17, 2011. [Link] Published in R. Smiraglia (Ed.) *Proceedings from North American Symposium on Knowledge Organization* (pp. 174-184), Vol. 3. Toronto, Canada. [Link]

Acknowledgments

I would like to thank the Institute of Jazz Studies at Rutgers University for their early and ongoing expert advice and support. A special thank to Chris Weller and Ben Fino-Radin whose contributions to the project design, development and technical support were invaluable. I would also like to thank the graduate assistants and student volunteers who contributed on everything from data analysis to visualization – Jared Negley, Sara Rubinow, Genevieve Szablya, and Jeff Walloch. I am also grateful to Quinn Lai for her administrative support.

Bibliography

Berners-Lee, T. (2006). Linked Data - Design Issues. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Brickley, D., & Miller, L. (2010, August 9). FOAF Vocabulary Specification 0.98. Retrieved from http://xmlns.com/foaf/spec/

Dumbill, E. (2002, June 1). XML Watch: Finding friends with XML and RDF. Retrieved from http://www.ibm.com/developerworks/xml/library/x-foaf.html

Fitzgerald, M. (2008). *Jazz archives in the United States* (Master's thesis). Retrieved from https://cdr.lib.unc.edu/

Hannemann, J., & Kett, J. (2010). Linked Data for libraries. *World Library and Information Congress 2010: 76th IFLA General Conference and Assembly, Gothenberg*. Retrieved from http://www.ifla.org/files/hq/papers/ifla76/149-Hannemann-en.pdf

Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the web into a global data space*. San Rafael, CA: Morgan & Claypool.

Malmsten, M. (2008). Making a library catalog part of the Semantic Web. *DC-2008: Proceedings of the International Conference on Dublin Core and Metadata Applications, Berlin*, 146-152. Retrieved from http://libris.kb.se/resource/bib/11306748

Oomen, J., & Aroyo, L. (2011). Crowdsourcing in the cultural heritage domain: opportunities and challenges. *Initiatives, 29*(July), 138-149.

Orlandi, F., & Passant, A. (2011). Modelling provenance of DBpedia resources using Wikipedia contributions [Special issue]. *Web Semantics: Science, Services and Agents on the World Wide Web*, *(9)*2, http://dx.doi.org/10.1016/j.websem.2011.03.002

Pattuelli, M.C., Weller, C., & Szablya, G. (2011). Linked Jazz: An exploratory pilot. *DC-2011: Proceedings of the International Conference on Dublin Core and Metadata Applications, The Hague,* 158-164. Retrieved from http://dcpapers.dublincore.org/index.php/pubs/article/view/3637

Reck, R. P., Sall, K. B., & Swanbeck, W. A. (2011). Determining the impact of Eric Clapton on music using RDF graphs: selected challenges of semantics across and within datasets. *Proceedings of Balisage: The Markup Conference 2011, Montréal, Canada, 7. doi:10.4242/BalisageVol7.Sall*

Vila Suero, D. (Ed.). (2011). Library Linked Data Incubator Group: Use Cases. Retrieved from the W3C Incubator Group Report website: http://www.w3.org/2005/Incubator/lld/XGR-lld-usecase-20111025/

Appendices

Appendix I. List of interview transcripts

Note: The list includes the full sample of 50 interview transcripts, including file sources and DBpedia identifiers. The pool of 12 transcripts from the first pilot are marked in red.

No.	Artist Name	File Source	DBPedia URI
1	Clark Terry	Hamilton_ClarkTerry.pdf	http://dbpedia.org/resource/Clark_Terry
2	Lionel Hampton	Hamilton_LionelHampton_04.18.2011.pdf	http://dbpedia.org/resource/Lionel_Hampton
3	Marian McPartland	Hamilton_MarianMcPartland_04.18.2011.pdf	http://dbpedia.org/resource/Marian_McPartland
4	Slide Hampton	Hamilton_SlideHampton.pdf	http://dbpedia.org/resource/Slide_Hampton
5	Chico Hamilton	Smithsonian_ChicoHamilton_04.18.2011.pdf	http://dbpedia.org/resource/Chico_Hamilton
6	Danny Barker	Smithsonian_DannyBarker.pdf	http://dbpedia.org/resource/Danny_Barker
7	Doc Cheatham	Smithsonian_Doc_Cheatham_JOHP.pdf	http://dbpedia.org/resource/Doc_Cheatham
8	Frank Foster	Smithsonian_FrankFoster.pdf	http://dbpedia.org/resource/Frank_Foster_%28musician%29
9	Artie Shaw	Smithsonian_joh_artie_shaw_04.18.2011.pdf	http://dbpedia.org/resource/Artie_Shaw
10	Jane Jarvis	Hamilton_JaneJarvis.pdf	http://dbpedia.org/resource/Jane_Jarvis
11	Mary Lou Williams	Rutgers_mlw_complete_FINAL_CW.pdf	http://dbpedia.org/resource/Mary_Lou_Williams
12	Billy Taylor	Smithsonian_Billy_Taylor.pdf	http://dbpedia.org/resource/Billy_Taylor_%28jazz_bassist%29
13	Benny Powell	Hamilton_Benny_Powell.pdf	http://dbpedia.org/resource/Benny_Powell
14	Ron Carter	Hamilton_RonCarter_04.18.2011.pdf	http://dbpedia.org/resource/Ron_Carter
15	Buddy Tate	Hamilton_Buddy_Tate.pdf	http://dbpedia.org/resource/Buddy_Tate_%28musician%29
16	Buster Williams	Hamilton_Buster_Williams.pdf	http://dbpedia.org/resource/Buster_Williams
17	Etta Jones	Hamilton_Etta_Jones.pdf	http://dbpedia.org/resource/Etta_Jones
18	Herbie Hancock	Hamilton_Herbie_Hancock.pdf	http://dbpedia.org/resource/Herbie_Hancock
19	James Moody	Hamilton_James_Moody.pdf	http://dbpedia.org/resource/James_Moody_%28saxophonist%29
20	Joe Williams	Hamilton_JoeWilliams.pdf	http://dbpedia.org/resource/Joe_Williams_%28jazz_singer%29
21	Milt Hinton	Hamilton_Milt_Hinton.pdf	http://dbpedia.org/resource/Milt_Hinton
22	Ray Drommond	Hamilton_Ray_Drummond.pdf	http://dbpedia.org/resource/Ray_Drummond
23	Benny Glson	Smithsonian_Benny_Golson.pdf	http://dbpedia.org/resource/Benny_Golson
24	Buddy DeFranco	Smithsonian_Buddy_Defranco.pdf	http://dbpedia.org/resource/Buddy_DeFranco
25	Dave Brubeck	Smithsonian_DaveBrubeck.pdf	http://dbpedia.org/resource/Dave_Brubeck
26	Bob Haggart	Hamilton_Bob_Haggart.pdf	http://dbpedia.org/resource/Bob_Haggart
27	Jimmy Scott	Smithsonian_Jimmy_Scott.pdf	http://dbpedia.org/resource/Jimmy_Scott
28	J. J. Johnson	Smithsonian_JJ_Johnson.pdf	http://dbpedia.org/resource/J._J._Johnson
29	Louie Bellson	Smithsonian_LouieBellson.pdf	http://dbpedia.org/resource/Louie_Bellson
30	Quincy Jones	Smithsonian_Quincy_Jones.pdf	http://dbpedia.org/resource/Quincy_Jones
31	Melba Liston	Smithsonian_Melba_Liston.pdf	http://dbpedia.org/resource/Melba_Liston
32	Roy Haynes	Smithsonian_Roy_Haynes.pdf	http://dbpedia.org/resource/Roy_Haynes
33	Gunther Schuller	Smithsonian_Gunther_Schuller.pdf	http://dbpedia.org/resource/Gunther_Schuller
34	Andy Kirk	University_of_Michigan_Andy_Kirk.pdf	http://dbpedia.org/resource/Andy_Kirk
35	Billy Eckstine	University_of_Michigan_Billy_Eckstine.pdf	http://dbpedia.org/resource/Billy_Eckstine
36	Johnny Griffin	University_of_Michigan_Johnny_Griffin.pdf	http://dbpedia.org/resource/Johnny_Griffin
37	McCoy Tyner	University_of_Michigan_McCoy_Tyner.pdf	http://dbpedia.org/resource/McCoy_Tyner
38	Roy Eldridge	University_of_Michigan_Roy_Eldridge.pdf	http://dbpedia.org/resource/Roy_Eldridge
39	Sam Rivers	University_of_Michigan_Sam_Rivers.pdf	http://dbpedia.org/resource/Sam_Rivers
40	Gerald Wilson	Smithsonan_Gerald_Wilson_.pdf	http://dbpedia.org/resource/Gerald_Wiggins
41	Delfeayo Marsalis	Smithsonian_Delfeayo_Marsalis.pdf	http://dbpedia.org/resource/Delfeayo_Marsalis
42	Abbey Lincoln	Smithsonian_Abbey_Lincoln.pdf	http://dbpedia.org/resource/Abbey_Lincoln
43	Yusuf Lateef	Smithsonian_Yusef_Lateef.pdf	http://dbpedia.org/resource/Yusef_Lateef
44	Phil Woods	Hamilton_Phil_Woods.pdf	http://dbpedia.org/resource/Phil_Woods
45	Gerald Wiggins	Hamilton_Gerald_Wiggins.pdf	http://dbpedia.org/resource/Gerald_Wiggins

46	Benny Waters	Hamilton_Benny_Waters.pdf	http://dbpedia.org/resource/Benny_Waters
47	Ed Shaughnessy	Hamilton_Ed_Shaughnessy.pdf	http://dbpedia.org/resource/Ed_Shaughnessy
48	Eddie Marshall	Hamilton_Eddie_Marshall.pdf	http://dbpedia.org/resource/Eddie_Marshall
49	Oscar Peterson	Hamilton_Oscar_Peterson.pdf	http://dbpedia.org/resource/Oscar_Peterson
50	Harold Ousley	Hamilton_Harold_Ousley.pdf	http://dbpedia.org/resource/Harold_Ousley

Appendix II. Tables depicting connections

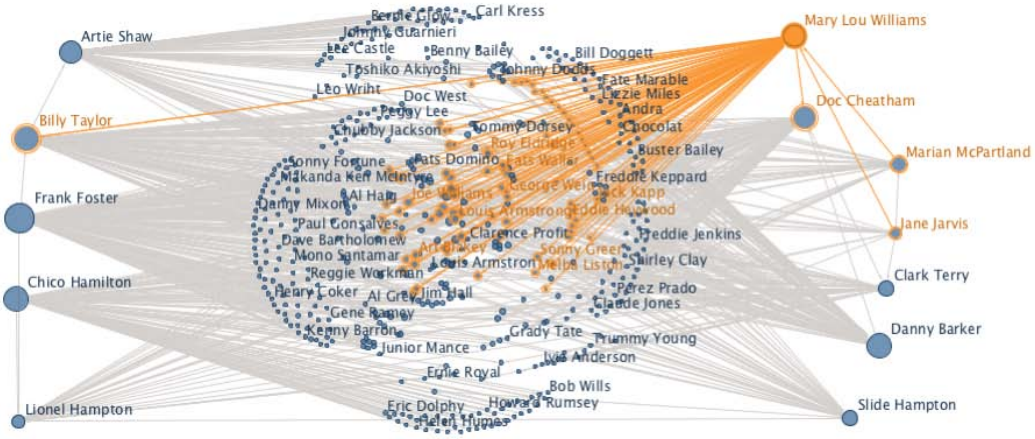
Artist	Total Connections
Artie Shaw	87
Billy Taylor	122
Chico Hamilton	108
Clark Terry	40
Danny Barker	112
Doc Cheatham	96
Frank Foster	183
Jane Jarvis	15
Lionel Hampton	26
Marian McPartland	34
Mary Lou Williams	85
Slide Hampton	44
TOTAL	**952**

Table 1. The number of connections among interviewees and jazz artists from the directory.

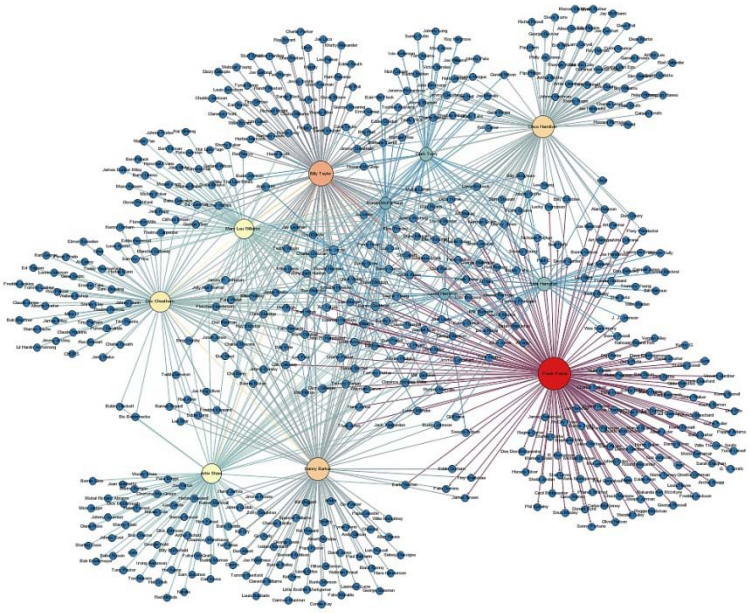
	Artie Shaw	Billy Taylor	Chico Hamilton	Clark Terry	Danny Barker	Doc Cheatham	Frank Foster	Jane Jarvis	Lionel Hampton	Marian McPartland	Mary Lou Williams	Slide Hampton
Artie Shaw		25	9	7	24	14	21	4	8	11	19	11
Billy Taylor			33	14	37	33	42	8	12	16	39	15
Chico Hamilton				12	19	12	33	3	10	10	14	11
Clark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

Table 2. Interviewees' shared connections.

Appendix III. Social network visualizations from first sample



1. Visualization of the network showing the 12 interviewees.



2. Visualization of the network structure.