

Personal Name Vocabularies as Linked Open Data: A Case Study of Jazz Artist Names

Journal of Information Science

XX (X) pp. 1-8

© The Author(s) 2012

Reprints and Permissions:

sagepub.co.uk/journalsPermissions.nav

DOI: 10.1177/0165551512455989

jis.sagepub.com



M. Cristina Pattuelli

School of Information and Library Science, Pratt Institute, New York, New York

Abstract

This paper describes the process of constructing a vocabulary of personal names of jazz artists in the form of Linked Open Data (LOD). Created as a name directory to support the development of the Linked Jazz project,¹ it provides a case study that demonstrates the value and the challenges of developing a domain-specific vocabulary tool that draws upon the semantics of DBpedia, a major LOD dataset. The paper also addresses possible strategies for enhancing the directory to make it a more robust personal name vocabulary.

Keywords

Linked Open Data; personal name vocabularies; DBpedia

1. Introduction

As the Linked Open Data (LOD) initiative continues to expand, vocabularies of various types and levels of complexity are becoming essential components of this new global and open web environment. LOD vocabularies, from metadata element sets to value vocabularies, have the potential to play an important role as semantic hubs that facilitate interlinking between heterogeneous datasets, assist with information integration, and help create new paths to information discovery. While LOD 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 case study presented in this paper contributes to LOD efforts in the cultural heritage sector through the development of a personal name vocabulary of jazz artists for LOD applications.

This personal name vocabulary was created as part of the Linked Jazz project, the goal of which is to develop a LOD dataset to represent and visualize the network of relationships among jazz musicians as described in archival documents. The Linked Jazz project has been developed in a prototype style with multiple phases, sub-phases, and iterative steps. One of the first phases of the project was the creation of the Linked Jazz Name Directory, a directory consisting of personal names of jazz artists. The Linked Jazz Name Directory was to be used as a reference tool for performing name matching and extraction from archival documents. DBpedia, a major LOD dataset, served as the source of name semantics. The development of the directory progressed through rounds of testing and subsequent adjustments and is currently in the process of being developed into a vocabulary enhanced with name authorities.

This paper begins with a description of the iterative process of building the personal name directory. It continues with a discussion of the final outcome and then addresses strategies for enhancing the directory and developing it into a more robust and comprehensive personal name vocabulary.

¹ <http://www.linkedjazz.org>

Corresponding author:

M. Cristina Pattuelli, School of Information and Library Science, Pratt Institute, 144 W. 14th Street, New York, NY 10011
mpattuel@pratt.edu

2. Related work

Vocabularies lie at the core of LOD functionality. They are key tools in facilitating the integration and reuse of content because of their capability to reduce semantic ambiguity and effectively support interlinking among different datasets. Because the development of LOD “depends on the ability of its practitioners to identify, re-use, or connect to already available datasets and data models” [1], vocabularies that support such functions are essential.

Various subject-specific LOD vocabularies have been developed in recent years and have been aligning data from their subject-specific datasets with other vocabularies. One example is AGROVOC, a “multilingual structured thesaurus created by [the Food and Agriculture Organization of the United Nations] FAO and the Commission of the European Communities since 1980 covering the fields of food, agriculture, forestry, fisheries, and other related domains” which is used worldwide for improved searching and indexing of subject-specific data [2]. AGROVOC is published as LOD and linked to twelve vocabularies, including EUROVOC, Library of Congress Subject Headings (LCSH), and DBpedia [3].

Projects that specifically focus upon the creation of LOD personal name vocabularies are somewhat scarcer. An example is the Social Networks and Archival Context project (SNAC), which seeks to enhance access to humanities resources by linking historic persons to archival descriptions, library catalogues, and authority files [4]. After pulling personal names, along with other information, from archival records, the SNAC project has matched these names with VIAF preferred and alternate names [4]. Another set of similar projects are those attempting to create unique identifiers for scholars and researchers from different countries to facilitate attribution for their contributions as well as to support name disambiguation. The Names Project, for example, is in the process of creating a name authority series for UK repositories which would allow for precise and persistent identification of some 45,000 of the UK’s top researchers whose work has been published in journal articles and thus does not fall under the authority control of libraries [5]. Other name authority projects with a similar focus include Open Researcher and Contributor ID (ORCID) [6] and LATTES in Brazil [7].

These projects exist in the greater context of the widespread conversion of library data to LOD and the resulting linking that has occurred among libraries and other vocabularies. Libraries have traditionally overseen the creation and implementation of controlled vocabularies to assist in information access and retrieval. In the last few years, many library and information professionals have been calling for the implementation of LOD in controlled vocabularies, arguing that this is an essential step in keeping this rich information relevant in an LOD environment [8, 9]. Libraries that have headed the move to LOD include the Norwegian National Library,² AMICUS,³ the Swedish National Library’s LIBRIS,⁴ the Hungarian National Library,⁵ and NTNU LOD Authorities.⁶ Now, the Virtual International Authority File (VIAF) and the Library of Congress’ controlled vocabularies are also available as LOD, opening immense opportunity for linking and further increasing the LOD environment. The Linked Jazz project situates itself at this juncture, bringing data otherwise hidden in archival records into the open information space of LOD and exploring the possibility of integrating a subject-specific name directory with name authorities to contribute a vocabulary of jazz artist names to the LOD community.

The role of controlled vocabularies, including authority files, has yet to be fully explored in the new context of LOD, in which new functions such as interlinking and data integration become key to discovery and navigation [10, 11]. As Wolven [12] points out, the library community has extensive experience in employing authority files as an effective method to identify people and corporate bodies as well as to link variant forms of names within their catalogues. As the web-as-discovery-space becomes increasingly global across community and institutional boundaries, it also becomes more heavily populated with names in all their diverse forms. To this end, it becomes essential to expand the traditional notion of identity and authority and to envision what Wolven calls a “broader framework for identification” [12].

2.1. Background

The creation of a directory of jazz artist names was one of the first steps in the development of the Linked Jazz project. The directory was intended to support the process of identifying personal names of jazz artists in textual documents,

² <http://www.nb.no/english>

³ <http://www.collectionscanada.gc.ca/amicus/index-e.html>

⁴ <http://libris.kb.se/?language=en>

⁵ http://nektar2.oszk.hu/librivision_eng.html

⁶ <http://www.ntnu.edu/ub>

specifically interview transcripts, through string matching. Under the best practice of reusing semantics whenever possible, we began investigating suitable sources of semantics that were already available. The Library of Congress Name Authority File (LC/NAF)⁷ was a natural candidate for name value sources due to the quality of the data and the availability as LOD. For similar reasons, the Virtual Internet Authority File (VIAF)⁸ was also considered. As general lexical resources, however, they both offer only limited coverage of names in the specific domain of jazz. In fact, they are both primarily focused on the names of authors of works held in libraries and, as such, lack the level of specificity required for contributing to the name directory. The traditional function of the library catalogue and the notion of literary warrant are the driving principles behind library name authorities. The IFLA Statement of International Cataloguing Principles points out the limits of using these principles as the basis of name authorities in the context of the web and the necessity of a broader approach that contains any personal name including those associated with journal articles and web pages [13]. Another compelling reason that prevented us from using library authorities was the impossibility of discerning and filtering out the names of jazz artists only.

Domain-specific lexical tools and jazz artist discographies, such as Michael Fitzgerald's Jazz Discography⁹ and the All Music Guide to Jazz [14], were also investigated, but did not fulfil essential requirements such as being publicly available or providing information in a LOD-compatible format.

The type of name value data we needed had to be in the form of RDF statements including the artist's name paired with a Uniform Resource Identifier (URI). The URI is, along with the Resource Description Framework (RDF), one of the building blocks of Linked Data technology [15]. The selection of sources of URIs was narrowed down to two major LOD datasets: MusicBrainz¹⁰ and DBpedia. While both datasets are extensive, heavily used, and stable, DBpedia was ultimately selected as the primary source of semantics because its URIs are human-readable and easier to query.

DBpedia is the result of a community effort to extract structured information from Wikipedia and make it openly available on the web [16]. DBpedia offers easy query access to large amounts of structured content via a SPARQL endpoint, a protocol service that enables querying SPARQL,¹¹ the standard query language for LOD. Although it is a general dataset, DBpedia also describes types of entities, including persons and places, in ways that make it suitable to serve as a reference value vocabulary for data mining and reuse [1].

2.2. Construction

As a first step in the construction of the directory, we queried the English language version of DBpedia 3.6 for jazz artist name values using SPARQL. The query process required several iterative rounds of assessment of results and subsequent adjustments. Indeed, various SPARQL queries had to be executed before the results reached a satisfactory level of comprehensiveness. For example, our original SPARQL query made use of `foaf:name`, which appeared the most suitable property to search for personal names. This search, however, yielded only a small subset of name values. The query was then expanded to include additional properties such as `rdfs:label` and `dbpprop:name` that significantly increased the number of results.

To be able to retrieve all of the names of jazz artists and only these names, we employed domain-specific entities and properties. Identifying all the semantic constructs contained in the DBpedia knowledge base that were pertinent to our domain and suitable for queries is in itself an interesting, albeit challenging, task. The data model underlying DBpedia is rather opaque, and a map of classes and properties populating the dataset is not immediately evident. Various classification schemes including the DBpedia Ontology¹² and Wikipedia categories are used to organize DBpedia data, but lack of consistency and semantic coherence still persists in the knowledge base [17].

We initially proceeded in a heuristic way by employing representation constructs that would seem relevant to jazz including `dbpedia:MusicalArtist` and `dbpedia:Jazz`. Random tests were conducted to assess accuracy and inclusiveness. The tests compared the results of our queries with entries in name authorities and jazz-specific databases. They revealed that the form of the names was generally correct and consistent with the LC/NAF as well as with entries in major jazz databases including The Jazz Discography by Tom Lord [18].

Omissions of major jazz musicians' names, however, were detected during the tests. In a few cases, names were missing because there was no corresponding Wikipedia entry for an artist and therefore no instance of a URI in

⁷ <http://id.loc.gov/authorities/names.html>

⁸ <http://viaf.org/>

⁹ www.JazzDiscography.com

¹⁰ <http://musicbrainz.org/>

¹¹ <http://www.w3.org/TR/rdf-sparql-query/>

¹² <http://wiki.dbpedia.org/Ontology>

DBpedia. There were also cases in which the artist had a Wikipedia entry, but no URI was found in the DBpedia knowledge base. The source of this type of omission was more challenging to detect and usually attributable to different types of human error.

The literature has extensively discussed issues of accuracy in the content provided by Wikipedia contributors. In a comparison of Wikipedia entries with those in *Encyclopedia Britannica*, Giles [19] shows that both sources suffer from major errors such as misinterpretation of concepts. Wikipedia is, however, more prone to inconspicuous errors including omissions or misleading assertions. Empirical studies discussed in Fallis [20] recognize the overall reliability of Wikipedia, but identify as a major source of error the incompleteness of the entries rather than their inaccuracy [21, 21].

One part of Wikipedia that has a major impact on the content of DBpedia knowledge base is the infobox. Typically located on the upper right-hand side of a Wikipedia page, infoboxes include structured information and serve as the main source of information from which DBpedia extracts data values to populate its knowledge base. Wu and Weld [23] point out that many of the errors found in the infobox stem from the fact that infobox templates are designed manually through a 'copy and edit' process which can lead to inaccuracies. They go on to explain that errors may also arise during the conversion of articles from one infobox class to another, a frequent occurrence as authors continually reclassify articles.

In the context of our project, the case of "Steve Coleman" offers a revealing example of the issues that can arise as a result of faulty or missing infoboxes. When trying to understand why this name was absent from our directory, we realized that this musician had an entry in Wikipedia, but no infobox on his page.¹³ This omission shows that the infobox is the chief source of DBpedia personal name values and thus critical in determining the completeness of the directory.

Understanding the rules underlying the use of the infobox helped to make sense of other types of omissions. While deemed fairly complete for the names by which artists were usually known, our directory only erratically included birth names and other name variants. This is primarily due to the fact that the "name" of an artist (or group) is one of two mandatory infobox fields that a Wikipedia author is required to provide when supplying information to the "musical artist" infobox template.¹⁴ The infobox template also includes fields for additional names such as the birth name and aliases. These fields, however, are all optional and are thus supplied inconsistently. This inconsistency is also reflected in our directory.

Human error is also responsible for the incorrect classification of musicians in Wikipedia, which in turn affects the quality of the DBpedia dataset. Multiple revisions of the queries were necessary to overcome issues related to inaccurate categorization of jazz musicians. This was the case with "Count Basie," the name of the famous jazz pianist was returned by a query that was expanded to include `dbpedia:Swing_music`, `dbpedia:Big_band_music` and `dbpedia:Piano_blues`, but surprisingly, his name did not come up when we just employed `dbpedia:Jazz`.

In fact, having our focus on the domain of jazz presented further challenges. Jazz is a type of music that defies definition. It is a quintessential example of a cross-genre type of music that spans and combines multiple sub-genres and styles from blues to bebop. These considerations led us to expand SPARQL queries to include related music styles and sub-genres such as *Swing_music*, *Jazz_fusion*, or *Skiffle*. This expansion allowed us to acquire name values that were not obtained when simply using the entity *Jazz*.

It should be noted that the query expansion, as expected, decreased precision and retrieved a number of names extraneous to our domain, including rappers and DJs. After query expansion, the directory increased considerably from 2,676 triples describing 2,367 individuals to 17,559 triples (+557%) describing 6,444 individuals (+172%).

For the immediate goals of the Linked Jazz project, comprehensiveness was valued over accuracy. As discussed earlier, the directory was intended to serve as a matching tool to perform the identification and extraction of jazz artists' names from interview transcripts. Greater inclusiveness was therefore valued in order to maximize matches. For the same reason, we could tolerate a good deal of redundancy because it would not affect the matching outcome.

Various cycles of data curation were later conducted to eliminate false matches and stray errors. Deduplication was also performed which led to the removal of a significant amount of redundant name values. Eliminating those data resulted in a dataset that, as of June 1, 2012, contained 3,389 triples corresponding to an equal number of unique URIs. An example of a triple from the name directory expressed in N-Triples format is shown in Listing 1.

¹³ This example refers to the results of the data extraction conducted in August 2011. An infobox for Steve Coleman has since been added to his Wikipedia entry (5 May 2012).

¹⁴ The other field is "background" (http://en.wikipedia.org/wiki/Template:Infobox_musical_artist).

```
<http://dbpedia.org/resource/Duke_Ellington>  
<http://xmlns.com/foaf/0.1/name> "Duke Ellington"
```

Listing 1. Example of a literal triple from the name directory.

3. Discussion

The name directory succeeded in its intended purpose to assist with name matching and extraction as part of a greater effort to build a dataset of triples representing connections among jazz artists. Overall, DBpedia was an adequate and appropriate source for personal name values. The primary advantage of using DBpedia lay in the richness of its knowledge base and in the depth of its domain specificity as compared to the more general Library of Congress vocabularies. From a sustainability viewpoint, another advantage is DBpedia's capacity to dynamically incorporate new names by mirroring the evolving nature of Wikipedia, which would in turn help to keep the directory current.

On the other hand, there are a number of limitations to using DBpedia as a source of data. As discussed earlier, human error, including incorrect classifications and missing content, has to be taken into consideration when using data extracted from a community-driven tool like Wikipedia. The challenges of relying on user-generated data sets where the quality of data may vary considerably have been discussed by Reck, Sall and Swanbeck [24]. Their study looked at the impact of Eric Clapton on pop music by querying semantic web datasets including DBpedia and MusicBrainz.

More problematic shortcomings stem from the underlying data model of DBpedia. The literature has addressed, although only in passing, issues related to the lack of a formal structure in DBpedia. This issue affects the quality of its data and has potentially negative implications for applications that rely on these data [25, 26].

4. Future work

4.1. From directory to vocabulary

While developed for a particular research purpose, the Linked Jazz Name Directory is a unique knowledge tool that has the potential to be used well beyond the goals of the project for which it was built. As discussed earlier, vocabularies, including personal name vocabularies, can work as semantic hubs to facilitate cross-dataset interlinking and data integration and thus enable new forms of information access and discovery.

In the context of the Linked Jazz project, data curation and maintenance are an ongoing process intended to keep the directory current and improve its quality. Additional steps can be taken to strengthen the directory and help it mature as a vocabulary tool for use by the larger LOD community. One such step would be to make it more comprehensive through the inclusion of alternate forms of names. The directory currently contains a limited set of name variants. For example, variants can be found for Miles Davis (e.g., Miles Dewey Davis) and Willie "The Lion" Smith (e.g., The Lion), but not for Duke Ellington. While alternate names, including aliases and birth names, are part of the Wikipedia infobox template, they are not supplied reliably by Wikipedia authors. This is not surprising considering that name variant fields are not mandatory in the "musical artist" infobox. While the lack of name variants did not affect our matching process in any significant way, this issue would need to be worked out should the directory be made available to the LOD community in an open publishing format.

One way to address the currently limited and inconsistent presence of name variants is to complement the name URIs in our directory with name identifiers from name authorities. The authors of DBpedia have highlighted some of the benefits of interlinking DBpedia with traditional knowledge tools such as hand-crafted ontologies [16]. Such interlinking would enable applications to take advantage of the formal semantics of these ontologies to enhance instance data from DBpedia.

Similarly, name authorities can serve to enrich and strengthen the Linked Jazz Name Directory. A crosswalk between identifiers from the Linked Jazz Name Directory and a name authority would provide a suitable method for performing such enhancement. This mapping would allow systems to cross-reference corresponding names and also derive name variants from the name authority that would otherwise be missing in the Linked Jazz Name Directory.

For such a crosswalk, we are considering the Library of Congress Authorities & Vocabularies and VIAF as suitable sources of authority-controlled personal names, despite the limitations of the domain coverage. As mentioned above, both the VIAF and the Library of Congress' controlled vocabularies, including the LC/NAF and the Library of

Congress Subject Headings (LCSH), have been made available as LOD. LC/NAF includes over eight million authoritative descriptions intended to identify persons, organizations, events, places, and titles. VIAF, a joint project of the Library of Congress, the French and German national libraries, and OCLC, has been released under an open license with no specific restrictions on its use. It combines authority records from several national libraries and includes additional authority files such as the Getty Union List of Artist Names.

Suitable methods to perform the crosswalk are currently under investigation. First, we plan to use `owl:sameAs`, a predicate of the OWL Web Ontology Language,¹⁵ for representing equivalence between the names in the directory and the authorities.

Resources are becoming available that assist in the identification of co-references and could help in creating name authority clusters. One of these tools is the `<sameAs>`¹⁶ service that uses `owl:sameAs` to retrieve a list of different URIs that denote the same entity. To specifically collect personal name authority URIs, the National Library of Norway has recently released an authority service that clusters personal name authority data from multiple identifier schemes.¹⁷ Listing 2 shows an example of results from querying the service for “Duke Ellington” that generates an alignment between DBpedia, LC/NAF and VIAF identifiers for the same entity:

```
<rdf:RDF>
  <rdf:Description rdf:about="http://dbpedia.org/resource/Duke_Ellington"/>
    <owl:sameAs rdf:resource="http://viaf.org/viaf/66651610"/>
    <owl:sameAs rdf:resource="d.loc.gov/authorities/names/n50080187.html"/>
  </rdf:Description>
</rdf:RDF>
```

Listing 2. Graph representing an example of alignment for “Duke Ellington.”

In the case of our directory, this type of alignment would enable the cross-referencing not only of preferred names across authorities, but also of any alternate name associated with these authority name entries. The Linked Jazz Name Directory would thus be enhanced with alternate forms of names either through interlinking or by importing these variants into our directory. Both LC/NAF and VIAF data have been modelled in SKOS (Simple Knowledge Organization System) [27], an RDF-based data framework to represent vocabularies in the context of the Semantic Web. Listing 3 shows a SKOS representation taken from the VIAF authority record for Duke Ellington, which includes a list of controlled and alternate headings from various other authority files. Here, the lexical labelling properties `skos:prefLabel` and `skos:altLabel` employed to represent controlled and alternate headings. The matching property `skos:exactMatch` is also included that serves to map to concepts from different schemes.

```
<skos:Concept rdf:about="http://viaf.org/viaf/sourceID/LC%7Cn++50080187#skos:Concept">
  <skos:inScheme rdf:resource="http://viaf.org/authorityScheme/LC"/>
  <skos:prefLabel>Ellington, Duke, 1899-1974</skos:prefLabel>
  <skos:altLabel>Ellington, Edward Kennedy, 1899-1974</skos:altLabel>
  <skos:altLabel>Ellington, Di□u□k, 1899-1974</skos:altLabel>
  <skos:altLabel>Turner, Joe, 1899-1974</skos:altLabel>
  <skos:altLabel>Greer, Sonny, 1899-1974</skos:altLabel>
  <skos:altLabel>Ellington, Obie Duke, 1889-1974</skos:altLabel>
  <skos:altLabel>Duke, Obie, 1889-1974</skos:altLabel>
  <skos:exactMatch rdf:resource="http://id.loc.gov/authorities/names/n50080187"/>
```

Listing 3. Graph representing part of the VIAF record for “Duke Ellington.”

Either the `owl:sameAs` property, as in Listing 2, or the `skos:exactMatch` property, as in Listing 3, could serve as mapping links. The two constructs differ in their degree of formality. While `skos:exactMatch` carries rather loose semantics, `owl:sameAs` expresses a more rigorous notion of logical equivalence. Asserting that an individual is `owl:sameAs` another individual implies that what is stated about one is also true of the other. This could trigger “undesirable entailments” and thus become a source of inconsistency [28]. Current practices of vocabulary mapping, however, employ one or the other property or both. Using both properties is viable because LOD vocabularies allow for the mixing and matching of properties from different schemas and properties with overlapping scope can easily coexist. As Stellato [29] points out, the two properties are based on vocabularies which are more interchangeable than commonly believed and which can be intertwined in a beneficial manner. In the context of our

¹⁵ <http://www.w3.org/TR/owl-guide/>

¹⁶ <http://sameas.org/>

¹⁷ http://data.bibsys.no/data/query_authority.html

project, the use of `sameAs` and `skos:exactMatch` will be evaluated based on the opportunity to reuse semantics already encoded in the authority files, the requirements of the applications, and in accordance with best practices.

Various benefits could be expected from meshing the Linked Jazz Name Directory with authority data to create a robust LOD vocabulary of jazz artists' names. An immediate advantage would be the increased level of comprehensiveness obtained by being able to cross-reference to alternate names. The quality of data would also be better ensured by relying on the trusted data provided by authority files. If our vocabulary were integrated in one view and made browsable via a user interface, the vocabulary could be used to assist data searching and navigation. This method would also carry over some of the beneficial properties of controlled vocabularies including the provision of context, essential to reduce ambiguity derived from homonymy and polysemy. The vocabulary would open up opportunities for interlinking to bibliographic data as well as to external LOD datasets. For example, the authority URIs dereference to bibliographic records and their metadata, making it possible to link a personal name to works associated with that name as well as to other data sources. Personal names could then serve as data points for interlinking to and navigating across multiple datasets and domains.

5. Conclusion

This paper describes the process of creating a LOD directory of personal names using the DBpedia knowledge base as a source of semantics. The paper also discusses methods of vocabulary mapping that could enhance the directory by complementing it with name authorities. As the LOD ecosystem continues to expand, vocabularies have become an integral part of this landscape. As semantic hubs, they facilitate information integration and the interlinking of heterogeneous datasets. The aim of this paper is to increase our understanding of the value and of the challenges that personal name vocabularies hold as Linked Open Data. Their role in the development of the LOD initiative has just begun to be recognized and deserves further exploration.

Acknowledgements

I would like to thank Chris Weller, Leanora Lange, and Sara Rubinow for their invaluable contributions to the project and to this paper.

An earlier version of this paper was presented at the International Conference on Trends in Knowledge and Information Dynamics (ICTK) held in Bangalore, India in July 2012.

Funding

The work presented herein was partly supported by a Online Computer Library Center (OCLC) and Association for Library and Information Science Education (ALISE) Library and Information Science Research Grant.

References

- [1] Isaac A, Waites W, Young J and Zeng, M. *Library Linked Data incubator group: datasets, value vocabularies, and metadata element sets: W3C incubator group report 25 October 2011*, <http://www.w3.org/2005/Incubator/ld/XGR-ld-vocabdataset-20111025/> (2011, accessed June 2012).
- [2] Rajbhandari S and Keizer J. The AGROVOC concept scheme: a walkthrough. *Journal of Integrative Agriculture* 2012. 11(5): 694-699.
- [3] Food and Agriculture Organization of the United Nations (FAO). AGROVOC Linked Open Data, <http://aims.fao.org/standards/agrovoc/linked-open-data> (2012, accessed June 2012).
- [4] Larson R and Janakiraman, K. Connecting archival collections: the Social Networks and Archival Context project. *Research and Advanced Technology For Digital Libraries Lecture Notes in Computer Science* 2011; 6966: 3-14.
- [5] Cross P, Danskin A, Hill A and Needham D. *Names Project, Phase Two Final Report*, <http://ie-repository.jisc.ac.uk/573> (2011, accessed June 2012).

- [6] Open Researcher and Contributor ID (ORCID). 'Welcome to ORCID.' Open Researcher and Contributor ID (ORCID), <http://about.orcid.org> (2012, accessed June 2012).
- [7] CNPq. 'About Lattes database.' Plataforma Lattes, <http://lattes.cnpq.br/english/index.htm> (2012, accessed June 2012).
- [8] Coyle, K. Library data in the web world. *Library Technology Reports* 2010; 46(2): 5-11.
- [9] Harper CA and Tillett BB. Library of Congress controlled vocabularies and their application to the Semantic Web. *Cataloging and Classification Quarterly* 2007; 43(3-4): 47-68.
- [10] Tillett B, Dechman L, and McLean L. *Linking to LCSH and LCC: controlled subject headings and classification systems through the web*. Slide lecture presented at the 77th IFLA General Conference and Assembly, San Juan, Puerto Rico, <http://conference.ifla.org/past/ifla77/149-tillett-en.pdf> (2011, accessed June 2012).
- [11] Guenther R. *LC's authorities and vocabularies web service: experimenting with Linked Data*, <http://metro.org/files/287> (2012, accessed June 2012).
- [12] Wolven R. (Speaker). *Why name identifiers* [ASP digital media recording]. NISO March two-part webinar: identifiers: new problems, new solutions. <http://www.niso.org/news/events/2010/nameid> (2010, accessed June 2012).
- [13] Danskin A. *Spelling it all out: FRAD, ISNI, RDA, VIAF automation and the future of authority control*. <http://www.cilip.org.uk/FileDownloadsLibrary/Groups/cig/Spelling%20it%20all%20out.ppt> (2009, accessed June 2012).
- [14] Erlewine M. *All music guide to jazz: the experts' guide to the best jazz recordings*. San Francisco: Backbeat Books, 1998.
- [15] Berners-Lee T. 'Linked Data', *Design issues*, <http://www.w3.org/DesignIssues/LinkedData.html> (2009, accessed June 2012).
- [16] Bizer C, Heath T, and Berners-Lee T. Linked Data: The story so far. *International Journal on Semantic Web and Information Systems* 2009; 5(3): 1-22.
- [17] Pattuelli MC and Rubinow S. Charting DBpedia: Towards a cartography of a major linked dataset. *The International Society for Knowledge Organization (ISKO 2012)*, Mysore, India, August 6-9, 2012.
- [18] Lord T. The jazz discography online (TJD Online), <http://www.lordisco.com> (2012, accessed June 2012).
- [19] Giles J. Internet encyclopaedias go head to head. *Nature*, 2005; 43(8): 900-901.
- [20] Fallis D. Toward an epistemology of "Wikipedia." *Journal of the American Society for Information Science and Technology*, 2008; 59(10): 1662-1674.
- [21] Devgan L, Powe N, Blakey B and Makary M. Wiki-surgery? Internal validity of Wikipedia as a medical and surgical reference. *Journal of the American College of Surgeons*, 2007; 205(3): S76-S77.
- [22] Bragues G. Wiki-philosophizing in a marketplace of ideas: evaluating Wikipedia's entries on seven great minds. *Media Tropes eJournal*, 2009; 2(1): 117-158. <http://www.mediatropes.com/index.php/Mediatropes/article/download/15767/12862>
- [23] Wu F and Weld DS. Automatically refining the Wikipedia infobox ontology. *17th World Wide Web Conference (WWW2008)* 2008: 635-644. <http://www2008.org/papers/pdf/p635-wu.pdf> (accessed June 2012).
- [24] Reck RP, Sall KB, and Swanbeck WA. *Determining the impact of Eric Clapton on music using RDF graphs: selected challenges of semantics across and within datasets*, <http://www.balisage.net/Proceedings/vol7/print/Sall01/BalisageVol7-Sall01.html> (2011, accessed June 2012).
- [25] Passant A. dbrec: music recommendations using DBpedia. In: Patel-Schneider PF, Pan Y, Hitzler P, Mika P, Zhang L, Pan JZ, et al. (eds.) *Proceedings of the 9th International Semantic Web Conference - ISWC 2010* 2010; 1380: 1-16.
- [26] Thakker DA and Stephens M. Press association images: image retrieval challenges. In: Müller H, Clough P, Deselaers T and Caputo B (eds.) *ImageCLEF: experimental evaluation in visual information retrieval*. New York: Springer, 2010. p. 469.
- [27] Isaac A, Phipps J and Rubin D. *SKOS use cases and requirements*, W3C working group note, <http://www.w3.org/TR/skos-ucr> (2009, accessed June 2012).
- [28] Miles, A and Bechhofer, S. skos:closeMatch, skos:exactMatch, owl:sameAs, owl:equivalentClass, owl:equivalentProperty. *SKOS Simple Knowledge Organization System reference*. <http://www.w3.org/TR/skos-reference/#L4858> (2009, accessed June 2012).
- [29] Stellato, A. Dictionary, thesaurus or ontology? Disentangling our choices in the semantic web jungle. *Journal of Integrative Agriculture* 2012; 11(5): 710-719.