The Knowledge Organization of DBpedia: A Case Study

M. Cristina Pattuelli and Sara Rubinow
School of Information and Library Science, Pratt Institute
New York, NY, USA

1. Introduction

Linked Data extends the traditional web by providing an open and unifying framework for the discovery, integration, and reuse of data. It has the potential to realize the vision of the Semantic Web by promoting interoperability through the interlinking of well-defined machine-readable data. One of the strengths of the Linked Data initiative lies in its technical infrastructure, which is based on a simple set of principles and open standards. These standards include the Uniform Resource Identifier (URI), which serves as the global identifier mechanism, and the Resource Description Framework (RDF), which acts as the model for expressing data in a common format (Berners-Lee, 2009). This lightweight framework is key to lowering the barriers to Semantic Web adoption.

The library and cultural heritage communities are actively participating in Linked Data research and development, particularly in the context of its open license version, Linked Open Data (LOD). LOD is seen as a promising technology for connecting data and metadata from heterogeneous sources and contexts in order to facilitate their integration and sharing (Baker *et al*., 2011).

Major efforts of the library community in LOD development are currently focused on converting bibliographic data, into RDF triples, the building blocks of Linked Data. The purpose of these efforts is to steer library data into a seamless search stream. Becoming part of a single global data space would be a significant departure from the traditional "walled garden" model of the bibliographic universe. Library, archive and museum data would be interlinked with billions of RDF triples from a variety of external datasets, most of which rely on relaxed knowledge structures and approximate semantics.

This new and boundless scenario necessitates a strategic reconsideration of both the nature of knowledge organization systems as well as the role those systems play in the LOD context. In this evolving information environment, knowledge structures are becoming increasingly dynamic and flexible, though often less coherent and consistent.

The literature on Linked Data development has just begun to address the implications of dealing with loosely formalized knowledge structures that

Forthcoming in *Journal of Documentation, (69)*6 (2013). Copyright © Emerald Group Publishing Limited.

produce significant amounts of "noisy" data. For example, issues related to the proliferation of identifiers assigned to the same entity, as well as to the inappropriate use of the OWL[i] identity property when managing co-references, have been discussed (Halpin *et al.*, 2010; Hayes, 2011; Uschold, 2010). A special issue of the *Journal of Web Semantics* (Schlobach and Knoblock, 2012) was recently devoted to the challenges of "dealing with the messiness" of the LOD semantic universe. Nonetheless, there is not yet a substantial enough body of research with which to frame an articulate and cohesive discussion on LOD data quality.

Additional studies are also needed to explore the knowledge models that underpin the major datasets populating the LOD landscape. This line of research will enable application developers, Linked Data publishers, vocabulary maintainers and other users to take full advantage of the functionality of these tools.

To this end, this paper investigates the semantic structure underlying DBpedia,[ii] one of the largest and most densely interlinked LOD datasets. DBpedia's knowledge base offers a wealth of Linked Data semantics, which is freely available for exploration, data querying and consumption. Its data organization, governed by multiple schemas and categorization systems, can appear opaque, making it difficult to query. Because it is not always apparent which semantic constructs are being used, it can be difficult to determine the terms which might be good candidates for queries (Auer*, et al.,* 2007).

With the aim to begin to develop a "cartography" of the DBpedia dataset, we selected the domain of jazz as the application scenario for the analysis.

The domain of music has long been a significant presence in the Linked Data ecosystem, as evidenced by prominent datasets such as MusicBrainz[iii] and the massive amount of RDF statements within DBpedia itself.[iv] Jazz was chosen due to the researchers' familiarity with the subject area, one which has served as the context of an ongoing research intended to apply LOD technology to the jazz history archival resources.[v] While the domain of jazz was useful for providing concrete points of reference, the analysis applies to the entire knowledge structure of the DBpedia.

2. DBpedia

DBpedia, which launched in 2007, and is maintained by the Free University of Berlin, the University of Leipzig and OpenLink Software,[vi] is a multilingual and cross-domain LOD dataset created by extracting structured information from Wikipedia. This information is then made openly available on the web in the form of RDF triples, the building blocks of Linked Data technology. Currently, the English version of DBpedia provides information about over 3.64 million resources, or "things," each referring to an article of Wikipedia and described by properties. DBpedia has become one of the largest repositories of Linked Data semantics. It is densely interlinked with many other open datasets and is heavily used as a semantic hub, as illustrated by the Linking Open Data cloud diagram (Cyganiak and Jentzsch, 2011).

The main source of structured data for DBpedia is the Wikipedia *infobox*, which is located in the upper right-hand corner of a Wikipedia article. Infoboxes present summary information about a Wikipedia article in a standardized format. Other sources of structured data used by DBpedia include article abstracts, category labels, external links, and geo-coordinates. Infoboxes, however, are key to understanding both the origin of the properties that populate the DBpedia knowledge base and to gaining a better sense of the structure underlying DBpedia.

Infoboxes are based on a wide range of templates that contain core information elements related to the subject of Wikipedia articles. More than 6,000 templates exist within English-language Wikipedia,[vii] and that number is constantly changing. Infobox templates are created and reused by Wikipedia contributors who also supply the documentation and the rules that determine their format and use. Although Wikipedia owes its size and comprehensive content to its inherently open structure, its lack of centralized administrative control has implications for the consistency and accuracy of how the information fields within infobox templates are filled as well as how those templates are utilized. As Bizer et al. (2009) note, Wikipedia authors may use different templates to describe the same topic; the same attribute can be named differently in different templates (e.g., `birthPlace` and `placeOfBirth`); and attribute values may be expressed using different formats.

2.1 DBpedia data extraction

The methods for extracting structured data determine the type of semantics that populate the DBpedia knowledge base. Two extraction methods, *generic* and *mapping-based,* are employed to acquire structured information from Wikipedia templates (Bizer *et al.*, 2009). The generic method is algorithm-based and consists of bootstrapping the complete coverage of the infobox content and retaining the same property names used in the infobox. The data obtained through this process are described by properties identified by the `dbpprop` prefix. Because the `dbpprop` properties are not part of any cohesive structure or conceptual model, they populate the dataset with significant redundancy and inconsistency.

In order to increase precision and consistency, an additional extraction process was later adopted, based on manually generating mappings between Wikipedia infobox templates and the DBpedia Ontology.[viii] This method was introduced in an effort to contribute a set of controlled, higher quality data to complement the raw, noisy dataset.

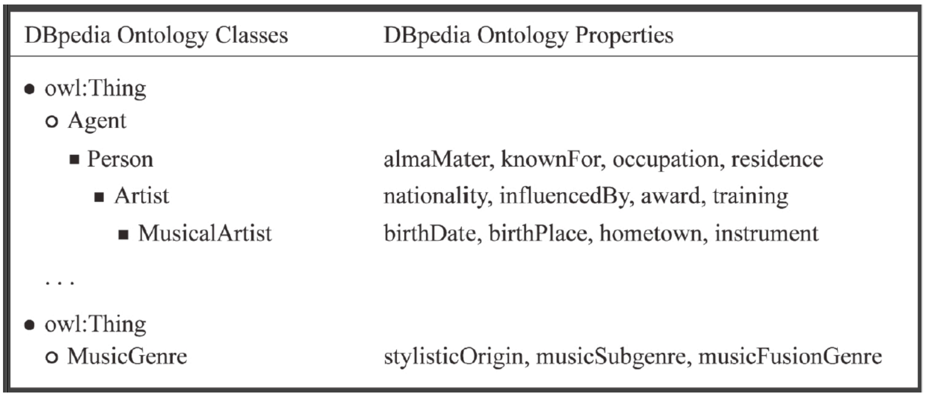
Like Wikipedia, DBpedia grants editing rights to anyone motivated to create manual mappings of Wikipedia infobox templates. Only a small portion of Wikipedia data has been mapped leaving the majority of Wikipedia data yet to be mapped.[ix] To advance this process, a crowdsourcing solution has been adopted to distribute the work across subjects and languages. Soon after editing rights were opened to the community at large in the summer of 2011, the DBpedia team noted the emergence of quality issues due to the proliferation of redundant mapping, expressing concerns that "the ontology is getting unclear and the benefits of standardisation get lost."[x] Achieving a balance between mapping coverage and ontology clarity appears to be an unmet goal.

2.2 The DBpedia Ontology

The DBpedia Ontology is a general ontology that covers multiple domains. It consists of a shallow class hierarchy of 320 classes and was created by manually deriving 170 classes from the most common Wikipedia infobox templates in the English edition. It also includes 750 properties resulting from mapping attributes from within these templates.

The DBpedia Ontology currently describes approximately half of the DBpedia entities (DBpedia, 2011). Because the DBpedia Ontology is built upon infobox templates, its classes suffer from a lack of logical consistency and present significant semantic gaps in their hierarchy. As Damova *et al.* (2010) point out, inconsistencies are shown, for example, in the degree of generality present in its upper-level classes that range from abstract concepts such as *person* or *event* to very specific ones like *protein* or *drug.*

In the course of exploring our application scenario, we discovered that the concept *jazz* is not included in the ontology. The broader domain of music is currently represented by eight classes: *MusicalArtist*, *MusicFestival*, *MusicGenre*, *Musical*, *MusicWork*, *Album, Single, Song.* Each class is part of an unrelated branch of the tree structure that has *owl:Thing* as its root. Table 1 shows a sample of the knowledge structure of the ontology.



DBpedia Ontology Classes	DBpedia Ontology Properties
● owl:Thing	
○ Agent	
■ Person	almaMater, knownFor, occupation, residence
■ Artist	nationality, influencedBy, award, training
■ MusicalArtist	birthDate, birthPlace, hometown, instrument
. . .	
● owl:Thing	
○ MusicGenre	stylisticOrigin, musicSubgenre, musicFusionGenre

Table 1. A sample of DBpedia Ontology classes and properties in the domain of music.

DBpedia data represented by the DBpedia Ontology are identified by the prefix `dbpedia-owl`. As with the uncontrolled `dbpprop` properties, a comprehensive dictionary of `dbpedia-owl` properties is not readily available. Infobox-based `dbpedia-owl` properties can be viewed on an infobox-by-infobox basis through the graphical user interface of the DBpedia Mapping Tool,[xi] a tool which allows the creation and editing of mappings.

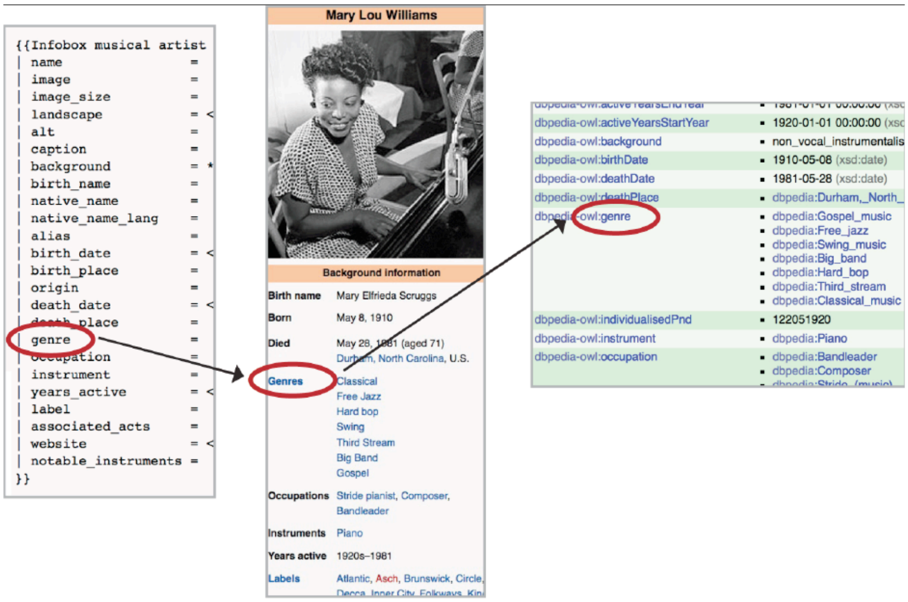
To attempt to understand how DBpedia properties, both ontology-based and uncontrolled, operate in the domain of jazz, we focused on the description of jazz musicians. In the absence of a general inventory of DBpedia properties, we loaded the *musical artist* infobox template[xii] in the DBpedia Mapping Tool in order to discover the full set of `dbpedia-owl` properties employed to describe jazz artists.

Table 2 shows the list of `dbpedia-owl` properties derived from the ontology-based mapping of the *musical artist* infobox template. These properties are presented alongside the corresponding `dbpprop` properties. While the `dbpprop` properties simply mirror the field name used in the Wikipedia infobox template field,[xiii] the `dbpedia-owl` properties reconcile multiple synonyms across infobox templates under a common term.

Properties from mapping-based extraction	Properties from generic extraction
dbpedia-owl:activeYearsEndYear	dbpprop:years_active
dbpedia-owl:activeYearsStartYear	dbpprop:years_active
dbpedia-owl:alias	dbpprop:alias
dbpedia-owl:associatedBand	dbpprop:associated_acts
dbpedia-owl:associatedMusicalArtist	dbpprop:associated_acts
dbpedia-owl:background	dbpprop:background
dbpedia-owl:birthDate	dbpprop:birth_date
dbpedia-owl:birthPlace	dbpprop:birth_place
dbpedia-owl:deathDate	dbpprop:death_date
dbpedia-owl:deathPlace	dbpprop:death_place
dbpedia-owl:genre	dbpprop:genre
dbpedia-owl:hometown	dbpprop:origin
dbpedia-owl:instrument	dbpprop:instrument
dbpedia-owl:occupation	dbpprop:occupation
dbpedia-owl:recordLabel	dbpprop:label

Table 2. Properties describing a *MusicalArtist* entity through the *musical artist* infobox template.

An example of infobox information as mapped to the DBpedia ontology is shown in Figure 1. The left box presents the wiki markup for the *musical artist* infobox template, which is displayed in the infobox (center) as viewed within the Wikipedia article for “Mary Lou Williams.” The properties resulting from the mapping-based extraction process are listed on the corresponding DBpedia resource page.



```
{{Infobox musical artist
| name                =
| image               =
| image_size          =
| landscape           = <
| alt                 =
| caption             =
| background          = *
| birth_name          =
| native_name         =
| native_name_lang    =
| alias               =
| birth_date          = <
| birth_place         =
| origin              =
| death_date          = <
| death_place         =
| genre               =
| occupation          =
| instrument          =
| years_active        = <
| label               =
| associated_acts     =
| website             = <
| notable_instruments =
}}
```

Mary Lou Williams

Background information	
Birth name	Mary Elfrieda Scruggs
Born	May 8, 1910
Died	May 28, 1981 (aged 71) Durham, North Carolina, U.S.
Genres	Classical Free Jazz Hard bop Swing Third Stream Big Band Gospel
Occupations	Stride pianist, Composer, Bandleader
Instruments	Piano
Years active	1920s–1981
Labels	Atlantic, Asch, Brunswick, Circle, Decca, Inner City, Folkways, Kin

dbpedia-owl:activeYearsEndYear	▪ 1981-01-01 00:00:00 (xsc
dbpedia-owl:activeYearsStartYear	▪ 1920-01-01 00:00:00 (xsc
dbpedia-owl:background	▪ non_vocal_instrumentalis
dbpedia-owl:birthDate	▪ 1910-05-08 (xsd:date)
dbpedia-owl:deathDate	▪ 1981-05-28 (xsd:date)
dbpedia-owl:deathPlace	▪ dbpedia:Durham,_North_
dbpedia-owl:genre	▪ dbpedia:Gospel_music ▪ dbpedia:Free_jazz ▪ dbpedia:Swing_music ▪ dbpedia:Big_band ▪ dbpedia:Hard_bop ▪ dbpedia:Third_stream ▪ dbpedia:Classical_music
dbpedia-owl:individualisedPnd	▪ 122051920
dbpedia-owl:instrument	▪ dbpedia:Piano
dbpedia-owl:occupation	▪ dbpedia:Bandleader ▪ dbpedia:Composer ▪ dbpedia:Stride_(music)

Figure 1. Infobox of “Mary Lou Williams” as mapped to the DBpedia ontology.

As discussed earlier, the concept *jazz* is not part of the DBpedia Ontology. *Jazz* is, however, a “thing” present in the DBpedia knowledge base because it is the subject of a Wikipedia article and, as such, is an entity in the DBpedia knowledge base. The Wikipedia entry for *jazz* includes a *music genre* infobox, from which DBpedia can extract structured data. As with all DBpedia entities, the DBpedia properties for *jazz* are displayed on a correspondng HTML DBpedia *resource webpage*, which lists data representing the entity in a human-friendly fashion. The data are presented in the form of property-value pairs.

It would be expected that the DBpedia resource page for *jazz* would include all the `dbpedia-owl` and `dbpprop` properties associated with any other *MusicGenre* entity (e.g., *instrument*, *stylisticOrigin*). Upon reviewing the *jazz* resource page,[xiv] however, none of `dbpedia-owl` infobox properties were found. The properties are present, however, on other DBpedia resource pages for entities that also employ the *music genre* infobox template, such as *Bebop*, *Swing music* and *Rhythm and blues*. This discrepancy does not seem to stem from the data source, since the infobox fields in the Wikipedia entry for *jazz* are filled appropriately. It is more likely that the properties were missed during the extraction process, which is not an uncommon occurrence.

2.3 Knowledge representation tools in DBpedia

In addition to the DBpedia Ontology, the DBpedia knowledge base is governed by a variety of knowledge representation tools including additional classification schemes and RDF vocabularies.

Two classification systems are particularly relevant and consistently used: Wikipedia categories and the YAGO ontology. Wikipedia uses categories to group articles that share similar subjects.[xv] Wikipedia categories are constantly evolving and currently number more than 740,000. Most categories are assigned manually by Wikipedia contributors and can be found listed as links at the bottom of a Wikipedia article. For example, the page about the jazz artist Mary Lou Williams currently displays 28 categories, ranging from *Musicians from Pittsburgh* and *Converts to Roman Catholicism* to *Deaths from bladder cancer*. Each category links to a *category page* containing an alphabetized list of links to other Wikipedia articles assigned to the same category. When available, the category page also lists related parent- and sub-categories.

Wikipedia categories are organized hierarchically, but they are not grounded in a strict taxonomic structure. For example, “Mary Lou Williams” is assigned simultaneously to three categories related to music composition, all of which should be in a subset relation: *American composers*, *Jazz composers,* and *Women composers.* As Suchanek, Kasneci and Weikum (2008) point out, this category system merely mirrors the “thematic structure” of a Wikipedia article rather than representing a cohesive knowledge conceptualization (p. 210).

The Wikipedia organizational system is the result of a collaborative effort that presents advantages as well as weaknesses. On one side, the Wikipedia authoring and editorial process ensures that the categories are continually updated to correspond with article content. On the other side, the system suffers from lack of consistency in its hierarchical structure and what Bizer et al. (2009) call a “rather loose relatedness between articles” (p. 157).

DBpedia makes use of Wikipedia categories to organize its entities. The hierarchical structure of the categories is represented in DBpedia by way of two different properties: `dcterms:subject` and `skos:broader`. The property `dcterms:subject` relates DBpedia entities to their corresponding categories. Each category is then related to its parent category through the `skos:broader` property (Mirizzi, Di Noia, Ostuni, and Ragone, 2012).

The YAGO ontology is another classification system introduced to provide DBpedia data with coherence and structural consistency. YAGO, the most recent iteration of which is called YAGO2, was developed at the Max Planck Institute for Informatics in Saarbrücken, Germany. YAGO was originally derived from the Wikipedia category system using the semantic lexicon WordNet.[xvi] More specifically, YAGO was created through the automatically generated mapping of Wikipedia categories to WordNet synsets (Suchanek, Kasneci and Weikum, 2007). It is a robust and extremely rich classification scheme with a deep hierarchical structure. As Bizer et al. (2009) note, while YAGO is very accurate, it is not immune to the errors and omissions that inevitably occur when ontologies are created using algorithm-based methods. DBpedia uses "YAGO as a taxonomic backbone to connect the facts to a coherent whole" (Suchanek et al., 2008, p. 2).

YAGO class instances are represented as values of the `rdf:type` property. The `rdf:type` property is also paired with values that are class instances of various external ontologies including OWL, schema.org[xvii] and UMBEL.[xviii] The ontology classes serve as connectors that facilitate the interlinking of web content and data, thereby providing context for these data. As Bizer (2009) points out, the interlinking of various ontologies with DBpedia makes it possible for applications to integrate data from all of these sources. Figure 1 shows clusters of DBpedia properties as well as external vocabularies that represent the DBpedia entity "Mary Lou Williams."

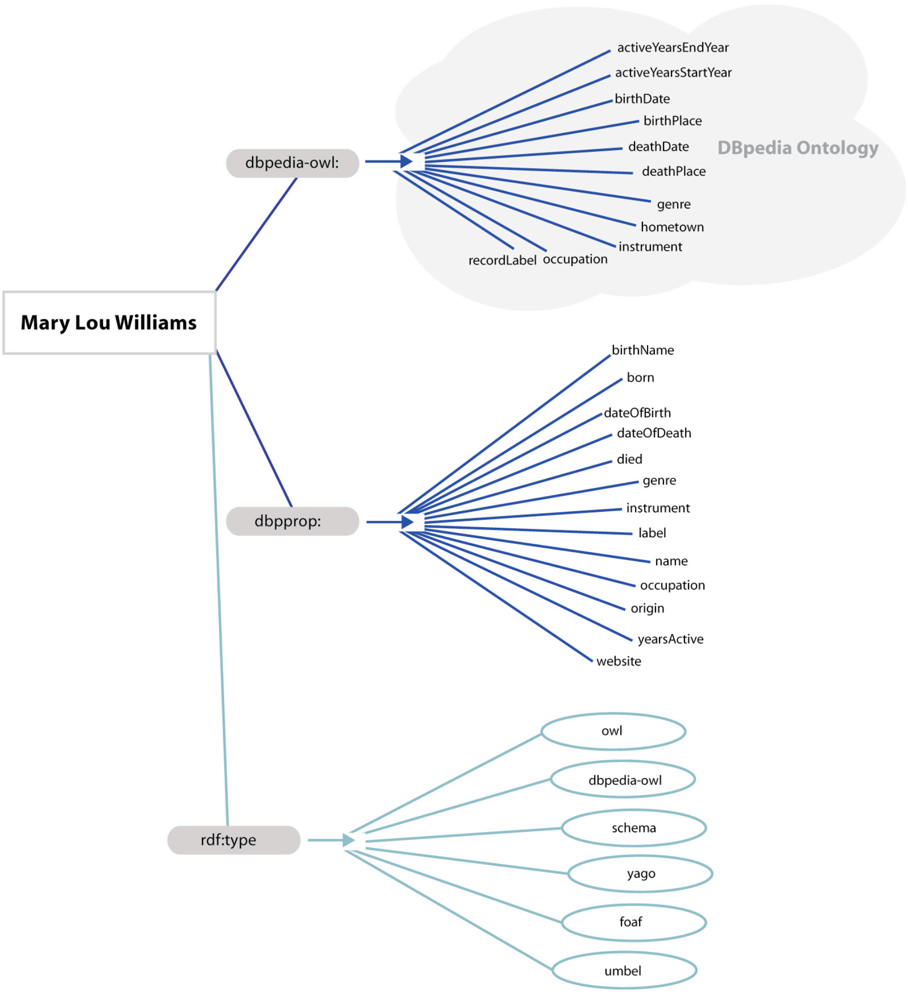


Figure 2. A sample of properties and vocabularies used to represent the DBpedia entity "Mary Lou Williams."

Both RDF native and RDF-based vocabularies are also employed to describe DBpedia data, such as the Dublin Core Metadata Element Set (e.g., `dc:description`) and the Friend of a Friend (FOAF) ontology (e.g., `foaf:name`, `foaf:givenName`, `foaf:page`). The use of properties from different external vocabularies is made possible by the unifying framework provided by RDF. The flexibility of the RDF model allows for the mixing and matching of properties from different namespaces without the need for agreement on the adoption of a specific schema. Multiple vocabularies can be used in a layered fashion, and properties with overlapping scope can coexist in the spirit of "cooperation without coordination" (Wood, 2011), challenging the classical notion of semantic parsimony.

Conclusion

This paper investigates the semantic structure underlying DBpedia, one of the largest datasets in the context of LOD. Our analysis attempts to shed light on a new type of knowledge representation environment that is in a constant state of flux, where different descriptive and classification approaches are employed concurrently.

The semantic constructs and schemes employed in this open and dynamic environment vary significantly in terms of their degree of formalization, stability, cohesiveness and consistency. As such, they challenge our tolerance threshold for data quality and our traditional notion of authority control. Unearthing the knowledge organization of DBpedia increases our practical understanding of the new semantic context provided by LOD. It also has the potential to be useful to LOD users. For example, by having an understanding of the range of entities and properties available, developers could formulate queries with higher precision, rather than use the commonly employed trial and error approach.

This analysis of DBpedia has the potential to open up a new area of research in the broader information and library science community to which the knowledge organization community can make a significant contribution. The decentralized interplay of vocabularies as well as the proliferation of noisy data have implications that have yet to be fully understood. As Dunsire *et al.* (2011) stress, there will be an increasing need to understand and leverage what is perceived as chaos, rather than fearing the presumed end of an existing order.

Notes

[i] http://www.w3.org/TR/owl-features/

[ii] http://dbpedia.org/ (release 3.7)

[iii] http://musicbrainz.org

[iv] Music albums represent the third largest set of RDF statements after persons and places (DBpedia, 3.7).

[v] http://linkedjazz.prattsils.org/

[vi] http://wiki.dbpedia.org/Team

[vii] http://mappings.dbpedia.org/server/statistics/en/ accessed on April 15, 2012.

[viii] http://wiki.dbpedia.org/Datasets

[ix] http://mappings.dbpedia.org/server/statistics/en/

[x] http://mappings.dbpedia.org/index.php/Mapping_Guide

[xi] http://mappings.dbpedia.org/index.php/MappingTool

[xii] http://en.wikipedia.org/wiki/Template:Infobox_musical_artist

[xiii] It should be noted that we excluded the properties `dbpedia-owl:abstract`, `dbpedia-owl:individualisedPnd`, `dbpedia-owl:thumbnail` and `dbpedia-owl:wikiExternalLinks` because they were not based on infobox template mappings

[xiv] http://dbpedia.org/page/Jazz

[xv] http://en.wikipedia.org/wiki/Help:Category

[xvi] http://wordnet.princeton.edu/

[xvii] http://schema.org/

[xviii] http://umbel.org/

References

Auer, S. et al. (2007), "DBpedia: A Nucleus for a Web of Open Data", in Aberer *et al.* (Eds.). *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea,* Springer, Berlin, pp. 722-735.

Auer, S. and Lehmann, J. (2007), "What have Innsbruck and Leipzig in common? Extracting semantics from Wiki content", Lecture *Notes in Computer Science*, Vol. 4519, pp. 503-517.

Baker, T. *et al*. (2011), "Library Linked Data Incubator Group Final Report, W3C Incubator Group Report 25 October 2011", available at: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/ (accessed 20 April 2012.

Berners-Lee, T. (2009), "Linked Data—design issues", available at: http://www.w3.org/DesignIssues/LinkedData.html (accessed 2 February 2012).

Bizer, C. *et. al.* (2009), "DBpedia—a crystallization point for the web of data", *Journal of Web Semantics*, Vol. 7 No. 3, pp. 154-165.

Bizer, C. (2009), "The Emerging Web of Linked Data", Intelligent Systems IEEE, Vol. 24, pp. 87-92.

Cyganiak, R. and Jentzsch, A. (2011), “The Linking Open Data Cloud Diagram”, available at: http://richard.cyganiak.de/2007/10/lod/ (accessed 7 March 2012).

Damova, M., Kiryakov, A., Simov, K. and Petrov, S. (2010), “Mapping the central LOD ontologies to PROTON upper-level ontology”, paper presented at the Fifth International Workshop on Ontology Matching, 7 November, Shanghai, China, available at: http://ceur-ws.org/Vol-689/ (accessed 23 March 2012).

DBpedia. (2011), available at: http://wiki.dbpedia.org/ (accessed 12 February 2012).

dhylandwood. (2011), “SemWeb Elevator Pitch”, [video online] available at: http://youtu.be/VzguipSLUYc (accessed 25 February 2012).

Doerr, M. and Iorizzo, D. (2008), “The dream of a global knowledge network—a new approach”, *ACM Journal on Computers and Cultural Heritage*, Vol. 1 No. 1, pp. 1-23.

Dunsire, G., Hillmann, D. I., Phipps, J. and Coyle, K. (2011), “A reconsideration of mapping in a semantic world”, paper presented at the International Conference on Dublin Core and Metadata Applications, 21-23 September, The Hague, The Netherlands, available at http://dcpapers.dublincore.org/index.php/pubs/article/view/3622/1848 (accessed 3 March 2012).

Halpin, H. *et al.* (2010), “When owl:sameAs isn’t the same: An analysis of identity in linked data”, paper presented at the 9th International Semantic Web Conference, 7-11 November, Shanghai, China, available at: http://iswc2010.semanticweb.org/pdf/261.pdf (accessed 5 March 2012).

Hayes, P. (2011), “On being the same: keynote address”, in Slavic, A. and Civallero, E. (Eds.), *Classification and ontology: formal approaches and access to knowledge: proceedings of the International UDC Seminar, 19-20 September, The Hague, The Netherlands*, Würzburg: Ergon Verlag, pp. 1-2.

Kobilarov, G., Bizer, C., Auer, S. and Lehmann, J. (2009), "DBpedia—a linked data hub and data source for web and enterprise applications", available at: http://www2009.eprints.org/228/ (accessed 30 January 2012).

Mirizzi, R., Di Noia, T., Ostuni, V. C. and Ragone, A. (2012), "Linked Open Data for content-based recommender systems", available at: http://sisinflab.poliba.it/semantic-expert-finding/papers/tech-report-1-2012.pdf (accessed 14 April 2012).

Schlobach, S. and Knoblock, C. A. (Eds.). (2012), "Dealing with the Messiness of the Web of Data" [Special issue], *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 14.

Suchanek, F., Kasneci, G. and Weikem, G. (2007), "YAGO: A core of semantic knowledge unifying WordNet and Wikipedia", paper presented at the Proceedings of the International World Wide Web Conference, 8-12 May, Banff, Canada, available at: http://www2007.org/papers/paper391.pdf (accessed 19 April 2012).

Suchanek, F., Kasneci, G., and Weikem, G. (2008), "YAGO: A large ontology from Wikipedia and WordNet", *Journal of Web Semantics*, Vol. 6 No. 3, pp. 203-217.

Uschold, M. (n.d.) "Proliferation of URIs, managing coreference", available at: http://ontologydesignpatterns.org/wiki/Community:Proliferation_of_URIs,_Managing_Coreference (accessed 7 April 2012).