Linked Data is Merely More Data

Prateek Jain[*], **Pascal Hitzler**[*], **Peter Z. Yeh**[†], **Kunal Verma**[†] and **Amit P. Sheth**[*]
[*]Kno.e.sis Center, Wright State University, Dayton, OH
[†]Accenture Technology Labs, San Jose, CA

Abstract

In this position paper, we argue that the Linked Open Data (LoD) Cloud, in its current form, is only of limited value for furthering the Semantic Web vision. Being merely a weakly linked "triple collection," it will only be of very limited benefit for the AI or Semantic Web communities. We describe the corresponding problems with the LoD Cloud and give directions for research to remedy the situation.

Where We Are

The recent emergence of the "Linked Data" approach for publishing data represents a major step forward in realizing Berners-Lee, Handler and Lassila's original vision of a web that can "understand and satisfy the requests of people and machines to use the web content"[1] – i.e. the Semantic Web (Berners-Lee et al. 2001). This new approach has resulted in the Linked Open Data (LoD) Cloud (Bizer et al. 2007), which includes more than 70 large datasets contributed by experts belonging to diverse communities such as geography, entertainment, and life sciences (Bizer, Heath, and Berners-Lee 2009). Table 1 lists some of the datasets available as a part of LoD Cloud.[2]

The interlinking of these diverse datasets promises a "Web of Data" that will enable users to easily navigate between these datasets in a manner analogous to how users currently navigate from one webpage to another in the "Web of Documents." Moreover, the LoD Cloud can significantly benefit both the AI and Semantic Web communities by enabling new classes of applications and enhancing existing tasks such as querying, reasoning, and knowledge discovery.

To exemplify, *a scientist interested in exploring the relationship between the presence of the spider "Agelenopsis emertoni" and weather patterns*, can do so easily with the help of the LoD Cloud as the Geospecies dataset gives information about the spider *"Agelenopsis emertoni,"* and the interlinking of Geospecies with Geonames makes it easy to explore the different kinds of information related to the locations where it can be found *(Wisconsin)*, the locations where it cannot be found *(Iowa, Minnesota*), and the topography of these regions. Thus, in this scenario, the interlinks might help in identifying and analyzing the topographical patterns related to Iowa and Minnesota which make it difficult for this spider to survive in those regions.

However, the current interlinks between datasets in the LoD Cloud – as we will illustrate – are too shallow to realize much of the benefits promised. If this limitation is left unaddressed, then the LoD Cloud will merely be more data that suffers from the same kinds of problems which plague the Web of Documents, and hence the vision of the Semantic Web will fall short.

What Is Needed

The growing number of datasets available on the LoD Cloud presents a challenge with regards to its usage, since on the one hand datasets such as DBpedia and Freebase offer massive amounts of information from diverse domains, while on the other hand there is no formal description of these or any other LoD Cloud components or their interlinking. We believe that the LoD Cloud can be transformed from "merely more data" to "semantically linked data" by addressing the shortcomings identified in the following.

Lack of Conceptual Description of Datasets Usage of LoD Cloud datasets requires a human being to utilize his or her cognizance to identify the domain of the datasets. To exemplify, currently *there is no mechanism to describe that Jamendo*[3] *captures music related information, whereas Geonames captures geographical information*. This is a serious drawback if we envision applications that could seamlessly harness the vast number of facts present in the cloud. Although some efforts have been made to devise a solution to describe the datasets (Quilitz and Leser 2008; Alexander et al. 2009), these approaches focus more on the statistical aspects of the datasets and do not cater to the requirements for capturing conceptual information. We believe the presence of a conceptual description will help in making knowledge discovery easy and systematic.

[1]http://en.wikipedia.org/wiki/Semantic_Web, retrieved October 15, 2009

[2]http://linkeddata.org/

[3]http://dbtune.org/jamendo/

Dataset	Description	Size in triples (approx)	Some datasets linked to
DBpedia	Information from Wikipedia	2.6 million	Geonames, US Census, Freebase
Geonames	Geographic data	8 million	DBpedia, Jamendo, FOAF Profiles
US Census	2000 US Census data	1 billion	GovTrack, DBpedia, Geonames
GovTrack	Information about US Congress	N/A	US Census
FOAFProfiles	Information about people	N/A	SIOC, Flickr Exporter, Geonames

Table 1: Some Datasets Part of LoD Cloud

Absence of Schema Level Links The LoD Cloud datasets lack schema level mappings and do not convey relationships between concepts of different datasets at the schema level. To exemplify, a feature in the Geonames schema can serve as a venue for an event, e.g. the current model identifies *"Atlanta in Georgia was the venue of 1996 Olympics"* at the instance level. This creates significant limitations with respect to the reasoning potential which knowledge on the schema level would provide.

Lack of expressivity The LoD Cloud is of very shallow expressivity as a knowledge base and thus hardly allows to make use of underlying formal semantics through reasoning. The LoD Cloud primarily consists of ground level RDF triples, and hence does not utilize rich expressive features provided by OWL or RDF Schema. To exemplify, *there is inconsistency related to the population of Barcelona between DBpedia and Geonames. This could be detected (and hence fixed) by declaring the properties dbpedia-owl:populationTotal and geonames:population to be functional*. Since instances of Barcelona in geonames and DBpedia are linked to each other using `owl:sameAs`, using an OWL reasoner, an inconsistency could be detected, since an instance cannot have multiple values for a functional property. The lack of such expressive features is a severe drawback as expressivity enhanced LoD Cloud could significantly help in knowledge discovery and thus promote the usage of the LoD Cloud in the scientific community and elsewhere.

The shortcomings identified above severly impact the usage and limit the applications that can be built using the LoD Cloud. To justify our arguments, the following section illustrates the impact of these shortcomings on an important requirement related to knowledge discovery, namely the seamless querying of the LoD Cloud.

Difficulties with respect to querying SPARQL (Seaborne and Prudhommeaux 2008) has emerged as the de-facto query language for the Semantic Web community. It provides a mechanism with which a user can express constraints and facts, and the entities matching those constraints are returned to the user. To ease this process from an infrastructural perspective, data contributors have provided public SPARQL endpoints to query the LoD Cloud datasets. However, the syntax of SPARQL requires users to specify the precise details of the structure of the graph being queried in the triple pattern. To illustrate, in order to formulate a query which spans multiple datasets such as *"Select artists within Jamendo who made at least one album tagged as 'punk' by a Jamendo user, sorted by the number of inhabitants of the places where they are based"*, the user has to be familiar with multiple datasets, and has to express the precise relationships between concepts in the RDF triple pattern, which even in trivial scenarios implies browsing at least two to three datasets. In our previous work (Jain et al. 2009) we made progress towards alleviating this obstacle. But with respect to a systematic querying of the LoD Cloud we believe that the following challenges make the process difficult and will have to be adressed.

- *Schema heterogeneity:* The LoD Cloud datasets cater to different domains, and hence have been modeled differently. To exemplify, *a user interested in music related information has to skim through at least three different datasets such as Jamendo, MusicBrainz, MySpace*. This is perfectly fine from a knowledge engineering perspective, but it makes the querying of the cloud difficult as it requires users to understand the various heterogeneous schemas. This stems from the ***Lack of Conceptual Description of the Datasets*** as pointed out above.
- *Entity disambiguation:* Often LoD datasets have overlapping domains and hence provide information about the same entity. To exemplify, both DBpedia and Geonames have information about the city of Barcelona. Although DBpedia references Geonames using the `owl:sameAs` property, from the perspective of querying this makes it difficult as it might confuse the user as to which is the best source to answer the query. This problem gets even more compounded when contradictory facts are reported for the same entity by different datasets. *For example, DBpedia quotes the population of Barcelona as 1,615,908, whereas according to Geonames it is 1,581,595*. One can argue this might be because of difference in the notion of *the city of Barcelona*. But that leads to another interesting question: *Is the `owl:sameAs` property misused in the LoD Cloud?* This issue is partly related to ***Lack of expressivity*** since there is no mechanism to perform verification of facts. Additionally, the LoD methodology prohibits reification of statements, thus disallowing assignment of context to statements.[4]
- *Ranking of results:* In scenarios where the results of the query can be computed and returned by multiple datasets, the result which should be ranked higher for a specific query becomes an interesting and important question. As presented above, the query related to *population of*

[4]Note that even OWL, in the forthcoming revision OWL 2 (Hitzler et al. 2009), allows for some simple metamodelling.

Barcelona can be answered by multiple datasets such as Geonames and DBpedia, but which one of them is more relevant in a specific scenario is a relevant question. This issue has been addressed from the perspective of popularity of datasets by considering the cardinality and types of the relationships in (Toupikov et al. 2009), but not from the perspective of requirements with regard to a specific query.

How To Get There

Some of the LoD Cloud shortcomings identified above can be resolved by providing a systematic and formal description of the LoD Cloud. There is an apparent lack of an ontology which formalizes and systematically captures the information contained in LoD Cloud datasets. Such an ontology would bring multiple benefits with respect to the use of the LoD Cloud by providing systematic descriptions of the domains captured by the datasets, schema level linking of the datasets, additional schema-level axioms, and hence also better reasoning capabilities. Typically, such an integration would make use of an upper level ontology.

Indeed, in the past the Semantic Web community has relied on upper level ontologies such as Cyc (Reed and Lenat 2002), SUMO (Niles and Pease 2001), or DOLCE (Masolo et al. 2002) to integrate heterogeneous knowledge bases. For applications, these ontologies have been integrated with domain specific ontologies (de Melo, Suchanek, and Pease 2008; Oberle et al. 2007) to provide advantages such as better knowledge discovery, reasoning, or consistency verification.

An upper level ontology typically describes the knowledge base at a very abstract level and thus may or may not convey schema-level knowledge for the grounded knowledge bases which are part of the LoD Cloud. The presence of diverse datasets indeed calls for an ontology which is sufficiently abstract to be able to link to the diverse LoD datasets, but at the same time is grounded enough to provide for easy mapping to LoD datasets. For transforming the LoD Cloud from "merely more data" to "semantically linked data" this integration should provide the following features:

Systematic and Formal Description of LoD Datasets An upper level ontology captures various domains at a fairly abstract level. However the LoD extension of this upper level ontology should create a bridge between the abstraction of the ontology and instantiations available in the LoD Cloud. This will help in providing systematic and formal descriptions of the various ground statements, the classes to which the instances belong, and for identifying schema level relationships. As such, it will go a long way in creating a semantic description of the cloud, and thus help in identifying relationships between datasets at the schema level, and hence facilitate appliations which need to perform reasoning over the cloud. Figure 1 depicts conceptually such an integration of SUMO with the LoD cloud.

Ease of Querying An integrated upper ontology will help for querying since the specific branches of the upper ontology will be linked to the LoD Cloud, hence the user knows which sections of the cloud to look for. It also leaves scope for automated mechanisms for propagating queries over the cloud. To exemplify, if a user specified a SPARQL query in terms of the concepts of the upper level ontology, the mechanism will allow the query to propogate down and query data from actual datasets.

Checking Inconsistencies in the LoD Cloud An upper level ontology with axioms can help in detecting inconsistencies plaguing the linked data cloud. This extension can help in verification of the information captured by the LoD Cloud and thus identify and filter any inconsistent data. Inconsistencies, such as population of London[5] can then be removed using this approach.

Ease of Maintainance and Extensibility Since the LoD Cloud continues to increase in size and will capture more diverse domains in the future, the extension should be easy to maintain to allow modifications, and should support extensibility to provide support for concepts which are not supported natively by the ontology.

We close with a note on scalability issues: While it could be argued, that an attempt to enhance the LoD Cloud with more expressive schema-level knowledge might be doomed from the start due to difficulty of dealing with very large amounts of schema knowledge in ontology reasoners, we believe that this is not necessarily the case. Recent advances, in particular those reported around the Billion Triple Challenges at the International Semantic Web Conferences,[6] show that reasoning over very large knowledge bases is within reach. Importing such reasoning into realistic applications over realistic datasets, as those in the LoD Cloud, however, requires further advances into reasoning with large volumes of noisy data, and indeed research efforts need to be undertaken to realize this. A general discussion of the issues involved in this can be found in (Hitzler 2009).

Conclusion and Future Work

We have outlined shortcomings of the LoD Cloud and have argued for the use of an upper level ontology to alleviate the shortcomings. We believe the road to nirvana for the LoD Cloud is based on the path we have envisoned. We intend to pursue the development of an upper level ontology along the lines we have identified, and a mechanism to query the LoD Cloud seamlessly.

Acknowledgement

This work is funded primarily by NSF Award:IIS-0842129, titled "III-SGER: Spatio-Temporal-Thematic Queries of Semantic Web Data: a Study of Expressivity and Efficiency" and secondarily by NSF ITR Award:071441, "Semantic Discovery: Discovering Complex Relationships in Semantic Web."

[5] http://iandavis.com/blog/2009/08/time-in-rdf-1

[6] http://challenge.semanticweb.org/

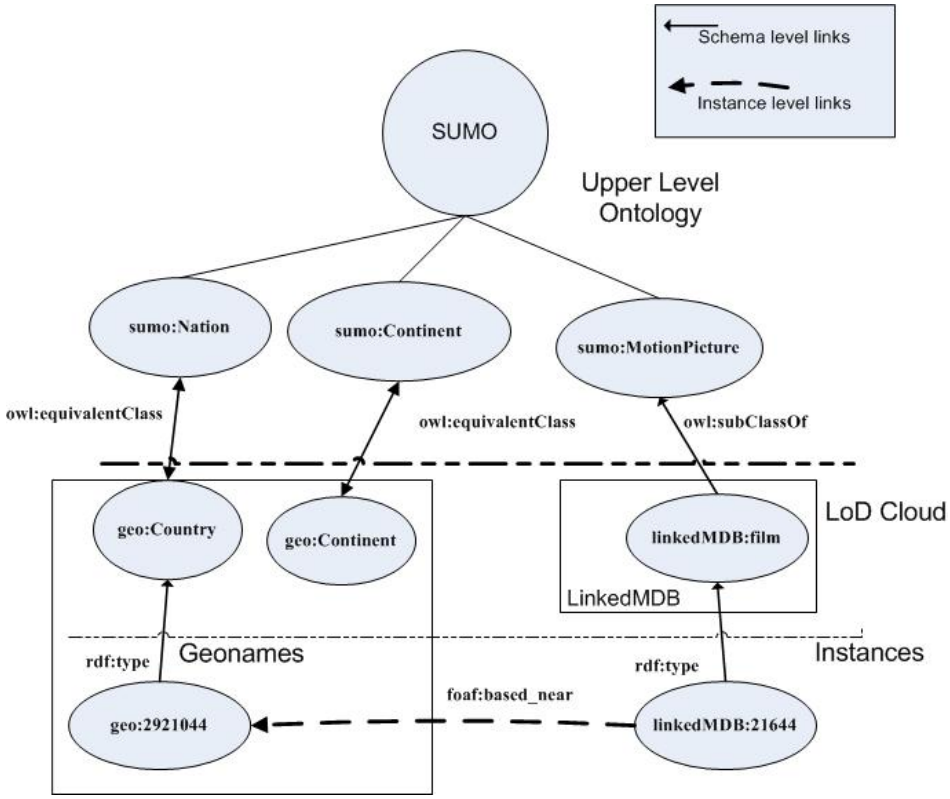

Figure 1: Possible LoD integration with SUMO

References

Alexander, K.; Cyganiak, R.; Hausenblas, M.; and Zhao, J. 2009. Describing Linked Datasets – On the Design and Usage of voiD, the 'Vocabulary of Interlinked Datasets'. In *Proceedings of the WWW2009 Workshop on Linked Data on the Web (LDOW2009)*. Available from http://events.linkeddata.org/ldow2009/.

Berners-Lee, T.; Hendler, J.; Lassila, O.; et al. 2001. The Semantic Web. *Scientific American* 284(5):28–37.

Bizer, C.; Heath, T.; Ayers, D.; and Raimond, Y. 2007. Interlinking Open Data on the Web. In *Demonstrations Track, 4th European Semantic Web Conference, Innsbruck, Austria*. Available from http://www.eswc2007.org/pdf/demo-pdf/LinkingOpenData.pdf.

Bizer, C.; Heath, T.; and Berners-Lee, T. 2009. Linked data – the story so far. *International Journal on Semantic Web and Information Systems* 5(3):1–22. To appear.

de Melo, G.; Suchanek, F.; and Pease, A. 2008. Integrating YAGO into the Suggested Upper Merged Ontology. In *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2008), Dayton, OH, USA, November 2008*, volume 1, 190–193. IEEE Computer Society.

Hitzler, P.; Krötzsch, M.; Parsia, B.; Patel-Schneider, P.; and Rudolph, S., eds. 2009. *OWL 2 Web Ontology Language: Primer*. W3C Proposed Recommendation 22 September 2009. Available from http://www.w3.org/TR/owl2-primer/.

Hitzler, P. 2009. Towards reasoning pragmatics. In Janowicz, K.; Raubal, M.; and Levashkin, S., eds., *Proceedings of the Third International Conference on GeoSpatial Semantics*. Springer. To appear.

Jain, P.; Yeh, P.; Verma, K.; Henson, C.; and Sheth, A. 2009. SPARQL query re-writing for spatial datasets using partonomy based transformation rules. In Janowicz, K.; Raubal, M.; and Levashkin, S., eds., *Proceedings of the Third International Conference on GeoSpatial Semantics*, 140–158. Springer. To appear.

Masolo, C.; Borgo, S.; Gangemi, A.; Guarino, N.; Oltramari, A.; and Schneider, L. 2002. The WonderWeb library of foundational ontologies. *LADSEB-Cnr, Padova, IT, Preliminary Report D* 17. Available from http://wonderweb.semanticweb.org/deliverables/D17.shtml.

Niles, I., and Pease, A. 2001. Towards a Standard Upper Ontology. In Smith, B., and Welty, C., eds., *Proceedings of the International Conference on Formal Ontology in Information Systems – Volume 2001*, 2–9. ACM New York, NY, USA.

Oberle, D.; Ankolekar, A.; Hitzler, P.; Cimiano, P.; Sintek, M.; Kiesel, M.; Mougouie, B.; Vembu, S.; Baumann, S.; Romanelli, M.; Buitelaar, P.; Engel, R.; Sonntag, D.;

Reithinger, N.; Loos, B.; Porzel, R.; Zorn, H.-P.; Micelli, V.; Schmidt, C.; Weiten, M.; Burkhardt, F.; and Zhou, J. 2007. DOLCE ergo SUMO: On Foundational and Domain Models in the SmartWeb Integrated Ontology (SWIntO). *Journal on Web Semantics* 5(3):156–174.

Quilitz, B., and Leser, U. 2008. Querying distributed RDF data sources with SPARQL. In *Proceedings of the 5th European Semantic Web Conference, ESWC 2008, Heraklion, Greece, June 2008*, volume 5021 of *Lecture Notes in Computer Science*, 524–538. Springer.

Reed, S., and Lenat, D. 2002. Mapping ontologies into Cyc. Technical report, Cycorp, Inc,. Available from http://www.cyc.com/doc/white_papers/.

Seaborne, A., and Prudhommeaux, E. 2008. SPARQL query language for RDF. *W3C Recommendation 15 January 2008*. Available from http://www.w3.org/TR/rdf-sparql-query/.

Toupikov, N.; Umbrich, J.; Delbru, R.; Hausenblas, M.; and Tummarello, G. 2009. DING! Dataset ranking using formal descriptions. In *Proceedings of the WWW2009 Workshop on Linked Data on the Web (LDOW2009)*. Available from http://events.linkeddata.org/ldow2009/.