Thoughts on the Complex Relation Between Linked Data, Semantic Annotations, and Ontologies

Krzysztof Janowicz
University of California Santa Barbara, USA
jano@geog.ucsb.edu

Pascal Hitzler
Wright State University, USA
pascal.hitzler@wright.edu

ABSTRACT

The relation between data, annotations, and schemata seems straightforward at first: Data are annotated with additional meta information according to some schemata in order to expose additional non-intrinsic characteristics relevant to the meaningful interpretation of said data. However, on closer examination, things are not as simple. Focusing on geo-information retrieval, we will try to disentangle the aforementioned relations. We will report from our own experience and from observations gathered by editing papers about ontologies and Linked Data for the Semantic Web journal.

Keywords

Semantic Annotation, Linked Data, Ontologies

Introduction

After peeling of the jargon, Linked Data is less about specific technologies but about a paradigm shift. Today's Web is about documents and simple links between them. These documents providing the structure and context for the inherent data and, thus, support their interpretation. In contrast, Linked Data are not bound to a specific document but can be freely combined outside of their original creation context. In theory, one can use Linked Data to answer complex queries that span multiple repositories and establish new links between data. Unfortunately, receiving meaningful results is more difficult than one may expect. While uncoupling data from documents eases accessibility it puts the burden on their interpretation. In theory, ontologies are supposed to provide reference frames to overcome these difficulties.

In a nutshell, our argumentation is as follows. **We believe that the sweet spot for ontologies is still unclear.** Are ontologies an additional layer on top of data models; are they data models themselves; are Linked Data entities instances of ontological classes or just annotated using ontologies which exist in their own realm; what difference does this make; what is the role of semantics & reasoning for querying and information retrieval; are labels all we need? While we cannot answers all these questions, we will illustrate the relation to data models and the role of semantics.

Ontologies and Data Models

While the term data model/schema is overloaded, we will follow the established tradition [9] to differentiate data models into conceptual models concerned with types of entities and their relationships on an abstract level, the more concrete logical models that introduce attributes, cardinalities, and so forth, without enforcing a particular implementation, and physical models which are implementation specific, e.g., by assigning primitive data types to attributes. Intuitively, ontologies are an additional layer on top of conceptual models. While conceptual models are not implementation specific, they are purpose-driven nonetheless. Thus, it has been argued that an additional layer of abstraction is required which is independent of specific applications, tasks, or viewpoints. It is often said that ontologies model the world, i.e., what exists. Similarly as logical models rest upon conceptual models, the latter should build upon ontologies. This can either be realized by extending & instantiating ontologies, or by relating to them, i.e., by semantic annotations.

(I) In the first case, conceptual models use classes and relations from ontologies and extend them to introduce purpose-specific aspects. For example, an ontology may define a *Place Of Interest* (POI) class and conceptual models may subclass it in different ways depending on whether they are used for a navigation system or a historical gazetteer. Entities such as the French Press Cafe in Santa Barbara, are instances of these classes. Intuitively, for this to function, such ontologies must be universal enough to act as a common foundation for different, purpose-driven conceptual models. At the same time they should not be overly generic to a degree where no interesting statements about the defined classes and relations are made. In the literature [4], these ontologies are known as top-level or foundational ontologies. In practice, however, it turns out that their level of abstraction is not suitable anymore for many use cases and certainly not for information retrieval. These ontologies introduce classes such as *Endurant* and *Perdurant* together with numerous complex ontological commitments. Consequently, additional types of ontologies have been proposed that rest upon the foundational level but are more tangible, namely domain, task, and application ontologies. Unfortunately, this does not address the real problems. First, how would such ontologies differ from conceptual models? Secondly, the emerging, constructed, cultural, and highly

contextual nature of *meaning*, does not harmonize well with static, context-free, and highly abstract foundational ontologies. Thus, based on our observations, it is not surprising that these ontologies only play a marginal role for Linked Data.

Alternatively one can argue that ontologies are conceptual models. In this case, one would expect that they remain on an abstract level and do not introduce attributes, cardinality constraints, and so forth. However, today's Semantic Web knowledge representation languages, such as the Web Ontology Language (OWL) or the Resource Description Framework Schema (RDFS), go even further by include XML Schema data types. These languages are clearly suitable for physical models and many ontologies make heavy use of data type properties. In fact, the distinction between different data model layers got lost in the Semantic Web community over the years; cf. [7]. This does not necessarily mean that one would need yet another modeling language; these aspects can be approached from a methodological perspective, e.g., by modeling patterns [2].

The real issue are not representation languages but that **models of the physical world do not necessarily make for good data models**; cf. [1]. For example, comparing the definition of *Person* in the Dolce Ultra Light (DUL) ontology[1] to schema.org one will notice that the latter lists *givenName*, *telephone* number, and *spouse* among other properties while DUL does not. The difference between modeling the physical world and data models becomes even more clear from an information retrieval perspective. An ontology engineer would argue that telephone numbers are not properties of persons but of devices in possession by these persons and model this accordingly. While this is certainly true, querying Linked Data based on such an ontology using the SPARQL Protocol and RDF Query Language would be cumbersome. Instead of matching against a single triple, one would have to retrieve the devices first and then query for their numbers. In almost all cases these devices are of no interest but would have to be introduced nonetheless – just imagine a phone book would have to list such devices.

(II) In the second case, data do not directly instantiate classes but are annotated using classes from an ontology. These data may be text from a Web page or encoded using some language, e.g, the Geography Markup Language. GML comes with its own specifications that clearly distinguish between data model layers. In case of the Open Geospatial Consortium this is exactly the difference between abstract specifications and implementation specifications. In these cases, the ontologies used for annotation form an entirely separate layer and the annotation relation has no formal semantics. In case of Linked Data this is not the intended approach. Entities such as Berlin are instances of a certain type, e.g., *City*[2] and link to other resources, e.g., using OWL language constructs such as *owl:SameAs*; cf. [6].

Summing up, we argue that **ontologies suitable for Linked Data querying necessarily lead to physical data models.** Unsurprisingly, based on our observations, this leads to a gap between knowledge engineers and their ontologies and Linked Data enthusiasts and their vocabularies. Expressive ontologies that support interesting reasoning are rarely used for massive datasets, while vocabularies widely used on the Web of Linked Data are mere taxonomies.

Semantics and Reasoning

Is there more to the above discussion than just an academic five-finger exercise? We believe that semantics beyond simple labels does matter. By reviewing existing Linked Data it becomes clear that most of them use ontologies/vocabularies that do not go beyond *surface semantics*, i.e., the ontologies merely consist of explicit subsumption relations together with some other relations without providing a detailed axiomatization.[3] For instance, such an ontology would merely state that *Restaurant* is a subclass of *POI* and has a *Name*. The lack of a deeper axiomatization prevents any interesting reasoning, e.g., that restaurants cannot have visitors outside of their opening hours. Such surface ontologies make the meaningful interpretation and querying of Linked Data a difficult and manually intensive task and reduce ontology alignment techniques to educated guessing.

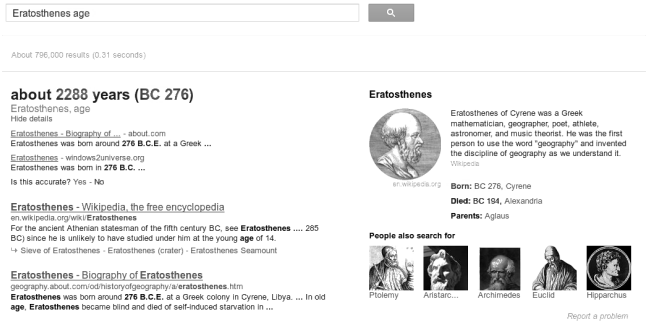

Figure 1: Querying for Eratosthenes' age using Google's Knowledge Graph in 2012.

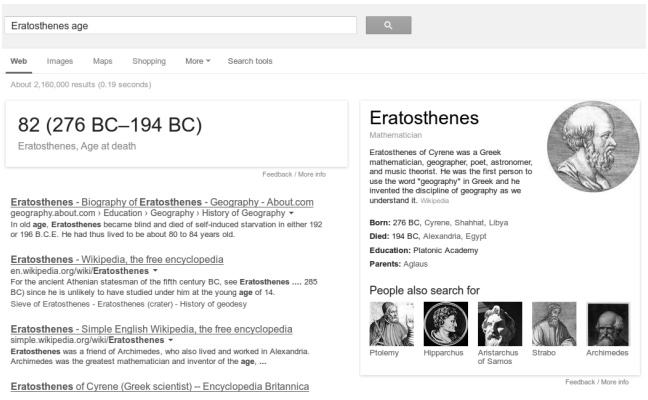

Figure 2: Querying for Eratosthenes' age using Google's Knowledge Graph in 2013.

To give an example for the power of semantics in query answering, compare the results from Google's Knowledge Graph in 2012 with the same query in 2013. Fig. 1 depicts a query for the age of Eratosthenes and its result – 2288

[1] See www.ontologydesignpatterns.org/ont/dul/DUL.owl
[2] See http://dbpedia.org/page/Berlin
[3] In the literature the terms lightweight & heavyweight ontology [3] are frequently used. We find them rather misleading as even minimal ontologies can convey detailed semantics.

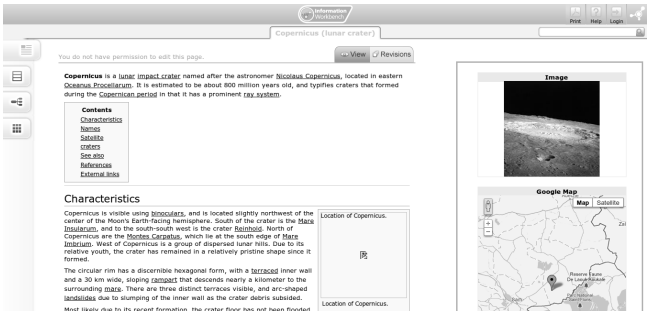

Figure 3: Fluidops maps the Copernicus crater to the Chad.

years. Note that the *born* and *died* information is available on the right. One year later, the query result (82) is displayed in Fig. 2. While the first result is not wrong, the so-called ontological commitments have clearly changed and the Google engineers decided to stop increasing the age of a human post mortem. It is worth noting that the query *Eiffel Tower age* will not work despite the opening date (*opened*) being available. Summing up, the Knowledge Graph does not only contain class and property labels but also rules that exploit the pair *born* and *died* to compute the age of humans.

Now let us consider an example where the semantics is not sufficiently captured and meaning is largely conveyed via human readable labels. Fig. 3 shows the Fluidops interface rendering Linked Data about the Copernicus crater from DBpedia. The crater is located on the Moon's surface but *geo:lat* and *geo:long* have been used to represent its centroid. While they are reserved for WGS84, this is not enforced by the W3C Basic Geo specs. While DBpedia ignores this fact, Fluidops rightfully renders whatever *geo:lat* and *geo:long* pair it finds using Google Maps. Consequently, the crater is placed near the city of Sarh in Chad.[4]

Consequences

So where is the sweet spot for ontologies that go beyond surface semantics? We have no simple answer but would argue that standardizing meaning is a misconception. Maybe, at least for Linked Data, it is time to give up on the idea of context-free ontologies as models of the physical world and instead define a multitude of purpose and data-driven micro-ontologies; cf. [5]. Research should focus on aligning and translating different perspectives expressed by these micro-ontologies and *ground* them [8] to foster interoperability. **Ontologies should be engineered based on the real data they are supposed to reflect and their axiomatization should be driven by the inference needs of typical queries**; see the age example. In contrast to the current state of the Knowledge Graph, these axioms (query answering rules, if you like) should be shared together with the data. Finally, we share Google's recent argument that maps should be personalized; the same is true for ontologies.

Acknowledgement. The first author acknowledges support from the Hellman Family Faculty Fellowship 2013. The second author acknowledges support by the National Science Foundation under award 1354778 *EAGER: Collaborative Research: EarthCube Building Blocks, Leveraging Semantics and Linked Data for Geoscience Data Sharing and Discovery.*

[4]See http://stko.geog.ucsb.edu/location_linked_data for more details, examples, and unintended consequences.

1. REFERENCES

[1] H. Couclelis. Ontologies of geographic information. *International Journal of Geographical Information Science*, 24:1785–1809, 2010.

[2] A. Gangemi and V. Presutti. Towards a pattern science for the semantic web. *Semantic Web*, 1(1):61–68, 2010.

[3] A. Gómez-Pérez and O. Corcho. Ontology languages for the semantic web. *Intelligent Systems, IEEE*, 17(1):54–60, 2002.

[4] N. Guarino. *Formal Ontology in Information Systems: Proceedings of the First International Conference June 6-8, 1998, Trento, Italy*. IOS Press, 1998.

[5] R. Guha. Micro-theories and contexts in CYC. Part 1: basic issues. MCC technical report ACTR-CYC-129-90, MCC Corp., 1990.

[6] H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, and H. S. Thompson. When OWL:sameAs isn't the same: An analysis of identity in linked data. In *The Semantic Web–ISWC 2010*, pages 305–320. Springer, 2010.

[7] W. Kuhn. Modeling vs encoding for the semantic web. *Semantic Web*, 1(1):11–15, 2010.

[8] S. Scheider. *Grounding geographic information in perceptual operations*, volume 244 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2012.

[9] D. Tsichritzis and A. Klug. The ANSI/X3/SPARC DBMS framework report of the study group on database management systems. *Information systems*, 3(3):173–191, 1978.