Position paper: Uncertainty reasoning for linked data

Dave Reynolds[1]

[1] Hewlett Packard Laboratories, Bristol
Dave.e.Reynolds@gmail.com

Abstract. Linked open data offers a set of design patterns and conventions for sharing data across the semantic web. In this position paper we enumerate some key uncertainty representation issues which apply to linked data and suggest approaches to tackling them. We suggest the need for vocabularies to enable representation of link certainty, to handle ambiguous or imprecise values and to express sets of assumptions based on named graph combinators.

Keywords: Uncertainty reasoning; linked open data; semantic web

1 Introduction

The need for reasoning over uncertain information within the semantic web occurs in many different situations. It can arise from intrinsic uncertainty in the world being modeled or from limitations of the sensing or reasoning agent itself (epistemic). The term *uncertainty* is often used to refer to many different notions including ambiguity, randomness, vagueness, inconsistency, incompleteness [1][9].

In recent years an approach to the semantic web, called *linked data*, has been developed and offers a promising route to practical and widespread semantic web uptake. It provides a set of design guidelines or patterns for how the semantic web technologies, and broader web architecture, can be used for sharing information. The existing guidelines and practices have no provision for representation of uncertainty; yet linked data is indeed fraught with many of these different types of uncertainty.

In this brief position paper we examine the ways in which uncertainty can occur in a linked data setting and sketch possible approaches to addressing the issues raised.

2 Linked data

Linked Data is a set of conventions for publishing data on the semantic web. It is based on principles outlined by Tim Berners-Lee [2]. These principle advocate the use of http URIs for naming entities, the publication of data about these URIs using the standards (RDF, SPARQL) and inclusion of links to other URIs so that agents can discover more information. While quite simple these guidelines, along with a growing body of practical advice [3], have led to publication and linking of many datasets in this form [4]. This has resulted in high profile commercial applications such as [5].

While not explicitly stated, the style of linked data places an emphasis on data sharing and simplicity, with corresponding less emphasis on depth of modeling and reasoning. Yet the intrinsic nature of the linked data approach leads to issues of uncertainty representation and reasoning. This is due to the emphasis on cross-linking

multiple data sources that have been independently developed and modeled. Uncertainty can arise from the instance linking process, from the mapping between different sources models and due to differing hidden assumptions in the underlying datasets. Yet the essence of linked data, and a large part of the reason for its uptake, is simplicity. The data is intended to be self-descriptive and accessible through simple link following and graph union or through SPARQL endpoints. Our challenge is to develop a common, easy to deploy, approach to uncertainty representation which can be applied to linked data sets without losing this simplicity.

3 Some sources of uncertainty in linked data applications

In this section we enumerate some key sources of uncertainty for linked data. We focus on the sources which directly result from the intrinsic nature of linked data – the cross-linking of independently developed RDF datasets.

3.1 Ambiguity resulting from data merging

In linked data, entities (Individuals) which co-occur with different URIs in different datasets are unified. This is achieved by publishing `owl:sameAs` relations between identified entities, either within the dataset or as a separate link set. The process of identifying such co-references is imperfect. Firstly, the co-references are typically found by a mixture of string matching, attribute matching, and type constraints, generally based on a statistical or machine learning algorithm [6]. Thus co-references are only identified with some probability (or less formal heuristic weighting). Yet the asserted links are binary and the strength of association is lost. Secondly, the nature of the entities is ambiguous in some datasets. For example, Wikipedia and thus DBPedia conflate the concepts of the City *Bristol* in the UK and the associated Unitary Authority. A co-reference link that identified the ambiguous DBPedia concept with one that specifically denotes the Unitary Authority would be an error in general, even though it may be an acceptable approximation in some situations.

3.2 Misalignment of precision and assumptions between merged sources

Many datasets in the linked data web publish property values for the entities they describe; for example, the *population* of the *City of London*. Yet those values are sometimes imprecise or dependent upon measurement assumptions that are not made explicit. For example, the *population* of a city depends on the time of the measurement, the measurement methodology and the precise definition of the boundary of the city; it is also subject to statistical uncertainty. As a result, at the time of writing, a linked data query on London returns a graph with four assertions on its population ranging from 7,700,000 to 8,500,000. One of these sources of variation, the time of measurement, is sometimes made explicit in data and indeed one of the four assertions is (indirectly) time qualified. However, such contextual qualification is not consistently available and, in any case, only accounts for one source of variation. Thus when datasets are linked the resulting union will often have multiple conflicting values for supposedly functional properties.

3.3 Misalignment of models

When linking datasets we also want to map the associated ontologies. This process is just as error prone as entity co-reference since the axiomatization of concepts in the ontologies is rarely so complete as to allow a unambiguous mapping. Errors in the ontology mapping can lead to global effects such as unexpected identification of related concepts. Determining and publishing such alignment errors is the subject of considerable research and is outside the scope of this paper.

3.4 Absence of source reliability information

Separate from the uncertainty arising from combination and linking of datasets then the datasets themselves can be uncertain or contain errors (either accidental or malicious). While this is true in general in the semantic web, the linked data approach implies broad cross linking with no provision for narrow scoping of link references. This exacerbates the problems of the veracity or trustworthiness of included datasets.

4 Mitigation approaches

We now discuss approaches to mitigate the effects of these uncertainty sources on the consumers of linked data. In keeping with linked data methodology we seek simple, broadly applicable, design patterns. In particular, we suggest the need for design patterns for making the uncertainty inherent in the linked datasets more explicit, and mechanisms to enable selective combination of datasets (so that problematic values or links can be omitted). In this a short position paper we only sketch the suggested approaches as a basis for discussion in the workshop.

4.1 Link vocabulary

The *link vocabulary* would provide a common representation for co-reference links, enabling publication of the link certainty information on which per-link inclusion decisions can be made. This can be achieved by extending the voiD ontology [8] with a concept *UncertainLinkSet* (as a subclass of `void:LinkSet`), and associated properties for describing the method used for deriving the link set. The *UncertainLinkSet* itself would contain n-ary relations (*WeightedLink*) comprising the link and associated link weight. Different subclasses of *WeightedLink* indicate different interpretations of the link weight (such as probabilistic or ad hoc).

4.2 Imprecise value vocabulary

The *imprecise value vocabulary* would provide a common representation for imprecise values that arise from data set merger, as discussed in 3.2. This would allow republication of merged datasets which explicitly show the variation in source data values. Returning to our example of the population of London the merged set might look like:

```
:London :population [a :ImpreciseValue;
   :samplevalue [:value 7700000; :source :s1; :context :y2009]
   :samplevalue [:value 7900000; :source :s2; :context :y2008]
   :estimatedValue 785123]
```

4.3 Override graphs

Finally we suggest the need for override graphs so that one agent can publish retractions and overrides to the link assertions or data assertions made by another.

The current approach to this, in linked data applications, is to partition data and link sets into named graphs [7]. For example, rather than include all the co-reference links directly in the same graph as the entity descriptions, we partition them into a separate named graph. In this way a RESTful access can see the union of the relevant graphs but a SPARQL endpoint can support selection of which graphs to include. This allows agents to avoid selected link sets or sub-sources but only at the grain size of the entire graph. To overcome this limitation we suggest extending the VoiD vocabulary to include graph combinators *difference*, *union* and *replace*. So one source can decide which subsets of the data and links to trust, and can then publish the assumptions it is making as a set of deltas over the source graphs. The difference graphs enable per-link and per-assertion changes to be expressed even if the underlying source only publishes the link set or data assertions as monolithic graphs.

5 Discussion

Of the issues in section 3 we have suggested an agenda for how to address some of them. The *link* and *imprecise value* vocabularies enable publication of link uncertainty (3.1) and value ambiguity (3.2) information in linked data sets. The vocabularies themselves would not remove the uncertainties, nor the problems of estimating them. However, simply having a means to publish this information is already a step forward. The suggested *graph combinators* would enable an agent to make and publish more selective data combinations, based on its interpretation of link strengths and data values. This does not solve the problems of deciding which parts of which sources to trust, but it does enable more effective sharing of such decisions.

References

1. Laskey, K.J., et al.: Uncertainty Reasoning for the World Wide Web, W3C Incubator Group Report, 31 March 2008. http://www.w3.org/2005/Incubator/urw3/XGR-urw3/
2. Berners-Lee, T.: Linked Data. http://www.w3.org/DesignIssues/LinkedData.html (2006)
3. http://linkeddata.org/
4. http://linkeddata.org/data-sets
5. Kobilarov, G., et al.: Media Meets Semantic Web --- How the BBC Uses DBpedia and Linked Data to Make Connections. Proc. of the 6th European Semantic Web Conference on the Semantic Web: Research and Applications (Heraklion, Crete, Greece). Lecture Notes In Computer Science, vol. 5554. Springer-Verlag, Berlin, Heidelberg, 723-737. (2009)
6. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering 19 (2007)
7. Carroll, J. J., Bizer, C., Hayes, P., and Stickler, P.: Named graphs, provenance and trust. In Proceedings of the 14th international Conference on World Wide Web (Chiba, Japan, May 10 - 14, 2005). WWW '05. ACM, New York, NY, 613-622. (2005)
8. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: voidD Guide : Using the vocabulary of Interlinked Datasets. http://rdfs.org/ns/void-guide (2009)
9. Kruse, R., Schwecke, E., and Heinsohn, J. 1991 *Uncertainty and Vagueness in Knowledge Based Systems*. Springer-Verlag New York, Inc.