Why the Semantic Web Should Become More Imprecise

Heiko Paulheim
Technische Universität Darmstadt
Hochschulstraße 10
64283 Darmstadt

Jeff Z. Pan
University of Aberdeen
King's College
Aberdeen AB24 3UE, UK

ABSTRACT

Recent Semantic Web research has been largely focused on precise ontologies and knowledge representation, leaving only little space for imprecise knowledge such as rules of thumb. The result is a Semantic Web which can answer requests almost perfectly with respect to precision, but provides only a low recall. In this position paper, we envision to address this issue with a stack of Semantic Web technologies that allow imprecise knowledge as an essential ingredient.

Categories and Subject Descriptors

I.2.3 [**Artificial Intelligence**]: Deduction and Theorem Proving—*Uncertainty, "fuzzy", and probabilistic reasoning*; I.2.6 [**Artificial Intelligence**]: Learning—*Induction*

General Terms

Theory

Keywords

Semantic Web, Linked Open Data, Uncertainty Reasoning, Machine Learning

1. INTRODUCTION

During the evolution of mankind, imprecise knowledge has played a vivid role in the survival of the human species. Every day, we make quick decisions based on imprecise knowledge. For example, the rule *lions are dangerous* has proven to be particularly useful for the survival of our kind, and thus has been selected as a useful piece of knowledge by the process of evolution. However, despite its use, the rule *lions are dangerous* is not entirely precise, since there may be single lions that are not dangerous, e.g., new born lion babies.

Research on the semantic web so far has been largely focused towards building solutions that favor highly precise knowledge. Reasoners are built to be sound and complete. Ontology engineering methodologies and design patterns are explored for making ontologies as precise as possible. The above example, formalized as the subsumption axiom $Lion \sqsubseteq DangerousAnimal$, would not be allowed when using the *OntoClean* methodology [5] – with $Lion$ being a rigid class, but $DangerousAnimal$ being non-rigid, one of the central design principles would have been violated. A rigid ontology engineering approach would state that the axiom *lions are dangerous* is not always correct, and thus, it should not be part of a precise ontology in order to avoid false reasoning results.

In other words: if our brains were ontology reasoners processing completely precise ontologies, our species would have been long gone from the surface of this planet.

In this paper, we argue that imprecise knowledge can add value to the Semantic Web as we know it, and that it could be an essential ingredient to building future semantic web applications. The rest of this paper is structured as follows: section 2 introduces a number of use cases of the Semantic Web which could benefit from imprecise knowledge. Section 3 discusses what a future Wemantic Web incorporating imprecise knowledge could look like. In sections 4 and 6, we show some ongoing research on automatically discovering and processing imprecise knowledge on the semantic web. We conclude with a summary and an outlook on future research.

2. USE CASES FOR IMPRECISE KNOWLEDGE

In classical information retrieval, there is the well-known trade-off of precision and recall: it is fairly simple to construct systems that perform well on either of those metrics, but optimizing both at the same time is a difficult task. As discussed above, semantic web research has been largely focused on representing and processing knowledge in a precise manner. Consequently, semantic web data often suffers from a recall problem.

In their pre-studies to building the prestigious *Watson* project, researchers at IBM also examined how approaches trying to look up information on structured knowledge sources would perform as a baseline. They reported that those approaches achieve a fairly high precision, but the recall would have been too low to get to the impressive results achieved by Watson – in fact, by only querying sources like DBpedia, Watson would only have been able to answer around 2% of the queries correctly [4, 7]. This shows that a good query answering system should be able to allow for a reasonable trade-off between precision and recall.

The same can be observed when experimenting with engines such as Wolfram Alpha[1], which can give quite precise answers: when asking, e.g., for the distance from Darmstadt to Aberdeen, the user

[1] http://www.wolframalpha.com/

is presented the exact distance in different units of measurement, plus additional information such as how long an aircraft would take to travel the distance. On the other hand, Wolfram Alpha often fails on comparatively simple questions, such as asking for the most famous work by William Shakespeare.

Today, looking up data on the semantic web will lead to a set of more or less correct, but probably not complete results. For example, querying DBpedia for instances of the YAGO category *UniversityTownsInGermany* yields seven results[2]. However, there are more than 40 of such towns in Germany. Since Linked Open Data follows the open world assumption, the answer retrieved is not incorrect. However, it is also only of limited use in query answering systems.

There are other datasets in Linked Open Data that motivate certain applications. For example, the *DrugBank* dataset[3] contains information about different medical drugs and their interactions. One could envision an application where a user can enter the drugs he or she is taking and ask for potentially harmful interactions. Again, in that case, recall is at least equally essential than precision: not reporting a potential interaction might lead to wrong self-medication, while reporting faulty interactions would, in the best case, just make the user check back with his or her doctor.

A scenario which demonstrates the trade-off of precision and recall even more drastically is the domain of emergency management. Recently, prototypes have been discussed that leverage Linked Open Data for supporting emergency response staff. One example is *MICI*, which uses information from Linked Geo Data to inform decision makers about potentially affected infrastructure. For example, if a fire is reported, schools and kindergartens nearby can be identified in order to prepare evacuation activities [16].

In that scenario, a high recall is even more important than a high precision. Missing to evacuate one school has more drastic effects than discarding one result that is not a school. Given that the type information flagging an object as a school is not complete, imprecise axioms such as $\exists placeOf.Course \sqsubseteq School$ may help increasing the recall, even though they may slightly reduce precision (e.g., private teachers giving courses at home may also be covered). This shows that even in unexpected scenarios (intuitively, emergency management requires precise knowledge), imprecise knowledge can be an asset.

These scenarios show that some applications built on Linked Open Data require a good recall, while a precision of 100% is often not that essential. Thus, allowing for imprecise information which increases recall, for the price of sacrificing a few percent of precision, seems to be desirable.

3. TOWARDS AN IMPRECISE SEMANTIC WEB

Leveraging imprecise knowledge on the Semantic Web requires efforts in different areas. First, imprecise knowledge has to be created. This may be done by humans coding some of their rules of thumb and implicit knowledge, but in order to foster a quick adoption, automatic approaches are highly favorable. Such approaches could, e.g., use frequent pattern mining on existing data (cf. section 4), or employ knowledge extraction from external sources, such as texts on the web.

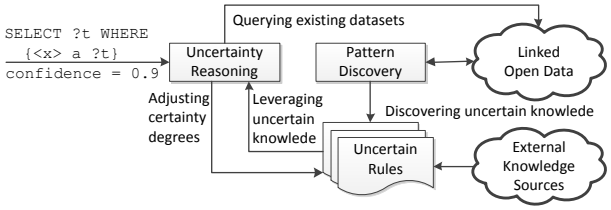

Figure 1: Enabling a SPARQL endpoint with different precision levels

Second, there need to be means to represent uncertain knowledge and uncertainty degrees on the level of individual axioms, both on the T-box and the A-box level. A possible option would be the adoption and extension of the Semantic Web Provenance vocabulary[4].

Third, imprecise knowledge needs to be processed in a useful manner. It is clearly not enough to present the user a set of possible statements with their degrees of precision. A possible implementation could provide SPARQL endpoints which allow for specifying a desired degree of precision, as shown in figure 1. Such an endpoint could be used to build applications satisfying different information needs, characterized by different trade-offs of precision and recall. A special reasoner capable of dealing with uncertain knowledge can both deliver the queried information at the demanded level of precision, as well as adjust the certainty degrees of the uncertain knowledge axioms.

Furthermore, using uncertain knowledge to derive new facts will be essential. Just as reasoning was proclaimed to be one of the key capabilities of the Semantic Web, new breeds of reasoners have to be able to cope with imprecise information (cf. section 6). This is particularly useful when combining different uncertain statements. Using our introductory example, the imprecise implications $Lion \sqsubseteq DangerousAnimal$ and $BigAnimal \sqsubseteq DangerousAnimal$ could be used to decide whether an actual instance of a lion is a dangerous animal.

A crucial point in creating and processing imprecise knowledge is the assignment of useful degrees of uncertainty. If uncertainty degrees are chosen badly, the entire approach can be flawed. One counter example are the degrees of similarity used in ontology matching [3], in particular those derived by syntactic matching algorithms. For example, the words *House* and *Mouse* may be similar, but stating $House \equiv Mouse$ with a precision of 0.8 would not be a useful piece of imprecise information, since most of the conclusions drawn from that axiom would be useless. Thus, meaningful numerical truth values for imprecise axioms are required.

4. CREATING IMPRECISE KNOWLEDGE

To minimize the ramp-up efforts, discovering good uncertain axioms is a desirable goal for the future uncertain Semantic Web. While various approaches for the automatic acquisition of uncertain knowledge are possible, a natural way is to use inductive learning, as it close to the way humans acquire their (imprecise) knowledge

[2]As of July 17th, 2012

[3]`http://www4.wiwiss.fu-berlin.de/drugbank/`

[4]Cf. `http://www.w3.org/2011/prov/wiki/ProvenanceRDFNamedGraph`

of the world: we observe the world and derive generalized rules from those observations. To pursue this approach, we have conducted a first experiment with mining uncertain rules from DBpedia.

The starting point of our considerations was a method for ontology learning on Linked Open Data. The underlying rationale was that linked open data has a lot of instance data, but only little schema information that can be exploited by reasoning in intelligent applications. Furthermore, the instance data is largely incomplete. One example is the mapping of DBpedia to YAGO. As shown in the above example about German university towns, this mapping is largely incomplete. Thus, it would be desirable to have rules which allow for the deduction of those missing mappings. For example, the axiom $GermanCity \sqcap \exists locationOf.University \sqsubseteq GermanUniversityTown$ could help completing this particular type of mapping.

As discussed in [20], a possible way of learning such rules on Linked Open Data is the use of association rule mining. Since association rule mining is a heuristic approach, it produces heuristic, yet inherently imprecise rules. Thus, we have pursued that approach to discover imprecise knowledge, using our framework *FeGeLOD*, which makes Linked Open Data accessible to machine learning tools [14].

For a preliminary study, we have used a dataset consisting of eight subsets of DBpedia, covering different domains and containing 1,000 randomly sampled instances each. The goal was to infer additional rules that help establishing additional links to the YAGO classification.

We have constructed three types of features on the dataset: direct type features (e.g., $Novel$) , unqualified relations (e.g., $\exists author.\top$), and qualified relations (e.g., $\exists author.AmericanWriter$). Using those features, we have mined association rules using four different levels of minimum confidence (1.0, 0.75, 0.5, and 0.25). As the number of mined rules is potentially large, we sampled a random subset of 250 rules from each dataset and level, and rated those rules manually, using three different rating values: correct, imprecise, and wrong.

Figure 2 depicts the evaluation of the mined rules. It can be observed that on some datasets (animals, books, movies), a considerable amount of useful, imprecise axioms can be learned. Examples include:

$$BirdsOfSuriname \sqsubseteq BirdsOfVenezuela, \quad (1)$$

$$BritishNovels \sqsubseteq \exists author.EnglishNovelists \quad (2)$$

$$TamilLanguageFilms \sqsubseteq \exists starring.TamilActors \quad (3)$$

Each of those axioms is correct to a certain extent, yet not 100% precise. Many birds living in Venezuela will also be observable in the neighboring Suriname, books written by English novelists are likely to be British novels (although they may occasionally be short story collections), and films starring Tamil actors are likely to be films in Tamil language, despite some Hollywood productions starring Tamil actors.

To further analyze the use of those imprecise rules, we have calculated the *productivity* of those axioms, defined as the number of additional axioms that may be derived by the rule. We define the productivity of an axiom $A \sqsubseteq B$ as the number of statements that fulfil the body, but not the head:

$$prod_{A \sqsubseteq B} := \#(B \sqcap \neg A) \quad (4)$$

For a better comparison, we define the *relative productivity* as

$$rprod_{A \sqsubseteq B} := 1 + \frac{\#(B \sqcap \neg A)}{\#A} \quad (5)$$

The relative productivity of an axiom describes the factor by which the recall of a query for instances of A could be increased when using all consequences induced by that axiom. For the above examples 1–3, the relative productivity is 1.91, 2.17, and 2.25. Thus, the recall about doubles in all three cases. This shows that the approach is feasible to produce rules of useful, imprecise knowledge.

5. REPRESENTING IMPRECISE KNOWLEDGE

Once imprecise information is discovered and created with appropriate measures of imprecision, it needs to be stored in order to allow for further processing. So far, there has not been a standard for representing uncertainty information in RDF and OWL. W3C set up an incubator group URW3-XG on uncertainty reasoning for the World Wide Web[5]; however, according to its final report[6], there seems to be some debate about what should be standardised by W3C. No working group, accordingly, was set up based on the work in the incubator group.

As imprecision can occur both on the instance and the schema level, mechanisms of an imprecise Semantic Web should be able to cope with both, ideally in a uniform way. Technically, to represent uncertainty, one could either aim at using existing language features in the Semantic Web, or to create new language extensions [2]. While numerous new extensions have been proposed[8], we believe that only downward compatible approaches using existing language features, at least to a certain extent, have the potential to become a building block of an imprecise semantic web.

Since RDF allows reification, i.e., statements about statements, a possible way would be to add reified statements representing uncertainty. Although reification is not well supported by many tools, and the use of reification for Linked Open Data is often explicitly discouraged[9], the advantage of this approach is that no new standards are required, except for a small vocabulary for representing the statements about imprecision, and that it is fully downward compatible with existing RDF and Linked Open Data.

As discussed above, there have been some efforts in the Provenance community to establish provenance information for named graphs. This would allow for flexible mechanisms for assigning imprecision measures[10]. An RDF graph would have to be broken down into several subgraphs, which could then be flagged with the appropriate measures. The same could be done for schemas. While this could lead to a larger number of named graphs and subgraphs, it would naturally fit with today's standardization efforts of the W3C.

[5] http://www.w3.org/2005/Incubator/urw3/

[6] [7]

[8] See http://www.w3.org/2005/Incubator/urw3/XGR-urw3-20080331/#appendixC for a comprehensive list

[9] cf. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

[10] Cf. http://www.w3.org/2011/prov/wiki/ProvenanceRDFNamedGraph

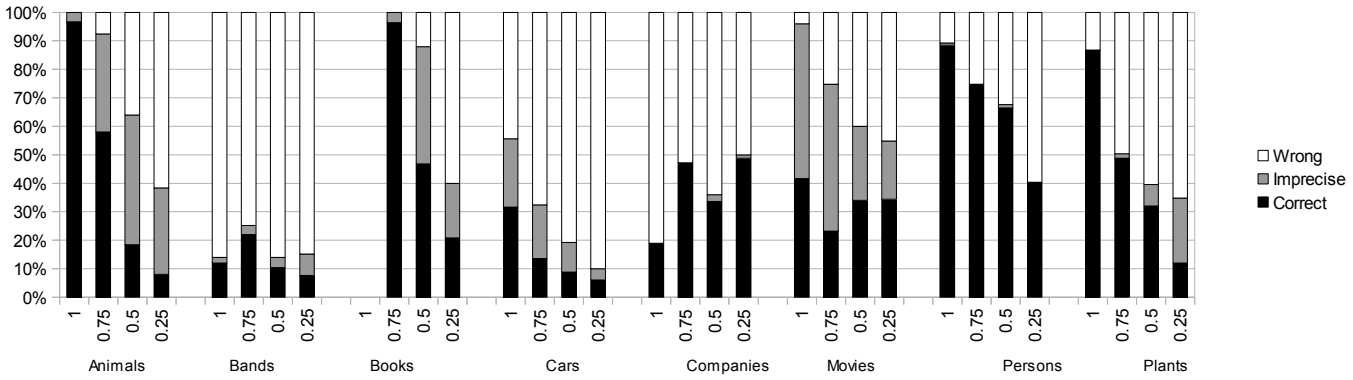

Figure 2: Evaluation of the mined rules on different subsets of DBpedia. For the books dataset, no rules with a minimum confidence of 1.0 were found.

Furthermore, it would be possible not only to add numerical values for imprecision, but also explanations on how the imprecise information was created (e.g., manually, by inductive learning, by crowd sourcing, etc.), which are a genuine part of provenance data.

6. PROCESSING IMPRECISE KNOWLEDGE

The Semantic Web needs to be able to deal with imprecise data and knowledge. Our envisioned imprecise semantic web can exploit existing uncertainty extensions of semantic web languages. In what follows, we first briefly summarize key approaches and then discuss their limitations, in terms of supporting the imprecise semantic web.

Handling imprecise data and knowledge in the semantic web is not a new topic. There is a series of international workshop on uncertainty reasoning for the semantic web.[11] The most well known approaches include fuzzy extensions [18, 19], probabilistic extensions [6, 11, 8] and possibilistic extensions [15]. Fuzzy extensions can be used to deal with vague concepts, such as 'Tall' and 'Dangerous'. The main difference between possibilistic extension and probabilistic extension lies in the fact that possibilistic logic is a qualitative representation of likelihood, whilst probabilistic extension is on quantitative aspects of likelihood. The W3C Uncertainty Reasoning for the World Wide Web Incubator Group (URW3-XG) published a report (see above) on existing solutions and methodologies of uncertainty extensions.

Reasoning on imprecise knowledge is useful for deriving new axioms based on known facts and imprecise axioms. For example, consider the two axioms:

$$\begin{aligned} BritishNovels &\sqsubseteq \exists author.EnglishNovelists \\ BritishNovels &\sqsubseteq EnglishBook \\ &\quad \sqcap \exists translation.GermanTranslation \end{aligned}$$

As discussed above, both of the axioms are imprecise. English novelists may occasionally publish works other than novels, and not every English book which has a translation to German is an English novel (although novels are more likely to be translated than other types of books). Given an English book b_0 which is both written by an English novelist and a German translation, an imprecise reasoner could calculate a high confidence score for the statement $BritishNovels(b_0)$ by combining both pieces of evidence.

This example shows how several pieces of imprecise knowledge can lead to new axioms which could not be derived using the traditional Semantic Web. However, there are some limitations of existing work on uncertainty extensions of semantic web languages:

- As discussed above, one first necessary step is an agreed standard representation of uncertain knowledge.
- Most uncertainty extensions of ontology languages suffer from high computational complexity; therefore, they might be ill-suited to provide real time reasoning services for semantic web applications. Having said that, there are some convincing work on scalable reasoning and/or query answering for the fuzzy extensions of tractable ontology languages, including fuzzy OWL 2 QL [12], fuzzy OWL 2 EL [17] and fuzzy OWL pD* [10]. More practical solutions should be provided on other extensions, such as probabilistic extensions. Given that RDF, as a sub-language of OWL 2 QL and OWL 2 RL, has been shown to be a good starting point [9, 21], it might be feasible to come up with some good solutions for tractable OWL profiles.
- Existing approaches deal with individual forms of uncertainty separately; however, they do not aggregate of different forms of uncertainty.

Solutions of the above limitations should be provided for the imprecise semantic web.

7. CONCLUSIONS

In this paper, we have discussed the use of imprecise knowledge on the Semantic Web. We have discussed an introductory examples showing that even in an unexpected domain such as emergency management, imprecise knowledge can add benefit to existing applications exploiting the Semantic Web. Furthermore, we have discussed a vision of a Semantic Web using both inductive and (imprecise) deductive techniques for serving information on different levels of confidence. Such a combination is useful for addressing individual applications' information needs defined by different trade-offs between recall and precision.

[11] http://http://c4i.gmu.edu/ursw/

Furthermore, we have shown given a glance at existing works in the area, showing an initial experiment for creating imprecise rules on DBpedia, and presented a short overview on approaches and initiatives that try to represent and process imprecise knowledge on the Semantic Web.

There are several crucial issues that have to be addressed to make the imprecise Semantic Web become useful. First of all, scalability is an important issue, both for discovering as well as for processing imprecise knowledge. When using inductive learning for discovering imprecise knowledge, *lazy learning* techniques [1] are a promising approach for addressing the scalability issue, since they do not require processing a whole dataset such as DBpedia in advance, but allow for deriving knowledge on demand for particular interesting instances. For processing imprecise knowledge, new breeds of reasoners dealing with imprecise knowledge are required to make that sort of knowledge usable for real-world use cases.

Assigning useful degrees of imprecision is a crucial point in creating and using imprecise knowledge. The confidence degrees delivered by a machine learning algorithm can be a candidate, but are not necessarily the best measure. Since human often rate the value of an imprecise rule differently from machines [13], augmenting our approach with humans in the loop, e.g., by using crowd-sourcing or games with a purpose, could help in creating more useful imprecise knowledge.

In this paper, we have introduced a number of examples for imprecise rules. These are rules that have a certain degree of imprecision, i.e., they are correct in a certain fraction of cases. However, other cases of imprecision could be useful. An example would be a rule such as *A movie with mostly American actors is an American Movie*, where the word *mostly* hints at a different type of imprecision.

For the future, we envision the creation of a model of imprecision that encompasses those different types of imprecision, and a stack of languages and tools that support a new imprecise Semantic Web based on those considerations. We are confident that such an imprecise Semantic Web could add much value to many knowledge-intense applications.

8. REFERENCES

[1] D. W. Aha. *Lazy Learning*. Springer, 1997.

[2] R. Dividino, S. Sizov, S. Staab, and B. Schueler. Querying for provenance, trust, uncertainty and other meta knowledge in rdf. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3), 2011.

[3] J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer, Berlin, Heidelberg, New York, 2007.

[4] D. A. Ferrucci. Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3.4):1:1 –1:15, may-june 2012.

[5] N. Guarino and C. A. Welty. An Overview of OntoClean. In *Handbook on Ontologies*, chapter 10, pages 201–220. Springer, 2nd edition edition, 2009.

[6] J. Heinsohn. Probabilistic Description Logics. In *Proc. of the 10th Annual Conference on Uncertainty in Artificial Intelligence*, pages 311–318, 1994.

[7] A. Kalyanpur, B. K. Boguraev, S. Patwardhan, J. W. Murdock, A. Lally, C. Welty, J. M. Prager, B. Coppola, A. Fokoue-Nkoutche, L. Zhang, Y. Pan, and Z. M. Qiu. Structured data and inference in deepqa. *IBM Journal of Research and Development*, 56(3.4):10:1 –10:14, may-june 2012.

[8] P. Klinov. Pronto: A non-monotonic probabilistic description logic reasoner. In *Proc. of ESWC2008*, pages 822–826, 2008.

[9] X. Lian and L. Chen. Efficient query answering in probabilistic rdf graphs. In *Prof. of SIGMOD Conference*, pages 157–168, 2011.

[10] C. Liu, G. Qi, H. Wang, and Y. Yu. Reasoning with Large Scale Ontologies in Fuzzy pD* Using MapReduce. *IEEE Comp. Int. Mag.*, 7(2):54–66, 2012.

[11] T. Lukasiewicz. Expressive probabilistic description logics. *J. of Artif. Intell.*, 172(6-7):852–883, 2008.

[12] J. Z. Pan, G. Stamou, G. Stoilos, S. Taylor, and E. Thomas. Scalable Querying Services over Fuzzy Ontologies. In *the Proc. of the 17th International World Wide Web Conference (WWW2008)*, 2008.

[13] H. Paulheim. Generating Possible Interpretations for Statistics from Linked Open Data. In *9th Extended Semantic Web Conference (ESWC)*, volume 7295 of *LNCS*, pages 560–574, 2012.

[14] H. Paulheim and J. Fürnkranz. Unsupervised Generation of Data Mining Features from Linked Open Data. In *International Conference on Web Intelligence and Semantics (WIMS'12)*, 2012.

[15] G. Qi, Q. Ji, J. Z. Pan, and J. Du. Extending description logics with uncertainty reasoning in possibilistic logic. *Int. J. Intell. Syst.*, 26(4):353–381, 2011.

[16] A. Schulz and H. Paulheim. Combining Government and Linked Open Data in Emergency Management. In *AI Mashup Challenge*, 2012.

[17] G. Stoilos, G. B. Stamou, and J. Z. Pan. Classifying Fuzzy Subsumption in Fuzzy-EL+. In *Proc. of the International Workshop on Description Logics*, 2008.

[18] G. Stoilos, G. B. Stamou, J. Z. Pan, V. Tzouvaras, and I. Horrocks. Reasoning with Very Expressive Fuzzy Description Logics. *J. Artif. Intell. Res. (JAIR)*, 30:273–320, 2007.

[19] U. Straccia. Reasoning within Fuzzy Description Logics. *J. Artif. Intell. Res. (JAIR)*, 14:137–166, 2001.

[20] J. Völker and M. Niepert. Statistical schema induction. In *Proceedings of the 8th extended semantic web conference on The semantic web: research and applications - Part I*, pages 124–138, 2011.

[21] A. Zimmermann, N. Lopes, A. Polleres, and U. Straccia. A general framework for representing, reasoning and querying with annotated semantic web data. *J. Web Sem.*, 11:72–95, 2012.