Data Management Issues on the Semantic Web

Oktie Hassanzadeh [*1], Anastasios Kementsietsidis [#2], Yannis Velegrakis [§3]

[*] *University of Toronto & IBM Research*
[1] oktie@cs.toronto.edu

[#] *IBM Research*
[2] akement@us.ibm.com

[§] *University of Trento*
[3] velgias@disi.unitn.eu

Abstract— **We provide an overview of the current data management research issues in the context of the Semantic Web. The objective is to introduce the audience into the area of the Semantic Web, and to highlight the fact that the area provides many interesting research opportunities for the data management community. A new model, the Resource Description Framework (RDF), coupled with a new query language, called SPARQL, lead us to revisit some classical data management problems, including efficient storage, query optimization, and data integration. These are problems that the Semantic Web community has only recently started to explore, and therefore the experience and long tradition of the database community can prove valuable. We target both experienced and novice researchers that are looking for a thorough presentation of the area and its key research topics.**

I. Introduction and Motivation

Over the last decade, we have witnessed a tremendous increase in the amount of data available on the Web. These data come from almost every field of human activity and include financial information, weather reports, news feeds, product information, geographical maps and the like. At the same time, the advent of Web 2.0 applications, such as Wikis, social networking sites and mashups, have brought new forms of data and have radically changed the nature of modern Web. They have transformed the Web from a publish-only environment into a vibrant place for information exchange. Naturally, Web users have also evolved and have gradually changed from being mainly data consumers to becoming active data producers themselves. Data dissemination agents have contributed further to the increase of the plethora of information on the Web.

Making sense of all these data is both important and increasingly challenging (by humans and machines alike). The goal of the Semantic Web [1] is to enhance the current Web by linking the data and enriching it with metadata in ways that facilitate both the understanding of data and the exploitation of its semantics [2], [3]. This explicit representation of data semantics is expected to enable a Web with new qualitative levels of service. In this context, new challenges emerge for semantic-aware data management systems. In particular, there is a growing need for representation models and successful communication mechanisms for the data semantics, for semantic-based (as opposed to value-based) query engines, and for semantic-aware presentation techniques of query results. Furthermore, the availability of Semantic Web services allows the development of modular, self-describing and self-contained applications that offer new challenges in retrieving, exchanging and composing data [4].

One key decision for the Semantic Web community, towards addressing its challenges and achieving its goals, was the adoption of the Resource Description Framework (RDF) [5] as the data model for representation and exchange in the new Web of data. RDF was chosen in spite of the popularity of the relational, and more recently of the XML model, and in spite of the maturity of these models in terms of both research and commercial system support. And for a while, there was limited availability of RDF data and RDF use was limited outside the Semantic Web academic cycle. However, this changed when governments (most notably from US [6] and UK [7]) and large companies and organizations have started using RDF as the *business* data model and representation format, either for semantic data integration (e.g. Pfizer [8]), search engine optimization and better product search (e.g. Best Buy [9]), or for representation of data from information extraction (e.g. BBC [10]). Further evidence of the RDF data explosion is clearly visible in the Linked Open Data (LOD) cloud [11], where Web data from a diverse set of domains like Wikipedia (DBpedia [12]), geographic locations, films (LinkedMDB [13] – a linked movie database), scientific data, and the like, are interlinked to provide one large data cloud. To store this data, a number of RDF data management systems have started to emerge, from which some rely on a relational back-end (e.g. Jena SDB [14], Virtuoso [15], Sesame [16], Oracle [17]), while others rely on newly developed *native* RDF storage techniques (e.g. Jena TDB [14], 4store [18]).

All this activity did not go unnoticed from the database community. The first research papers towards the development of database systems that can handle vast amounts of RDF data have already started to appear in major database conferences [19], [20], [21], [22]. Yet, there is a lot of work left to be done. Traditional data management approaches have a lot to gain by incorporating semantic information into their frameworks. Existing data integration, exchange and query solutions are typically based on the values of the data that are actually stored in the repositories, and not on the semantics of these values. Incorporation of semantics will improve query accuracy, offer better sharing and permit more efficient distribution services. Integration of new content, on-the-fly generation of mappings, querying on loosely structured data, keyword searching on structured data repositories, and entity identification, are some of the areas that can benefit from the presence of semantic knowledge alongside the data.

Although the data management community has traditionally focused on issues related to performance, scalability and query expressiveness, there have been strong evidences for an emerging interest towards the incorporation of semantic information in different data management processes. For instance, the University of Pennsylvania Workshop on Information Integration [23] has recognized that methods for query answering, matching and mapping need to be adapted to support ontologies. Ontologies can be successfully used to improve schema mapping tools [24]. Different techniques have also been developed for storing and querying different kinds of metadata, such as data quality parameters [25], provenance [26], superimposed information [27], or annotations [28], [29], while relevant values have been used to enhance database querying [30]. The data complexity and the difficulty of understanding its full semantics have led to the development of techniques that follow an approach similar to the one used in the Semantic Web. For instance, we have already seen work on keyword searching in relational databases [31], [32]. Furthermore, query processing in the recent area of dataspaces [33] assumes that very often users interact with the system in an exploratory nature, pose imprecise queries, define mappings in a *pay-as-you-go* fashion, and expect to receive highly heterogeneous and very often probabilistic query results. Even the data model typically used in dataspaces is triple-based, as in RDF, and the query mechanism is not based on structures like tuples, columns, elements, or values, but on the notion of entities, which is the backbone of the Semantic Web data.

Based on the above, it is becoming apparent that the Semantic Web brings new research challenges and opportunities for data engineering. First, by incorporating ontologies and other semantic models and reasoners in existing query and integration engines, we can improve the currently offered functionalities. Second, the RDF model offers new research directions to explore in storing and querying voluminous RDF data. One of our goals is to raise awareness of all these opportunities and motivate researchers in working towards these (and other) research directions. It is important to note here that the development of RDF is reminiscent of the story of XML. It also has been introduced by another community but turn out to be a popular topic for research that is active even to this day.

While motivating database researchers to work on Semantic Web problems is clearly one of our goals, a higher-level goal is to bring the Data Management and Semantic Web communities together. So far, the two communities have followed relatively independent lines of research (by publishing to different venues). The first has focused mainly on scalability and performance while the second on semantic interoperability of data on the Web. We believe that the success of modern Web applications highly depends on the contributions of both. Thus, we expect that we have excellent opportunities to bring together researchers and practitioners from both areas. The opening and invited talk of PODS 2009 [34] (a talk on a Web of concepts) has been an indication that the database community has started to keep an eye on the ideas and issues of the Semantic Web community. ICDE has been hosting for the last 2 years a Workshop dedicated on the intersection of the Semantic Web and of Databases: the Data Engineering Meets the Semantic Web (DESWeb). VLDB 2010 had a similar workshop Semantic Data Management (SemData) and SIGMOD 2011 hosted the Semantic Web Information Management (SWIM) workshop. The fact that the best paper award in VLDB 2007 was about storing RDF [21], and one of the PODS 2011 tutorials was on "Querying Semantic Web data with SPARQL" [35] are further indications of a growing interest.

II. Outline

Our objective is three-fold. First, to educate the audience on the goals of the Semantic Web and the approaches that have been taken so far. It intends to provide a generic overview of the different phases of publishing, querying, discovering, and integrating Semantic Web data, and draw the link to the respective works in the data management community. The second objective is to identify and analyze a list of the research opportunities in the area. The final objective is to provide a complete, unified and systematic presentation of all the efforts that have been proposed on managing RDF data and on querying using SPARQL.

We offer an introduction to the basic concepts of RDF and SPARQL. We provide an overview of the process of publishing data on the Web, and will use this process as a roadmap of the main topics that will be discussed in the tutorial. These topics include the following:

- RDF (schema) and semantics
- Schema/ontology matching and mappings
- Storage strategies for RDF
- (Efficient) query processing in SPARQL
- Semantic linking of RDF data
- Benchmarking performance and data quality

We include not only an overview of important related works but also a list of open research questions.

References

[1] N. Shadbolt, T. Berners-Lee, and W. Hall, "The semantic web revisited," *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 96–101, 2006.

[2] O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang, "A framework for semantic link discovery over relational data," in *CIKM*, 2009, pp. 1027–1036.

[3] S. Hassas Yeganeh, O. Hassanzadeh, and R. J. Miller, "Linking Semistructured Data on the Web," in *WebDB*, 2011.

[4] J. Cardoso, Ed., *Semantic Web Services: Theory, Tools, and Applications*, ser. Information Science Reference. IGI Global, 2007.

[5] Resource Description Framework (RDF), http://www.w3.org/RDF/.

[6] Data.gov, http://www.data.gov/.

[7] Data.gov.uk, http://www.data.gov.uk/.

[8] Making a Semantic Web Business Case at Pfizer, http://www.semanticweb.com/news/making_a_semantic_web_business_case_at_pfizer_161731.asp.

[9] Best Buy jump starts data web marketing, http://www.chiefmartec.com/2009/12/best-buy-jump-starts-data-web-marketing.html.

[10] BBC World Cup 2010 dynamic semantic publishing, http://www.bbc.co.uk/blogs/bbcinternet/2010/07/bbc_world_cup_2010_dynamic_sem.html.

[11] C. Bizer, T. Heath, and T. Berners-Lee, “Linked data - the story so far,” *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009.

[12] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, “DBpedia—a crystallization point for the web of data,” *Web Semantics*, vol. 7, no. 3, pp. 154–165, 2009.

[13] O. Hassanzadeh and M. P. Consens, “Linked movie data base,” in *Proceedings of the WWW2009 workshop on Linked Data on the Web (LDOW2009)*, 2009.

[14] “Jena: a semantic web framework for java,” http://jena.sourceforge.net.

[15] “Virtuoso universal server,” http://virtuoso.openlinksw.com.

[16] J. Broekstra and et al., “Sesame: A generic architecture for storing and querying RDF and RDF schema,” in *ISWC*, 2002.

[17] “Oracle database semantic technologies,” http://www.oracle.com.

[18] “4store - scalable RDF storage,” http://4store.org/.

[19] S. Alexaki, V. Christophides, G. Karvounarakis, and D. Plexousakis, “On storing voluminous rdf descriptions: The case of web portal catalogs,” in *WebDB*, 2001, pp. 43–48.

[20] C. M. Wyss and E. L. Robertson, “Relational languages for metadata integration,” *ACM Trans. Database Syst.*, vol. 30, no. 2, pp. 624–660, 2005.

[21] D. J. Abadi, A. Marcus, S. R. Madden, and K. Hollenbach, “Scalable semantic web data management using vertical partitioning,” in *VLDB*, 2007, pp. 411–422.

[22] T. Neumann and G. Weikum, “The rdf-3x engine for scalable management of rdf data,” *VLDB J.*, vol. 19, no. 1, pp. 91–113, 2010.

[23] University of Pennsylvania, “Report on the Workshop on Information Integration,” 2007.

[24] Y. An, A. Borgida, R. J. Miller, and J. Mylopoulos, “A semantic approach to discovering schema mapping expressions,” in *ICDE*, 2007, pp. 206–215.

[25] J. Widom, “Trio: A system for integrated management of data, accuracy, and lineage,” in *CIDR*, 2005, pp. 262–276.

[26] D. Bhagwat, L. Chiticariu, W. C. Tan, and G. Vijayvargiya, “An annotation management system for relational databases,” *VLDB J.*, vol. 14, no. 4, pp. 373–396, 2005.

[27] D. Maier and L. M. L. Delcambre, “Superimposed information for the internet,” in *WebDB (Informal Proceedings)*, 1999, pp. 1–9.

[28] F. Geerts, A. Kementsietsidis, and D. Milano, “Mondrian: Annotating and querying databases through colors and blocks,” in *ICDE*, 2006, p. 82.

[29] P. Buneman, S. Khanna, and W. C. Tan, “On propagation of deletions and annotations through views,” in *PODS*, 2002, pp. 150–158.

[30] S. Bergamaschi, C. Sartori, F. Guerra, and M. Orsini, “Extracting relevant attribute values for improved search,” *IEEE Internet Computing*, vol. 11, no. 5, pp. 26–35, 2007.

[31] J. X. Yu, L. Qin, and L. Chang, “Keyword Search in Databases,” *Synthesis Lectures on Data Management*, vol. 1, no. 1, pp. 1–155, 2009.

[32] A. Markowetz, Y. Yang, and D. Papadias, “Keyword Search Over Relational Tables And Streams,” *ACM TODS*, vol. 34, no. 3, pp. 1–51, 2009.

[33] X. Dong and A. Y. Halevy, “Indexing dataspaces,” in *SIGMOD Conference*, 2007, pp. 43–54.

[34] N. N. Dalvi, R. Kumar, B. Pang, R. Ramakrishnan, A. Tomkins, P. Bohannon, S. Keerthi, and S. Merugu, “A web of concepts,” in *PODS*, 2009, pp. 1–12.

[35] M. Arenas and J. Pérez, “Querying semantic web data with sparql,” in *PODS*, 2011, pp. 305–316.