Semantic Web Query Languages

James Bailey, University of Melbourne, http://www.csse.unimelb.edu.au/
François Bry, University of Munich, http://pms.ifi.lmu.de/
Tim Furche, University of Munich, http://pms.ifi.lmu.de/
Sebastian Schaffert, Salzburg Research, http://www.schaffert.eu/

SYNONYMS

Web Query Languages; Ontology Query Languages

DEFINITION

A number of formalisms have been proposed for representing data and meta data on the Semantic Web. In particular, RDF, Topic Maps and OWL allow one to describe relationships between data items, such as concept hierarchies and relations between the concepts. A key requirement for the Semantic Web is integrated access to data represented in any of these formalisms, as well the ability to also access data in the formalisms of the "standard Web", such as (X)HTML and XML. This data access is the objective of *Semantic Web query languages*. A wide range of query languages for the Semantic Web exist, ranging from i) pure "selection languages" with only limited expressivity, to fully-fledged reasoning languages, and ii) from query languages restricted to a certain data representation format, such as XML or RDF, to general purpose languages that support multiple data representation formats and allow simultaneous querying of data on both the standard and Semantic Web.

HISTORICAL BACKGROUND

The importance of Semantic Web query languages can be traced back to the roots of the Semantic Web itself. In its original conception, Tim Berners-Lee viewed the Semantic Web as allowing Web-based systems to take advantage of "intelligent" reasoning capabilities (4):

"The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users .... For the Semantic Web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning."

As the representation format for the Semantic Web has grown to cover XML, RDF, Topic Maps and OWL, there has been a corresponding growth in query languages that support access to each of these kinds of data.

SCIENTIFIC FUNDAMENTALS

A number of techniques have been developed to facilitate powerful data retrieval on the Semantic Web. This article follows the classification and taxonomy given in (1), which provides a comprehensive survey of the area. Several categories of query languages can be distinguished, according to the format of the Semantic Web data they can retrieve: i) Query languages for XML, ii) Query languages for Topic Maps, iii) Query languages for RDF and iv) Query languages for OWL.

XML Query Languages: Although not a primary format, it is possible to specify information on the Semantic Web using XML. Hence query languages for XML are applicable to Semantic Web data. Most query and

transformation languages for XML specify the structure of the data to retrieve using either of two approaches. In the navigational approach, path-based queries over the XML data are specified and the W3C standardised languages XPath, XSLT and XQuery are well known instances of this scheme. In the example based approach, query patterns are specified as "examples" of the XML data to be retrieved. Languages of this kind are mainly research languages, with some well known representatives being XML-QL (7) and Xcerpt (3; 15).

Topic Maps Query Languages: Several different query languages for Topic Maps data exist, with representatives being tolog (9), AsTMA (2) and Toma (11). tolog was selected as the initial straw man for the ISO Topic Maps Query Language and is inspired from logic programming, also having SQL style constructs. AsTMa is a functional query language, in the style of XQuery, whereas Toma combines both SQL syntax and path expressions for querying.

RDF Query Languages can be grouped into several families, that differ in aspects such as data model, expressivity, support for schema information, and type of queries. Principal among these families is the "SPARQL Family". This originated with the language SquishQL (12), which evolved into RDQL (12) and then was later extended to the language SPARQL (14). These languages all "regard RDF as triple data without schema or ontology information unless explicitly included in the RDF source". SPARQL currently has W3C Candidate Recommendation status as being the "Query Language for RDF". In particular, SPARQL has facilities to: i) Extract RDF subgraphs, ii) Construct a new RDF graph using data from the input RDF graph queried, iii) Return "descriptions" of the resources matching a query part, iv) Specify optional triple or graph query patterns (i.e., data that should contribute to an answer if present in the data queried, but whose absence does not prevent an answer being returned), v) Test the absence, or non-existence, of tuples. The general format of a SPARQL query is:

`PREFIX`	Specification of a name for a URI (like RDQL's `USING`)
`SELECT`	Returns all or some of the variables bound in the `WHERE` clause
`CONSTRUCT`	Returns a RDF graph with all or some of the variable bindings
`DESCRIBE`	Returns a "description" of the resources found
`ASK`	Returns whether a query pattern matches or not
`WHERE`	list, i.e., conjunction of query (triple or graph) patterns
`OPTIONAL`	list, i.e., conjunction of optional (triple or graph) patterns
`AND`	boolean expression (the filter to be applied to the result)

Another family of languages for RDF, the "RQL family", consists of the language RQL (10), and its extensions such as SeRQL (5). Common to this family is support for the combination of both data and schema querying. The RDF data model which is used slightly deviates from the standard data model for RDF and RDFS, disallowing cycles in the subsumption hierarchy and requiring both a domain and a range to be defined for each property. RQL itself has a large number of features and choices in syntactic constructs. This results in a complex, yet powerful language, which is far more expressive than other RDF query languages, especially those of the SPARQL family.

A number of other types of query languages for RDF also exist, using alternative paradigms. These include query languages using reactive rules, such as Algae (13) and deductive languages such as TRIPLE (6) and Xcerpt (3; 15). The last of these is noteworthy, as it combines querying on both the Standard Web (HTML/XML), with querying on the Semantic Web (e.g. RDF, TopicMaps) and also allows pattern-based, incomplete specification of queries.

OWL Query Languages: Query languages for OWL are still in their infancy compared to those for RDF. OWL-QL (8) is a well known language for querying OWL data and is an updated version of the DAML Query language. Its design targets the assistance of query-answering dialogues between computational agents on the Semantic Web. Unlike the RDF query languages, it focuses on the querying of schema rather than instance data. An RDF language such as SPARQL, may of course be used to query OWL data, but it is not well suited to the task, not being designed to be aware of OWL semantics.

Several themes emerge from considering the design of the various Semantic Web Query languages (1).

- ● *Choice of querying paradigm:* Semantic Web query languages express basic queries using either the path based (navigational) or logic based (positional) paradigm.

- ● *Choice of variable type:* When Semantic Web query languages have variables, they almost always are logical variables, as opposed to variables in imperative programming languages.

- ● *Provision of Referential Transparency and Answer-Closure.* Referential Transparency (i.e., within the same scope, an expression always means the same), a well known trait of declarative languages, is striven for by Semantic Web query languages. Answer closedness is a property that allows answers to queries to be themselves used as input to queries and is a key design principle of the languages SPARQL and Xcerpt.

- ● *Degree of Incompleteness*: Many Semantic Web query languages offer means for incomplete specifications of queries, a reflection of the semi structured nature of data on the Semantic Web.

- ● *Reasoning Capabilities.* Interestingly, but not surprisingly, not all XML query languages have views, rules, or similar concepts allowing the specification of other forms of reasoning. Surprisingly, the same holds true of RDF query languages. Many authors of RDF query languages see deduction and reasoning to be a feature of an underlying RDF store offering materialisation, i.e., completion of RDF data with derivable data prior to query evaluation. This is surprising, because one might expect many Semantic Web applications to access not only one RDF data store at one Web site, but instead many RDF data stores at different Web sites and to draw conclusions combining data from different stores.

KEY APPLICATIONS*

Like classical query languages such as SQL, the first key application of Semantic Web query languages is the efficient and scalable access, classification, analysis and transformation of large collections of data in a Web format such as XML, RDF, OWL, or Topic Maps. Whereas classical query languages are most often used for accessing a single, centralised database, Semantic Web query languages need to be able to access also remote databases and data sources. This opens up new application scenarios, potentially utilising any of the vast number of the data sources available on the Web.

For example, one might query researcher and publication information integrated over various sources, such as DBLP, Citeseer, IEEE and Cordis, combine that data with course and lecturer information from the Semantic Web School and then even further correlate it with the US census data. All these resources would be far too large to download individually and query locally, but they provide interfaces known as *endpoints*, that can be used to select the relevant portions via a Semantic Web query interface. Another example application is the W3C Amaya browser, which can be used to enrich Web pages visited by a user, with annotations contained in remote data sources. The annotations relevant to a given Web page are accessed by querying an annotation server using Algae (13), an RDF query language similar to SPARQL. In such scenarios, the ability of RDF (and to some extent, XML) to define the names and concepts used in a database, reason about them and to map them to names and concepts used in another database, is essential. This clearly separates the use of Semantic Web query languages from the use of classical query languages for centralised databases.

Increasingly, current Web applications (often referred to as Web 2.0 applications) contain a Javascript-based user interface which is separate from the data processed by the application itself. Thus, the user interface can be loaded once and data then requested from the origin server or other data sources on the Web as required. Web query languages for XML, RDF, JSON and Topic Maps are now becoming recognized as the ideal interfaces between the client user interfaces of Web 2.0 applications and data sources, since they can target just the data that is needed in the current state of the application. Web query languages allow flexible, but fine-grained access to the required data, rather than the coarse-grained access provided by other solutions.

FUTURE DIRECTIONS

Most RDF query languages are RDF-specific, and even specifically designed for one RDF serialisation, which of course limits their applicability. It is to be hoped that in the future, there will be an evolution towards data format "versatile" languages, capable of easily accommodating XML, RDF, Topic Maps and OWL, without requiring "serialisation consciousness" from the programmer.

The method of query evaluation in current Semantic Web query languages is either backtracking-free logic programming (as used by positional languages) or set-oriented functional query evaluation. It seems likely these two paradigms may converge in future Semantic Web query languages. Language engineering issues, such as abstract data types and static type checking, modules, polymorphism, and abstract machines, have not yet made their way into Semantic Web query languages, as they did not in database query languages. This situation opens avenues for promising research of great practical, as well as theoretical relevance.

DATA SETS*

There are a number of SPARQL endpoints that can be browsed on the Web. These provide RDF data which can be viewed and then queried using a SPARQL client:

- The 2000 US Census Data endpoint: http://www.rdfabout.com/demo/census/
- The Semantic Web School endpoint: http://sparql.semantic-web.at/
- A compilation of endpoints including DBLP, Citeseer, IEEE and Cordis: http://www.rkbexplorer.com/

A collection of concrete query language use cases for accessing RDF data can be found in the W3C RDF Use Case document at http://www.w3.org/TR/rdf-dawg-uc/. A use case collection is also included in (1).

URL TO CODE*

The D2R Server is a utility for publishing relational databases on the Semantic Web and can be found at: http://sites.wiwiss.fu-berlin.de/suhl/bizer/d2r-server/

Annotea is a project that aims to assist collaboration via shared semantic meta-data. The Annotea-Server with Amaya Browser and Algae QL can found at: http://www.w3.org/2001/Annotea/

CROSS REFERENCE*

HTML, Ontologies, OWL, RDF, RDFS, Semantic Web, Topic Maps, XML, XPath, XQuery, XSLT.

RECOMMENDED READING

[1] James Bailey, François Bry, Tim Furche, and Sebastian Schaffert. Web and Semantic Web Query Languages: A Survey. In *Reasoning Web, LNCS 3564, Springer-Verlag*, pages 35–133, 2005.

[2] Robert Barta. AsTMa 1.3 Language Specification. Technical report, Bond University, 2003.

[3] Sacha Berger, François Bry, Tim Furche, Benedikt Linse, and Andreas Schroeder. Beyond XML and RDF: The versatile Web query language Xcerpt. In *Proceedings of the 15th International World Wide Web Conference (WWW)*, pages 1053–1054, 2006.

[4] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web—A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, pages 29–37, 2001.

[5] Jeen Broekstra and Arjohn Kampman. SeRQL: A Second Generation RDF Query Language. In *Proc. SWAD-Europe Workshop on Semantic Web Storage and Retrieval*, 2003.

[6] Stefan Decker, Michael Sintek, Andreas Billig, Nicola Henze, Peter Dolog, Wolfgang Nejdl, Andreas Harth, Andreas Leicher, Susanne Busse, José Luis Ambite, Matthew Weathers, Gustaf Neumann, and Uwe Zdun. TRIPLE - an RDF rule language with context and use cases. In *Proc. of Rule Languages for Interoperability*, 2005.

[7] Alin Deutsch, Mary Fernandez, Daniela Florescu, Alon Levy, and Dan Suciu. A Query Language for XML. *Computer Networks*, 31(11–16):1155–1169, 1999.

[8] Richard Fikes, Patrick Hayes, and Ian Horrocks. OWL-QL – A Language for Deductive Query Answering on the Semantic Web. *Journal of Web Semantics*, 2(1):19–29, 2004.

[9] Lars Marius Garshol. tolog - a topic maps query language. In *Proceedings of the First International Workshop on Topic Maps Research and Applications (TMRA) LNCS 3873*, pages 183–196, 2005.

[10] Gregory Karvounarakis, Aimilia Magkanaraki, Sophia Alexaki, Vassilis Christophides, Dimitris Plexousakis, Michel Scholl, and Karsten Tolle. Querying the Semantic Web with RQL. *Computer Networks and ISDN Systems Journal*, 42(5):617–640, August 2003.

[11] Martin Lacher and Stefan Decker. RDF, Topic Maps, and the Semantic Web. *Markup Languages: Theory and Practice*, 3(3):313–331, December 2001.

[12] Libby Miller, Andy Seaborne, and Alberto Reggiori. Three implementations of SquishQL, a simple RDF query language. In *Proc. of the International Semantic Web Conference*, pages 423–435, 2002.

[13] Eric Prud'hommeaux. Algae RDF Query Language. http://www.w3.org/2004/05/06-Algae/, 2004.

[14] Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language for RDF. Candidate recommendation, W3C, June 2007, http://www.w3.org/TR/rdf-sparql-query/.

[15] Sebastian Schaffert and François Bry. Querying the Web Reconsidered: A Practical Introduction to Xcerpt. In *Proc. Extreme Markup Languages*, August, 2004.