RDF graphs to be enhanced by declaring semantic constraints between classes and properties. Such constraints are interpreted under the open-world assumption [2], propagating instances from one relationship to another. For example, in a database where "Océan is a pancake house" and "a pancake house is a restaurant", we can infer that "Océan is a restaurant". Querying the database for restaurants should also return all the pancake houses! Our framework is centreed on RDF, thus it natively supports RDF semantics when querying.

Explore data though analytics

Even for an experienced database developer, understanding a new dataset is always a challenge, as each brings its own set of features, which may be particularly subtle in the case of semantic-endowed data such as RDF. To facilitate their understanding, datasets are generally published along with a schema. But how does one understand the schema? Contemporary published schemas tend to be complex and can be seen as datasets in their own rights. While still small compared to the data, they can be real puzzles for the analyst. Working with RDF, one can seamlessly query the schema and the data, for example ask for all the relationships linking people to other entities. Our model allows analytics not only over the data, but over the schema as well.

Data cubes, no longer a dictatorship

In order to perform data warehouse analysis, one must first establish the dimensions and measures according to which to analyze the facts. Data cubes are built as a result of aggregating the measures along the dimensions. For instance, when asking "what are the total sales for region Lorraine in autumn 2013?", the sales are a measure, while region and period represent dimensions. However, such a warehouse cannot answer the query "how many regions registered sales in autumn 2013?", since region is a dimension, and relational data cubes do not allow aggregating over the dimensions. In contrast, our framework is very flexible, allowing a choice of dimensions and measures at data cube (query) time, not at data warehouse design time.

WaRG models the analytical schema of an RDF warehouse as a graph. Each node represents a set of facts, modelling a new RDF class. The edges connecting these nodes are defined independently and correspond to new RDF properties. The instances of these classes and properties, modelling the data warehouse contents to be further analyzed, are intentionally defined in the schema, following the well-know "Global As View" approach for data integration. For more details we refer the interested reader to [3].

Our ongoing work includes RDF analytical schema recommendation and efficient algorithms for massively parallel RDF analytics.

Link: https://team.inria.fr/oak/warg/

References:

[1] W3C, Resource Description Framework, http://www.w3.org/RDF/.
[2] S. Abiteboul, R. Hull, V. Vianu: "Foundations of Databases", Addison-Wesley, 1995.
[3] D. Colazzo, F. Goasdoué, I. Manolescu, A. Roatiș: "Warehousing RDF Graphs", in "Bases de Données Avancées", 2013, http://hal.inria.fr/docs/00/86/86/16/PDF/paper.pdf.

Please contact:

Alexandra Roatiș
Inria Saclay and LRI, Université Paris-Sud
http://www.lri.fr/~roatis/
E-mail: alexandra.roatis@lri.fr

The Web Science Observatory - The Challenges of Analytics over Distributed Linked Data Infrastructures

by Wendy Hall, Thanassis Tiropanis, Ramine Tinati, Xin Wang, Markus Luczak-Rösch and Elena Simperl

Linked data technologies provide advantages in terms of interoperability and integration, which, in certain cases, come at the cost of performance. The Web Observatory, a global Web Science research project, is providing a benchmark infrastructure to understand and address the challenges of analytics on distributed Linked Data infrastructures.

The evolution from the Web of documents to the Web of data, social networks, and crowdsourcing has opened up new opportunities for innovation driven by analytics on public datasets, on online social network activity, as well as on corporate or private datasets [1]. These opportunities are evidenced by the emergence of a number of Web Observatories [2] that not only collate and archive, but attempt to provide analytics on such datasets (http://www.nextcenter.org:8080/ugcp/live/observer, http://www.truthy.indiana.edu). These developments have been accompanied by an evolution of data literacy that has been taking place in parallel; people are no longer just consumers of documents on the Web but also contributors of content, data, and applications. The open data movement has shown the potential of crowdsourcing data and applications, while linked data technologies have established their potential for dataset interoperability and integration. The basic associative structure of linked data has advanced the idea of regarding the Web as a global dataspace that, in the future, may be queried as if it were one giant database. Dataspaces are a generic data management abstraction, which helps to derive and maintain relationships between a large number of heterogeneous but interrelated data sources. Relationships are regarded as integration hints until they are reviewed and approved. The overall data management and integration effort is distributed among various

stakeholders and performed in an evolutionary fashion [3].

But is that enough? Analytics on the Web of data and social networks present developers with a dilemma between data warehouse architectures with high performance, big data analytics on the one hand, and analytics on distributed, diverse, separately maintained datasets with potentially lower performance on the other hand. This is where certain challenges for Linked Data emerge; they concern both the representation of diverse datasets as Linked Data and the performance analytics over, potentially distributed, Linked Data infrastructures.

Some of those challenges are explored as part of the Web Observatory (WO) project, which was initiated under the auspices of the Web Science Trust (WST, http://webscience.org). The WO aims to provide a distributed global resource in which datasets, analytic tools and cross-disciplinary methodologies can be shared and combined to foster interdisciplinary research in the context of Web Science [2]. This effort involves many different communities including the Web Science Trust network of laboratories (WSTnet, http://webscience.org/wstnet-laboratories/), other major research groups in this area, government agencies, public sector institutions, and the industry. The WO project has grown as a bottom-up effort, and aims to enable interoperability among datasets and analytics on a large, distributed scale. It involves the crowdsourcing, publication and sharing of both datasets, and of querying, analytic or visualisation tools. At the same time, it involves the development of appropriate standards to enable the discovery, use, combination and persistence of those resources; effort in the direction of standards is already underway in the W3C Web Observatory community group (http://www.w3.org/community/webobservatory/).

The Web Observatory infrastructure that is deployed at the University of Southampton includes different datastore technologies, however, a significant part of it are Linked Data stores. To that end, the infrastructural deployment involves (i) converting large datasets into 5-star linked data formats (2 http://www.w3.org/DesignIssues/LinkedData.html), (ii) supporting distributed queries over Linked Data stores, and

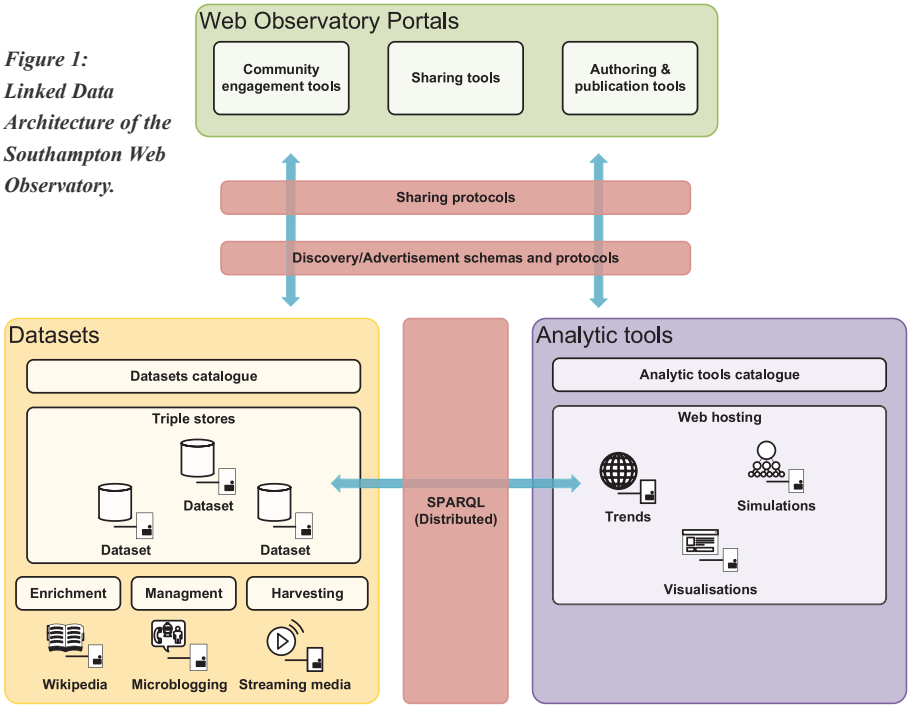

Figure 1: Linked Data Architecture of the Southampton Web Observatory.

(iii) supporting analytic and visualisation tools over distributed data stores and datasets. Essentially, this infrastructure (shown in Figure 1) will not only provide for valuable analytics over distributed resources but it will also provide for benchmarking analytics on Linked Data.

The datasets that are made available in Linked Data formats for this infrastructure include open data, licensed data and private data that can be accessed only by authorised parties. They include microblogging activity, access to Web 2.0 services (such as Wikipedia), Wellbeing data, and Web of data services (such as USEWOD). Given the volume of some of these datasets (e.g. microblogging data) and the use of diverse and distributed data stores, different approaches to optimisation for the performance of analytics are explored. These involve:

- Representation of datasets in Linked Data formats (e.g. representing microblogging activity, weighted graphs)
- Representation of qualitative data in Linked Data formats
- Enrichment of Linked Data stores for analytics
- Distributed Linked Data query optimisation
- Dataset management and security in Linked Data stores
- Dataset provenance and preservation.

As well as creating a platform for Web Science research, the Web Observatory project will provide insights into performance and optimisation for analytics over distributed Linked Data infrastructures on varying scales of data store size and distribution. It will propose appropriate approaches to data representation, distributed queries and integration in such environments. Contributions to standardization for Web Observatories and Linked Data can also be anticipated. The biggest test for the this activity is the extent to which Linked Data infrastructures can be used not only in an efficient manner but also to support researchers across disciplines engaging in interdisciplinary work.

Link:
http://webscience.org/web-observatory/

References:

[1] W. Hall, T. Tiropanis: "Web evolution and Web Science", Computer Networks, 56(18), 3859–3865, 2012, doi:10.1016/j.comnet.2012.10.004

[2] T. Tiropanis et al: "The Web Science Observatory", Intelligent Systems, IEEE, 28(2), 100–104, 2013, http://dx.doi.org/10.1109/MIS.2013.50

[3] T. Heath, C. Bizer: "Linked Data: Evolving the Web into a Global Data Space (1st edition)", Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool, http://linkeddatabook.com/editions/1.0/

Please contact:

Thanassis Tiropanis
University of Southampton
E-mail: tt2@ecs.soton.ac.uk