Data exchange for annotated documents

François Goasdoué and Ioana Manolescu

LRI (Université Paris-Sud) and Oak team, Inria Saclay

February 28, 2013

Context: annotated documents Tools for authoring electronic content and sharing it through the Internet are very widely adopted by now. First blogs, and then social networks, grew more or less in parallel with the major media providers' move towards allowing users to record their opinions next to the articles. These technical means to hold back-and-forth conversations, as well as the Linked Open Data movement[1], in which public and private institutions publish their data for transparency and accountability reasons, have sparked new usages of the Web and in particular of Web data. As another example, public figures rely heavily on social networks to communicate their positions to the public, collect feedback, survey opinion trends etc. At the same time, individuals interested in a particular topic or action (e.g., the economic impact of a certain policy) can comb the Web for bits of information, connect, interpret, annotate and re-share them. Such data gathering and fact checking have come at the core of "data journalism", pioneered, e.g., in Europe by The Guardian[2] and growing through efforts such as FactCheck[3], Politifact[4], and similar French sites[5].

The increased interest of the public at large for reading, analyzing, commenting, and crossing information raises the need for integrated, generic and open tools to gather and share content and annotations. Wide-audience tools must be *integrated* to spare the user the effort of gluing several software components. They need to be *generic* to handle many types of input. For instance, if one wants to analyze some tweet stream, one very likely needs to archive and analyze Web pages, RSS feeds or domain ontologies. Finally, they should be *open* w.r.t. the supported data formats, so that they are not only information sinks but also information sources, and w.r.t. the architecture, so that they can be easily customized and extended.

Approach: the XR model To address this need, within the Oak team, we have proposed XR, a language and model based on the W3C standards XML for representing structured Web documents, RDF for encoding facts and more generally Semantic Web data, and RDF Schema for encoding knowledge (i.e., ontologies) [1], under the single paradigm of *annotated documents*. In [1, 2] we have defined the language and its associated query language, XRQ; two works currently under evaluation focus on the efficient evaluation of XR queries, and on applying XR in a scenario where controversial facts and theories are discussed through the Web. More details about our project can be found at http://tripleo.saclay.inria.fr/xr.

Subject description The proposed topic consists of devising a *data exchange language for XR*, possibly inspired from the well-established body of research on *schema mapping* languages. The idea is the following.

In a distributed setting, some users authors documents while others users and/or program author annotations over them. In turn, other users have other views of the world whereas they are interested in certain subsets (or certain combination queries) to be evaluated over existing XR data sources, that is, over existing documents and/or annotations. Further, some users may express how existing data sources relate to each other, e.g., how annotations or content from different users are correlated. These tasks are typically accomplished through *data exchange* programs, which consists of *mapping rules*. An early formal study of the data exchange problem is provided in [3]; the problem has been intensely studied since.

The work to be accomplished can be decomposed in the following steps:

- Getting familiar with the basic technologies such as XML and RDF, the XR model and the XRQ language;
- Studying the closest works in the literature that address data exchange for XML and/or RDF;

[1] http://linkeddata.org
[2] http://guardian.co.uk/data
[3] http://www.factcheck.org
[4] http://www.politifact.org
[5] http://www.liberation.fr/desintox, http://decodeurs.blog.lemonde.fr

- Proposing a mapping language suitable to the XR model, which must consider both the explicit querying of documents and annotations, and the implicit reasoning present in the XR model through its RDF component;
- Studying the formal properties of data exchange programs specified in this language;
- Realizing a prototype implementation of an XR data exchange system.

Interesting candidates should have good background in databases and/or knowledge representation. Familiarity with Web standards (XML, RDF) is a plus, but these can be learned on site. Demonstrable good academic results in topics such as logics, artifficial intelligence and advanced algorithms are also strong arguments in favor of a candidate.

The project is part of our activity within the DigiCosme *LabEx* (French network of excellency in IT in the larger Saclay area, including Université de Paris Sud), to which the team participates. The research can be carried in either French or English. For more information:

- Team web site: http://team.saclay.inria.fr/oak
- Supervisors' coordinates: fg@lri.fr (http://www.lri.fr/ goasdoue), ioana.manolescu@inria.fr (http://pages.saclay.inria.fr/ioana.manolescu)
- More of our projects in the area: http://tripleo.saclay.inria.fr/xr
- DigiCosme LabEx web site: http://digicosme.lri.fr

Bibliography

[1] F. Goasdoué, K. Karanasos, Y. Katsis, J. Leblay, I. Manolescu, and S. Zampetakis. Growing Triples on Trees: an XML-RDF Hybrid Model for Annotated Documents. In M. Brambilla, F. Casati, and S. Ceri, editors, *First International Workshop on Searching and Integrating New Web Data Sources*, Seattle, United States, Sept. 2011.

[2] F. Goasdoué, K. Karanasos, Y. Katsis, J. Leblay, I. Manolescu, and S. Zampetakis. Growing Triples on Trees: an XML-RDF Hybrid Model for Annotated Documents. In *Journées de Bases de Données Avancées*, Rabat, Morocco, Oct. 2011.

[3] P. G. Kolaitis, J. Panttaja, and W. C. Tan. The complexity of data exchange. In *PODS*, pages 30–39, 2006.

Five recent team publications related to the topic

[1] Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, and Pierre Senellart. *Web Data Management and Distribution*. Cambridge University Press, Dec 2011.

[2] François Goasdoué and Marie-Christine Rousset. Robust module-based data management. *IEEE Transactions on Knowledge and Engineering*, December 2011.

[3] François Goasdoué, Konstantinos Karanasos, Julien Leblay, and Ioana Manolescu. View Selection in Semantic Web Databases. *Proceedings of the Very Large Databases Endowment (PVLDB)*, 5(2):97–108, 2011.

[4] Asterios Katsifodimos, Ioana Manolescu, and Vasilis Vassalos. Materialized view selection for XQuery. In *ACM SIGMOD International Conference on Management of Data*, pages 565–576, 2012.

[5] Ioana Manolescu, Konstantinos Karanasos, Vasilis Vassalos, and Spyros Zoupanos. Efficient XQuery Rewriting using Multiple Views. In *IEEE International Conference on Data Engineering (ICDE)*, 2011.