# Sharing Distributed Knowledge on the Web

Serge Abiteboul

**Collège de France, INRIA Saclay & ENS Cachan, France**

**Abstract**

To share information, we propose to see the Web as a knowledge base consisting of distributed logical facts and rules. Our objective is to help Web users finding information, as well as controlling their own, using automated reasoning over this knowledge base towards improving the quality of service and of data. For this, we introduce Webdamlog, a Datalog-style language with rule delegation. We mention the implementation of a system to support this language as well as standard communications and security protocols.

**1998 ACM Subject Classification** H.2 Database

**Keywords and phrases** Knowledge base, distributed data, world wide web, deduction.

**Digital Object Identifier** 10.4230/LIPIcs.CSL.2012.6

**Category** Invited Talk

## 1 Overview

Information of interest may be found on the Web in a variety of forms, in many systems, and with different access protocols. Today, the control and management of the diversity of data and tasks in this setting are beyond the skills of casual users. Facing similar issues, companies see the cost of managing and integrating information skyrocketing. We are interested here in the management of Web data. Our focus is not on harvesting all the data of a particular user or a particular application domain and then managing it in a centralized manner. Instead, we are concerned with the management of Web data *in place* in a distributed manner, with a possibly large number of autonomous, heterogeneous systems collaborating to support certain tasks. We describe ongoing works concerned with the foundations of such data management based on *declarative distributed data management*.

Centralized data management has matured with relational database systems, enabled by the combination and cooperation of a very active research community and a very successful industry. The success of the field rests on solid formal foundations that combine existing tools, e.g., first-order logic for specifying queries and dependencies, with others that were developed from scratch, e.g., query optimization or concurrency control. As a result, centralized data management systems are now very reliable and the corresponding science is well-developed.

Now consider a user of the Web today. Such a user typically needs to manage the following kinds of information: data (documents, photos, music, etc.), metadata (personal ontologies, other's, etc.), data localization (e.g., where friends place their photos), access rights (e.g., list of friends who have access to private photos), credentials (login, passwords, etc.), temporal and provenance information, other kinds of knowledge (replication policy, trust in others, beliefs, etc.). The information resides on many devices (smartphone, laptop, TV box, etc.), many systems (mailers, blogs, Web sites, etc.), many social networks (Facebook, Picasa, etc.) as well as in the "data rings" [5] of family members, friends, associations, companies, and health, tax or other organizations.

Computer Science Logic 2012 (CSL'12).
Editors: Patrick Cégielski, Arnaud Durand; pp. 6–8

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Clearly, relational databases are not providing adequate support for managing the diversity of such information. First, the exchange standard for the Web is based on data trees (HTML, XML, JSON) and not on relations. Then, the information is by nature distributed between huge number of autonomous systems, unlike in relational systems where it is centralized or distributed within a few tightly controlled machines. Finally, a critical dimension of the problem is the imprecise, uncertain, noisy, possibly contradicting nature of data in this setting.

Our thesis is that managing the richness and diversity of data residing on the Web can be tamed using a holistic approach based on a *distributed knowledge base*. Our approach is to represent all Web information as *logical facts*, and Web data management tasks as *logical rules*. A variety of complex data management tasks that currently require intense work and deep expertise may then greatly benefit from the automatic reasoning provided by inference engines, operating over the distributed Web knowledge base.

The kinds of reasoning tasks we are envisioning, and that are to be captured by *rules*, therefore include:

- Accessing information. Knowledge is used to localize data, e.g., find which systems hold the information of interest. Also, when a new source of information is discovered, some simple reasoning may be required to understand how it should be used, and how to obtain access rights.
- Peer's policy. Each peer specifies its own policy, which includes choices such as where to store/search for particular information, which data to serve to other peers, and which data to replicate. Such policies in social networks are typically defined based on information such as the composition of user groups ("circles" in Google+ terminology).
- Ontology processing. A particular source may structure its information in a particular manner or even describe it using RDF or RDFS. Reasoning is necessary to query this information. In particular, when accessing different information sources, knowledge is needed to align their concepts and relations.
- Quality management. Reasoning may be needed to assess the truthfulness of some data or to choose between contradicting information. This is related to evaluating the confidence one has in some data, the trust in sources, and, more generally, the beliefs of a particular peer or user.
- Knowledge acquisition and dissemination. These are central issues in this context. Knowledge acquisition, i.e., acquiring new facts and rules and evaluating their quality, should provide principled mechanisms that protect against (i) accepting any kind of information that is published by anyone on the Web and (ii) revising opinions too easily and in an ad-hoc manner (e.g., believing the last person who spoke).

We propose to express the peer's logic in Webdamlog, a datalog-style rule-based language. In WebdamLog, peers exchange facts (for information) and rules (in place of code). The use of declarative rules provides the following advantages:

- Peers may perform automatic reasoning using the available knowledge;
- Because the model is formally defined, it becomes possible to prove (or disprove) desirable properties in the spirit of [1] that uses logic to describe access control protocols;
- Because the model is based on a Datalog-style language, it can benefit from optimization techniques, as in [7, 9];
- Because our model represents provenance [3] and time, we can better control the quality of data; and
- Because the model is general, it can represent wide variety of scenarios and protocoles, which is the reality of today's Web.

**Acknowledgment.** This work has been partially funded by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013); ERC grant Webdam, agreement 226513. A more detailed presentations of the results of the project may be found on the Webdam Web site at `http://webdam.inria.fr/`.

## References

1 Martín Abadi. Logic in Access Control (Tutorial Notes). In Alessandro Aldini, Gilles Barthe, and Roberto Gorrieri, editors, *Foundations of Security Analysis and Design V*, volume 5705, chapter 5, pages 145–165. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

2 Serge Abiteboul, Meghyn Bienvenu, Alban Galland, and Emilien Antoine. A rule-based language for Web data management. In *Proceedings of the Symposium on Principles of Database Systems*, 2011.

3 Serge Abiteboul, Alban Galland, and Neoklis Polyzotis. A model for web information management with access control. In *Proceedings of the International Workshop on the Web and Databases*, 2011.

4 Emilien Antoine, Alban Galland, Kristian Lyngbaek, Amélie Marian, and Neoklis Polyzotis. [Demo] Social Networking on top of the WebdamExchange System. In *Proceedings of the International Conference on Data Engineering*, 2011.

5 Serge Abiteboul, Neoklis Polyzotis, The Data Ring: Community Content Sharing, Conference on Innovative Data Systems Research, 2007.

6 Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. Corroborating Information from Disagreeing Views. In *Proceedings of the International Conference on Web Search and Data Mining*, 2010.

7 Joseph M. Hellerstein. Datalog redux: experience and conjecture. In *Proceedings of the Symposium on Principles of Database Systems*, 2010.

8 Serge Abiteboul, T-H. Hubert Chan, Evgeny Kharlamov, Werner Nutt, and Pierre Senellart, Capturing Continuous Data and Answering Aggregate Queries in Probabilistic XML. In *ACM Transactions on Database Systems, vol. 36, nº 4*, 2011.

9 Boon Thau Loo, Tyson Condie, Minos Garofalakis, David E. Gay, Joseph M. Hellerstein, Petros Maniatis, Raghu Ramakrishnan, Timothy Roscoe, and Ion Stoica. Declarative networking: language, execution and optimization. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, SIGMOD '06, pages 97–108, New York, NY, USA, 2006. ACM.