THE WEB OBSERVATORY:

A Middle Layer for Broad Data

Thanassis Tiropanis,[1] Wendy Hall,[1] James Hendler,[2] and Christian de Larrinaga[3]

The Web Observatory Project

The Web Observatory project[1] is a global effort that is being led by the Web Science Trust,[2] its network of WSTnet laboratories, and the wider Web Science community. The goal of this project is to create a global distributed infrastructure that will foster communities exchanging and using each other's web-related datasets as well as sharing analytic applications for research and business web applications.[3] It will provide the means to observe the digital planet, explore its processes, and understand their impact on different sectors of human activity.

The project is creating a network of separate web observatories, collections of datasets and tools for analyzing data about the Web and its use, each with their own use community. This allows researchers across the world to develop and share data, analytic approaches, publications related to their datasets, and tools (Fig. 1). The network of web observatories aims to bridge the gap that currently exists between big data analytics and the rapidly growing web of "broad data,"[4] making it difficult for a large number of people to engage with them.

The Gap Between Big Data and the Web of Data

The promise of big data has been well established for various industrial sectors in financial, innovation, and productivity terms. The technical challenges of coping especially with the velocity and volume aspects of big data are being addressed by ongoing research. However, often there are underpinning assumptions in those techniques. One assumption is that the data stores on which big data techniques are applied are all accessible under a single administrative domain. Another assumption is that the data analytics team employed by any particular enterprise is the one best able to drive innovation from their datasets.

On the other hand, the evolution of the Web to the web of data has taught us a number of lessons. A first lesson is that decentralized infrastructures have more potential to generate externalities and foster innovation,[5] and contribute to innovation and wealth. The second lesson is that people are the decisive actors in the network effects that enable infrastructures and innovation to spread; it is people who co-create the Web by contributing and engaging with documents, content, data, and applications.[6] However, many of today's big data infrastructures are limited to a centralized or distributed infrastructure that is under a single administrative domain, and the data analytics team that engages with the datasets involves a limited number of people who have been granted access to it. This is the current model in enterprises engaging with analytics, and there is currently competition among those enterprises to attract analytic talent. As McKinsey expects that there will be shortage of analytic talent in the future,[7] one can further conclude that small and medium enterprises (SMEs) will be at a disadvantageous position in this competition for analytic talent. This could be detrimental to innovation that is increasingly data-driven; it has long been supported that SMEs are a significant contributor to innovation.[8]

In addition, many of the datasets that are used in various big data installations are common, but they are replicated on each installation. They can be open data or licensed data of the same provenance. The cost of maintaining multiple copies of the datasets is significant on a global scale, as is the cost of cleaning

[1]*Southampton University, Southampton, United Kingdom.*
[2]*Rensselaer Polytechnic Institute, Troy, New York.*
[3]*Internet Society, Geneva, Switzerland.*

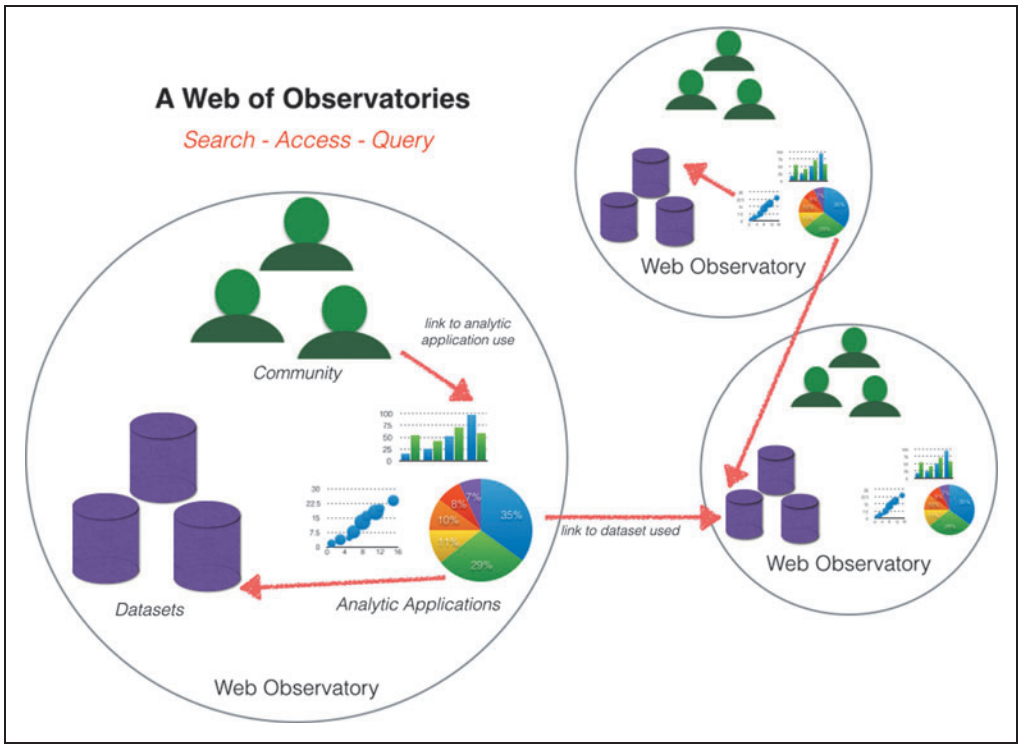

FIG. 1. The web of observatories as the middle layer between enterprise-wide Big Data architectures and the web of data.

or curating the many instances of the same dataset. Certain providers such as Datasift[9] provide access to curated cloud-based versions of datasets; however, those datasets will still have to be copied on local big data installations before analytics can be run on them.

When we consider these issues, it becomes clear that there is a gap that potentially inhibits the vision of big data, and especially broad web data, to reach its full potential; this gap lies in the ability to have distributed infrastructures for big data analytics that will (1) enable the query of common datasets where they are, without the need for replication unless desired; (2) provide opportunities for more people to engage with those datasets' and (3) enable the reuse of analytic tools that can be applied for analysis on different datasets. This gap calls for a different kind of global, more weblike infrastructure that will bridge the potential of sharing datasets on the Web (as evidenced by the growing amount of open data available) and the big data analytics that are used within the enterprise. Creating this middle layer between the web of data and the big data in the enterprise (see Fig. 2) is the goal of the Web Observatory project.

> "FROM A SOCIETAL POINT OF VIEW, THE PROJECT CAN PROMOTE DIGITAL LITERACY AND EMPOWER MEMBERS OF THE PUBLIC TO ACCESS DATA SOURCES ON MATTERS OF SCIENCE, GOVERNMENT, AND THE ECONOMY."

The Web Observatory as the Middle Layer

To bridge this gap, the emerging web of observatories is enabling people to discover and access not just their own data, but also datasets and analytic applications hosted on a web observatory supported by another community, often in a different administrative domain. Researchers are beginning to develop and share resources across connected web observatory instances; those applications can make use of one or more datasets or reuse other analytic applications. In this increasingly capable web of observatories, it is now possible to query or search across datasets that are open or to which one has been granted access (and where the data cannot be shared, to find the critical metadata that is needed to get access). Analytic applications on the web of observatories can point to the data sources they use and descriptions of the methods they employing. From a business point of view, in the web of observatories it will be easier to gain access to datasets without need of replication, to gain access to analytic applications for reuse, and to get a number of experts and public to engage with datasets and analytic applications that a research entity or company wishes to publish

and share. This can bring down costs of big data infrastructure for an enterprise and, by reducing the barriers to entry, enable smaller enterprises to better engage with data-driven innovation. From a societal point of view, the project can promote digital literacy and empower members of the public to access data sources on matters of science, government, and the economy.

In making this web of observatories possible, the following principles are required in the deployment of web observatory instances:

- Resources related to web observatories (projects, datasets, analytic applications, and people) need to have unique identifiers (preferably unique resource identifiers [URIs] or application programming interface [API] access points).
- There needs to be explicit links between analytic applications and the dataset(s) that they use.
- There need to be explicit links between web observatory resources and related use, scholarship, and discourse.
- Metadata should be published for all available resources in a web observatory instance.
- Datasets and analytic applications hosted or listed on a web observatory instance can be public or private; the publisher needs to control who can gain access to them.

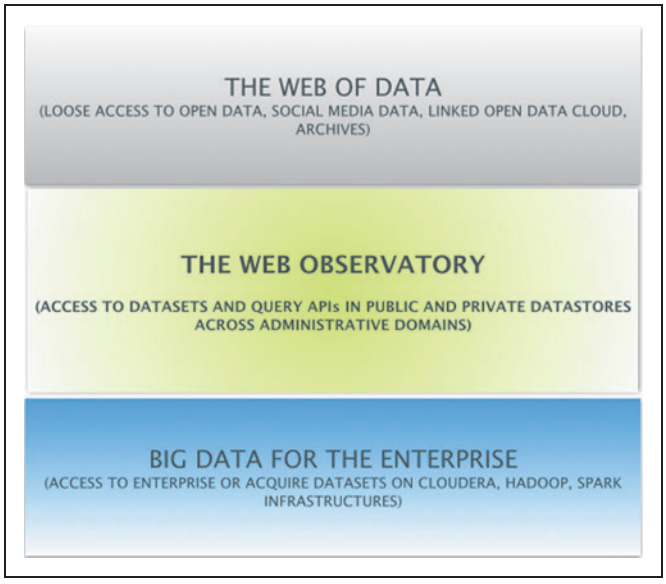

FIG. 2. Web of Observatories: linking components of Web Observatory instances.

- It should be possible to enable access for identified individuals (or to applications using their credentials) to specific datasets or applications hosted in local or remote datasets.
- It should be possible to support distributed queries across web observatory instances and to make computational resources on each instance available to that end.

There are significant technical challenges to achieving this vision. From a computing perspective, these challenges relate to the performance of machine-learning algorithms and statistics when applied on a large volume of data in a remote location. They also relate to providing for distributed queries that will address performance and bandwidth issues when querying more than one remote dataset. There are also challenges to designing the interfaces and APIs that enable the development of reusable analytic applications. Another set of challenges relate to providing ways to describe, discover, and access data in a secure way to these public and/or shared analytic resources. To achieve this, the project depends on providing adequate metadata for datasets on a level of granularity that will inform how they are to be used by applications, and this requires consensus and standards. This context that metadata can capture and deliver needs to be on a level that is neither too simple to be meaningful nor too detailed to be easily created or of value across different kinds of communities of practice. There also should be consensus on which datasets need to be treated as ephemeral; in the context of web observatories, metadata can include information as to which datasets are being utilized by applications and thus enabled more informed decisions.

"IN THE LONG-TERM, THERE IS A NEED FOR SOFTWARE THAT WILL CONSTANTLY MONITOR PRIVACY AND CONFIDENTIALITY ON THE WEB OF OBSERVATORIES, ALTHOUGH FOR NOW THAT IS BEING DONE BY HUMAN EFFORT."

This middle layer of enabling big data analytics across administrative domains will require addressing these challenges in a way that the tradeoff between cost savings and reduced performance due to remote access is meaningful for certain sectors. From a privacy perspective, the individuals who publish or share resources on web observatories need to be aware of issues of privacy and confidentiality; individuals should be warned in advance of possible violations should they add a specific dataset or application to the web of observatories and be able to withdraw such resources if or when such issues arise. In the long-term, there is a need for software that will constantly monitor privacy and confidentiality on the web of observatories, although for now that is being done by human effort.

The deployment of the web of observatories has been based on a bottom-up approach where this tradeoff is becoming clear

and meaningful on each incremental step. The first step of this approach involves the development of web observatory instances as large collections of datasets and analytic tools with which a community of people can engage. Such instances are currently primarily sharing data, but increasingly they are allowing remote access to datasets hosted on different platforms. They also now enable the creation of datasets by communities engaging with data stored on Hadoop installations, cloud platforms, and more recently the Internet Archive[10] and datasets derived therefrom.

As a key step in enabling this vision, an extension to the schema.org vocabulary is being used to describe web observatory instances and their three constituent parts: projects (which represent communities of people), datasets, and analytic applications and tools.[11] Emergent and existing web observatory instances embed this vocabulary on their sites, making resource descriptions accessible to web crawlers and search engines, thus supporting their discovery and use across communities. Progress on distributed query optimization has been made primarily for SPARQL queries, but there is increasing work on optimization techniques for additional types of data stores. Such techniques can focus on optimizing specific types of popular data mining algorithms and datasets (e.g., many Twitter™ collections are proposed for sharing, modulo the Twitter terms of service). Critical components that make use of optimization techniques can be shared across web observatory installations providing for further harmonization and further integration on the global web of observatories maintaining the bottom-up approach. Standardization helps support the reuse of analytic applications on different datasets and the detailed description of datasets and tools to support workflows and composition of analytics on the web of observatories. (A community group for those involved in this effort is supported by the World Wide Web Consortium.)[12]

The next steps in continuing to evolve this emerging distributed set of web observatories focus on the harmonization of access control mechanisms to resources in different domains and on the provision of levels of access control that will be able to distinguish specific types of queries, users, and slices of the data that those users can access. In addition, the development of components to monitor and safeguard privacy and confidentiality will be essential as the volume of resources on the web of observatories increases. One differentiator between these efforts and the earlier eScience Grid[13] is the focus on the use of web-based mechanisms for control and sharing, rather than the point-to-point agreements needed on the Grid because of the specificity and costs of the specialized high-performance architectures supported in that project.

The Web Observatory project continues to strive to bring together diverse communities engaging with data science and web-based resources. As technology progresses, we expect to go beyond the currently available sets of social media resources, linked data, and open government data now supported. We expect that the "Internet of Things" will provide for a dramatic increase in data volume, which, while further justifying the rationale of reducing redundancy with the web of observatories, will add to the challenges of optimization and security. It will also require harmonization of standardization efforts that will involve the Web (W3C), the Internet (IETF), and the e-Science communities. Current discussions to this effect have begun within the Research Data Alliance[14] as well as the previously cited W3C effort. As with many "open web" projects, the eventual success of this project requires the availability of networking and computing infrastructures that will aim to equally empower people regardless of the part of the world in which they live or the circumstances under which they were born. The undertaking of the Web Observatory project is ambitious, but it will be key in fostering data-driven innovation, economy, and data literacy on a worldwide scale.

Acknowledgments

The authors would like to acknowledge the contribution of Tat-Seng Chua (National University of Singapore), Noshir Contractor (Northwestern University), Nigel Shadbolt (University of Southhampton), and Dave De Roure (University of Oxford and leader of the W3C community group) to this work and that of all the WSTnet laboratories[15] and industrial supporters of the Web Observatory activities.

Author Disclosure Statement

No competing financial interests exist.

References

1. http://webscience.org/web-observatory/ (Accessed August 20, 2014).
2. http://webscience.org (Accessed August 20, 2014).
3. Tiropanis T, Hall W, Shadbolt N, et al. The Web Science Observatory. IEEE Intell Syst 2013; 28:100–104.
4. Hendler J. Broad data: Exploring the emerging web of data. Big Data 2013; 1:18–20.
5. Berners-Lee T. Long live the Web: Call for continued open standards and neutrality. Scientific American, December 2010.
6. Hall W, Tiropanis T. Web evolution and Web Science. Comput Netw 2012; 56:3859–3865.
7. Manyika J, Chui M, Brown B, et al. Big Data: The Next Frontier for Innovation, Competition, and Productivity. New York: McKinsey Global Institute, 2011, pp. 1–156.
8. Rothwell R, Zegveld W. Innovation and the Small and Medium Sized Firm. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship, 1982.
9. http://datasift.com (Accessed August 20, 2014).
10. https://archive.org (Accessed August 20, 2014).
11. DiFranzo D, Erickson JS, Gloria MJKT, et al. The web observatory extension: Facilitating web science collaboration

through semantic markup. In Proceedings of www 2014, April 7–11, 2014, Seoul, South Korea, pp. 475–480.
12. www.w3.org/community/webobservatory/ (Accessed August 20, 2014).
13. Foster I, Kesselman C, Tueckle S. The anatomy of the grid. Int J High Perform Comput Appl 2001; 15:200–222.
14. https://rd-alliance.org/ (Accessed August 20, 2014).
15. http://wstweb1.ecs.soton.ac.uk/wstnet-laboratories/about-wstnet/ (Accessed August 20, 2014).

Address correspondence to:

Thanassis Tiropanis
Southampton University
University Road
Southampton SO17 1BJ
United Kingdom

E-mail: tt2@ecs.soton.ac.uk