

The Web Science Observatory

Thanassis Tiropanis¹, Wendy Hall¹, Nigel Shadbolt¹, David De Roure²,
Noshir Contractor³ and Jim Hendler⁴

The World Wide Web is the largest information fabric in history. People shop, date, trade and communicate with one another using it. Scientists and researchers cannot imagine their work without it. The Web is ubiquitous and pervasive, and like all things that become commonplace, we take it for granted. However, over the past few years there has been a growing recognition that the ecosystem that is the Web needs to be treated as an important and coherent area of study—this is Web Science.

Web Science emerged in 2006 [1] as an interdisciplinary area to study the evolution of the Web and understand how it informs, shapes and is shaped by human activity. Bringing together researchers from different disciplines including computer science, the social sciences, law and engineering, effective Web Science methods adopting both quantitative and qualitative approaches are becoming established across the growing research community.

The launch of Web Science was ambitious and timely, with developments since 2006 ushering new challenges and opportunities both in terms of scale and scope. The success of online social networks, the emergence of new genres of social media such as micro blogging, and the explosive growth of data made available on the Web offer an unprecedented opportunity for the conduct of Web Science at *scale* [2,3]. Notwithstanding the computational challenges they create, data on the use of search engines (such as Google analytics) or on topics trending in online services such as YouTube, Twitter [4] and Wikipedia provide an increasingly accurate reflection of, and explanation for, social attitudes and behaviours.

In addition to scale, the conduct of Web Science is presented with challenges and opportunities in terms of *scope*. Even though it was launched in the West, the Web is increasingly a tapestry of multilingual and multicultural complexity. This diversity is revealing the computational limits of traditional search technologies and big data in Web Science research. As a result, they have spurred the development of new methodologies as well as visual-analytical and predictive tools that utilise archival as well as synthetic data. They offer exciting new opportunities for Web Science to

investigate and explain cultural differences and similarities [5].

To keep pace with the Web's growing scale and scope, Web Science research demands the development of new theories, the availability and interpretation of relevant data, effective and scalable multilevel analytical methods, and considerable computational infrastructure. This then is the motivation for the *Web Observatory*: an environment that will enable the next generation of interdisciplinary Web Science research involving mixed methods at a global scale. It seeks to empower researchers by providing a distributed, collaborative, scalable and sustainable online environment to share data, analytical methods and visualisation tools to explore the socio-technical evolution of the Web. In this respect, the Web Observatory project differs from other endeavours by focusing specifically on data *about the Web* (rather than *all* data on the Web), placing significant emphasis on inter-disciplinary analysis and providing multi-level analytics on a global scale.

Observing the Web

The Web Observatory is envisioned as a global data resource and an open analytics environment to nurture Web Science research. In order to understand the factors that have driven the Web's growth, to examine its current condition and anticipate future developments, the Web Observatory aims to provide a distributed archive of data and activity on the Web, as well as methodologies and tools to explore its evolution in the past and through time. By doing so, the Observatory enables:

- Live monitoring of the state of the Web in terms of topologies, resources, links and activity. These include the availability of Web resources (documents and data), relationships among resources and activity in the context of social networks and open data, and state-of-the-art visual-analytics to advance *explanatory* models of the co-evolution of Web and society.
- Simulation of the state of the Web at specific points in the past and development of *predictive* models of Web evolution and of our engagement with or via the Web.

¹ Electronics and Computer Science, University of Southampton, UK; {tt2,wh,nrs}@ecs.soton.ac.uk

² Oxford e-Research Centre, University of Oxford, UK; david.deroure@oerc.ox.ac.uk

³ Northwestern University, USA; nosh@northwestern.edu

⁴ Rensselaer Polytechnic Institute, USA; hendler@cs.rpi.edu

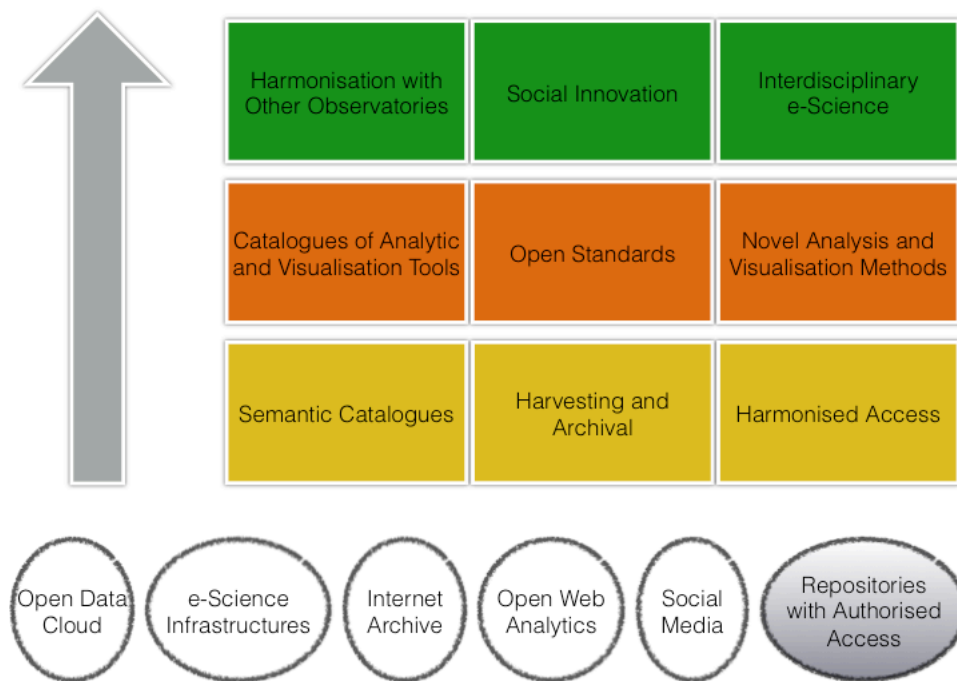


Figure 1. The bottom-up approach of building the Web Observatory.

In order to realise its vision, the Web Observatory is conceived as a multi-stakeholder endeavour. It entails substantial intellectual and resource investments by academia, industry and governments. In return, all of these stakeholders can accrue substantial benefits from the theoretical and practical insights gleaned on topics such as Web economics, social networks, privacy, trust, and Web infrastructure. Given the pressing need for developing Web analytics capability and capacity [6], the Web Observatory will empower government, academia and industries – large and small – with greater access to data, analytical tools and visualisations. It will assist policy makers, legislative bodies and other significant stakeholders to ensure safeguarding of privacy, societal values and the future of the Web. Finally, it will facilitate the emergence of standards and tools for the harmonisation of existing infrastructures that have been deployed. Therefore the foundations of the Web Observatory are:

- Access to distributed repositories of data related to the use of the Web, open data, online social network data and Web archives.
- Harmonised access to distributed repositories of visual-analytical tools, to support a variety of quantitative and qualitative research methods that are interoperable with either published or private datasets.
- Shared methodologies for facilitating the harvesting of additional data sources and the development of novel analytical methods and visualisation tools to explore how the Web informs, shapes and is shaped by human

activity, to address societal challenges and to promote innovation.

- Provision of a forum for discussion about an ethics framework on the archiving and processing of Web data and relevant policies.
- A data licencing framework for archived data and the results of processing those data.

Building the Observatory

In order to be scalable and sustainable, the Web Observatory must be built as a distributed environment for collaborative knowledge sharing rather than a vast warehouse. Realising this vision of a Web Observatory for Web Science requires a bottom-up approach involving a sequence of activity streams (Figure 1). The first activity stream involves identifying existing repositories and archives that contain relevant data; these come from research laboratories and other global efforts such as open data repositories, e-Science infrastructures, the Web Foundation and the Internet Archive. This activity involves the publication of Semantic Catalogues to locate and describe existing datasets that are available for use. Identified datasets come from businesses and organisations, including university labs; those parties may wish to share their datasets or advertise them and only grant access to authorised parties. Provenance [7] will be essential to track the generation of data products and usage rights, and the Observatory has a keen eye to emerging practice in topics such as automated policy handling, non-consumptive access, anonymisation, privacy-preserving linkage and differential privacy.



Figure 2. The NUS Tsinghua NEXT Observatory.

The second activity stream in building the Web Observatory involves the identification and sharing of tools to visualise, analyse and harvest large distributed datasets. The goal of this activity is to make the capabilities offered by these tools easily accessible to the larger Web Science community. The added value of being able to use these tools on the larger corpus of data as the Web Observatory grows, further establishes its value to industry, government and academia.

These two streams of activity are being coordinated by the Web Science Trust, a global not-for-profit organisation that promotes Web Science and leverages Web Science research resources of the 15 University Labs affiliated with the Trust (WSTNet) from Brazil, China, Korea, Europe and the US¹. Current work includes the deployment of harvesters to collect data on the use of Web resources, schemas to describe datasets, deployment of data repositories based on existing platforms such as EPrints and qualitative and quantitative methods to establish trends and influences across cultures².

Beyond WSTNet, a number of efforts are underway in different parts of the world, providing online analysis and visualisation of social, business, government and research activity; one example is the NEX-Live Observatory by the National University of Singapore and Tsinghua University [8] (Figure 2). The work within the Web Observatory is set to help establish the necessary momentum, standards and infrastructure to enable a Web of observatories with the potential for global impact, which leads to the third activity stream. To this end the Web Observatory is forging partnerships with the W3C, the Open Data Institute (ODI) in the UK, Fraunhofer, the Web Foundation and a growing list of industry collaborators.

A W3C community group³ has been established to foster discussion on the standardisation that will be necessary to enable interoperability between available resources and on identifying the opportunities for industry and global government agencies to contribute large scale systems, expertise and datasets. This enables the Observatory to operate not only as a 'lens' into

¹ <http://webscience.org/WSTNet.html>

² <http://thewebobservatory.org>

³ <http://www.w3.org/community/webobservatory>

human activity but also as a decentralised, distributed infrastructure for sharing data and analysis: a Web of observatories that will be more than the sum of its parts, opening new ways for conducting research and promoting innovation.

Using a bottom-up approach to building the Web Observatory invites two sets of challenges in the CS and Intelligent Systems arena. *First*, it necessitates the identification of open standards and protocols to promote interoperability for gathering and sharing data. With this effort under way, it will be possible to deploy new visual-analytical tools that will support mixed methods and interdisciplinary Web Science. Data analysis pipelines that can be automated stand to assist the scientist by mitigating the drudgery of managing the increasing volume of Web information flows. Hence building the Web Observatory will necessitate effective research communication to support reproducible and reusable digital methods [9]. *Second*, building the analytic tooling in the Web Observatory brings a set of challenges in the areas of databases, provenance and query languages for semi-structured data, record linkage, ontologies and inference, computational algorithms, high performance computing, computational workflows, machine learning, knowledge representation, policy reasoning and systems architecture.

The impact of the Observatory

Building the Web Observatory has the potential for transformative societal impact. But the relevance of this effort to multiple stakeholders – academia, government and industry – and their sustained engagement in a partnership, is key to the Web Observatory's success in helping address societal challenges. Benefits to industry and academia will encourage them to contribute additional data and tools and thus promote network effects that will help make the Web Observatory a vibrant global resource. The engagement of government will ensure that public sector information is made available in a structured and harmonised way, allowing governments to leverage industrial and academic activity at scale. A collateral benefit is that governments, industry, and academia will be able to use the Observatory to crowd-source analytical talent globally, and deposited data stands to accumulate in value when it is utilised.

By undertaking the challenge of harmonising data and analysis infrastructures and engaging in standardisation, the Web Observatory is paving the way for further integration of structured information on the Web. Building on open data initiatives, the Observatory will provide the necessary tools for publishing structured information and its integration with public and private resources. This standardisation effort, coupled with engagement

with communities in data repositories, e-Science and big data analytics, will provide a framework for the harmonisation of e-Infrastructures that go beyond Web Science. As such, even though the Web Observatory is focused on data *about* the Web, its standardisation efforts will enable the development of observatories about *all* data on the Web.

To understand and enable the evolution of Web to help address grand societal challenges, we need to observe the Web at scale across space and time, and that requires a globally distributed and collaborative Web Observatory.

References

1. Berners-Lee, Tim; Hall, Wendy; Hendler, James; Shadbolt, Nigel and Weitzner, Danny (2006). *Creating a Science of the Web*. *Science*, 31 (5788), 769-771.
2. Lazer, David; Pentland, Alex; Adamic, Lada; Aral, Sinan; Barabási, Albert-László; Brewer, Devon; Christakis, Nicholas; Contractor, Noshir; Fowler, James; Guttman, Myron; Jebara, Tony; King, Gary; Macy, Michael; Roy, Deb and Van Alstyne, Marshall (2009). Computational social science. *Science*, 323 (5915), 721–723. doi:10.1126/science.1167742
3. Aral, Sinan and Walker, Dylan (2012) Identifying Influential and Susceptible Members of Social Networks. *Science*, 337 (6092), 337-341.
4. Mustafaraj, Eni and Metaxas, Panagiotis (2010). From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search. In Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, April 2010.
5. Scott Golder and Michael W. Macy (2011). Diurnal and Seasonal Mood Vary with Work, Sleep and Daylength Across Diverse Cultures. *Science*, 333 (5061), 1878-1881.
6. Manyika, James; Chui, Michael; Brown, Brad; Bughin, Jacques; Dobbs, Richard; Roxburgh, Charles and Hung-Byers, Angela (2011) Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute
7. Groth, Paul and Moreau, Luc (editors) (2012). PROV-Overview: An Overview of the PROV Family of Documents, W3C Working Draft 11 December 2012
8. Chua, Tat-Seng; Luan, Huanbo; Sun, Maosong; Yang, Shiqiang; , "NExT: NUS-Tsinghua Center for Extreme Search of User-Generated Content," *MultiMedia, IEEE* , vol.19, no.3, pp.81-87, July-Sept. 2012 doi: 10.1109/MMUL.2012.39
9. FORCE11 (2012) "Force11 Manifesto", http://www.FORCE11.org/white_paper