Some Challenges for the Web Observatory Vision: Field Notes from a Southampton-Tsinghua-KAIST Collaboration

Evangelia Papadaki, Abby Whitmarsh, Eamonn Walls
Web Science Doctoral Training Centre
University of Southampton, UK
ep11g12@soton.ac.uk, aw3g09@soton.ac.uk, ew1g12@soton.ac.uk

ABSTRACT
This paper outlines some challenges for the Web Observatory vision with reference to field notes from a student exchange and research collaboration in December 2013 between the University of Southampton, Tsinghua University and KAIST. These field notes outline a methodological narrative of the practical challenges that we faced in using the Web Observatory in collaborative research. It is suggested that these challenges particularly come in the form of technical, organizational and legal issues. The paper concludes with some proposals for the future of the Web Observatory vision.

Categories and Subject Descriptors
K.4.2 [Computers and Society]: Social Issues

Keywords
Web Observatory; Field Notes; Sina Weibo; Corruption

1. INTRODUCTION
The Web Observatory started under the WSTn (Web Science Trust network) and builds on open data initiatives [3, 10]. The objective is to build in bottom-up fashion a distributed environment facilitating greater access to datasets and interoperable analytic and visualisation tools [10]. The Web Observatory is still an idea in progress – there are few strictly agreed definitions or standards, and processes to move towards such standards are still ongoing. The Web Observatory has been thought of as part of the vision of the evolution of the Web in general and Web Science in particular [10]. However there are a number of difficult and ongoing questions surrounding big data management that may have an influence on the future development of the Web Observatory [2]. There have been calls in the literature for wider discussion in the Web Observatory community to begin to define relevant criteria by which data might be assessed and improved over time [1, 5, 8].

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.
ACM 978-1-4503-2622-3/14/06.
http://dx.doi.org/10.1145/2615569.2615646

2. FIELD NOTES FROM A STUDY USING THE WEB OBSERVATORY
This paper refers to field notes of a study using the Web Observatory 'in the wild'. This study was part of a collaboration and student exchange between the University of Southampton, KAIST and Tsinghua University in December 2013. The aims and objectives of the project were chosen with guidance from professors at the University of Southampton. Using datasets from various Web Observatories and hosted at the University of Southampton Web Observatory, the project aimed to discover how political corruption was discussed and reported on social networking sites (SNS) in China, such as Sina Weibo.

A dataset from Sina Weibo which had been harvested by researchers at Tsinghua University had been made available to members of our research group. After examining the data it became apparent that the dataset we were using had been filtered to only two topics, neither of which were related to corruption. This was the first challenge to the project - not knowing what data was available in a Web Observatory nor the provenance of the data. For the research week at the University of Southampton our research group instead collated manual figures from Twitter and Sina Weibo. A list of official hashtags which were used on Sina Weibo to report corruption was supplied to University of Southampton academics. Using these hashtags we were able to produce some visualisations showing the frequency of use of the different hashtags by month.

We received access to a different dataset while taking part in the research week at Tsinghua University. However the provenance of the dataset was not discovered until after the research week was over. This appears to be a recurring issue for Web Observatories in general. Once we had access to some meaningful data we began the process of querying the database, and came upon our next challenge. The dataset is called weibo_2012 and is hosted at mdb-001.ecs.soton.ac.uk in a mongoDB database. Our team had no experience with using mongoDB and had to learn quickly how to query, view and group data to make it meaningful. This may suggest that researchers who wish to use data hosted at Web Observatories in its current form will require a relatively high level of technical skills. Querying the dataset we found the number of weibos in the dataset was 5,279,579. Such a large dataset was slow to query and our original plan of creating a visualisation which could be queried dynamically had to be abandoned and we decided to create visualisations on data that we had already filtered. Many of these decisions were influenced by purely practical considerations – limitations in computing resources, time, and labour all strongly influenced the direction of the project. Using a python script we created a smaller dataset based on ten search terms with and without hashtags.

3. CHALLENGES FOR THE WEB OBSERVATORY VISION

3.1 Technical Challenges

In order for the Web Observatory to operate as a common infrastructure for sharing data, effective research communication is required for the identification of interworking standards, which will allow datasets to be used effectively across the Web Observatory community [7]. A key concept for the design of a globally distributed Web Observatory is 'interoperability' [3]; querying across different platforms and datasets can be achieved only by addressing the challenges of data harmonization, standardization and the development of shared methodologies for facilitating data harvesting. What is more, given that the Web Observatory needs to be collaborative, it is recommended that the activities of Web Observatory researchers adhere to best practices of technology support for scientific method deriving from previous collaborative activities [8]. Issues of reusability could be addressed by documented and principled methods based on best practices such as Linked Data technologies [1].

3.2 Organizational Challenges

Given that the quality of the data is closely intertwined with the purposes of each funding project that creates it, questions arise as to whether this data could meet the needs of projects that want to reuse them [4]. Even if this first hurdle can be surmounted, accessing the datasets raises major privacy concerns imposing significant burdens on the implementation of the Web Observatory vision. Depending on the terms of service of the data source many datasets may not be entirely open; such datasets may require anonymization at source, data security, restricted access or even legal approval in order to be collected and shared [9]. Therefore, a variety of competing Internet players need to undertake the responsibility of ensuring that access to data sources is ethically controlled; it is essential that ethics processes are placed into the heart of Web Observatory governance structures instead of being managed around the edges [4, 9].

3.3 Legal Challenges

Under the terms of the Copyright, Designs and Patents Act 1998 (CDPA), databases are treated as a class of literary work and may therefore receive copyright protection; copyright protection afforded to a database as a whole should be distinguished from any protection that its individual components may attract. Copyright law aims at rewarding the author's intellectual creativity by protecting his work against copying without licence or permission; that is why, protection is limited to databases containing a sufficient degree of creativity in the selection and/or the arrangement of the data. To sum up, a body considering sharing data should consider whether the databases potentially qualify for copyright/database right protection, who is the owner of the databases, whether additional contractual protection is required with the party data is being shared, whether there are any licences to use the databases, whether it is complying with its obligations under data protection legislation etc. The borderless nature of cyberspace renders the sharing of online data even more complicated given that legal outcomes may differ between countries and thus the abovementioned rights may be enforceable in one country but not another.

4. CONCLUSION

Our experience of using the Web Observatory 'in the wild' suggests that many challenges lie ahead for the Web Observatory vision. The Web Observatory must demonstrate an ability to provide value to its stakeholders if it is to succeed in facilitating data exchange and research [1]. This point applies both on the individual and on the organizational level. "Whatever metric is used to value return on investment it need not be a financial value in itself but must be translatable into one. Otherwise, the funding of exchange systems like Observatories will remain fundamentally detached from the value created in the exchange" [1]. If the goal of Web observatories is to encourage and facilitate the sharing of data, it is essential that there is some incentive for data to be shared. It is imperative that the value of data exchanges can be framed in terms of measurable return on investment [1].

5. ACKNOWLEDGMENTS

This study was part of an ongoing collaboration between the University of Tsinghua, People's Republic of China, KAIST Advanced Institute of Science and Technology, Republic of Korea and the University of Southampton, UK.

6. REFERENCES

[1] P. Booth, P. Gaskell, and C. Hughes, "The Economics of Data: Quality, Value & Exchange in Web Observatories," in *Proceedings of the 22Nd International Conference on World Wide Web Companion*, Republic and Canton of Geneva, Switzerland, 2013, pp. 1309–1316.

[2] V. Borkar, M. J. Carey, and C. Li, "Inside 'Big Data Management': Ogres, Onions, or Parfaits?," in *Proceedings of the 15th International Conference on Extending Database Technology*, New York, NY, USA, 2012, pp. 3–14.

[3] I. Brown, W. Hall, and L. Harris, "From Search to Observation," in *Proceedings of the 22Nd International Conference on World Wide Web Companion*, Republic and Canton of Geneva, Switzerland, 2013, pp. 1317–1320.

[4] T. Davies, "Web Observatories: The Governance Dimensions", *Open Data Impacts Blog*, October 9, 2013, http://www.opendataimpacts.net/2013/10/web-observatories-the-governance-dimensions/.

[5] D. De Roure, C. Hooper, M. Meredith-Lobay, K. Page, S. Tarte, D. Cruickshank, and C. De Roure, "Observing Social Machines Part 1: What to Observe?," in *Proceedings of the 22Nd International Conference on World Wide Web Companion*, Republic and Canton of Geneva, Switzerland, 2013, pp. 901–904.

[6] E. Diaz-Aviles, "Living Analytics Methods for the Web Observatory," in *Proceedings of the 22Nd International Conference on World Wide Web Companion*, Republic and Canton of Geneva, Switzerland, 2013, pp. 1321–1324.

[7] C. Gallen, "Some Considerations for a Web Observatory", in *1st International workshop on Building Web Observatories, ACM Web Science,* May 2013.

[8] H. Glaser, "Observing Observatories: Web Observatories should use Linked Data", in *1st International workshop on Building Web Observatories, ACM Web Science,* May 2013.

[9] M. J. K. Gloria, D. L. McGuinness, J. S. Luciano, and Q. Zhang, "Exploration in Web Science: Instruments for Web Observatories," in *Proceedings of the 22Nd International Conference on World Wide Web Companion*, Republic and Canton of Geneva, Switzerland, 2013, pp. 1325–1328.

[10] W. Hall and T. Tiropanis, "Web evolution and Web Science," *Computer Networks*, vol. 56, no. 18, pp. 3859–3865, Dec. 2012.