From Search to Observation

Ian Brown
Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, UK
+44 (0)23 8059 5000
icb1g12@soton.ac.uk

Wendy Hall
Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, UK
+44 (0)23 8059 5000
wh@soton.ac.uk

Lisa Harris
School of Management
University of Southampton
Southampton, SO17 1BJ, UK
+44 (0)23 8059 5000
l.j.harris@soton.ac.uk

ABSTRACT
In this paper, we propose a set of concepts underlying the process and requirements of *observation:* that is, the process of employing web observatories for research. We refer to observation as a new concept, distinct from search, which we believe is worthy of study in its own right and note that the process of observation moves the focus of information retrieval away from universal coverage and towards improved quality of results and thus has many potential facets not necessarily present in traditional search.

Categories and Subject Descriptors
H.5.3 [**Group and Organisation interfaces**]:

General Terms
Design, Human Factors, Standardization, Theory.

Keywords
Web Science, Web Observatory, Observatory models.

1. INTRODUCTION
The concepts and purpose of Internet search have become established in the minds of users through such leading providers as Google, Yahoo, Alta Vista, Bing and many others. Search was developed as a necessary and specific response to the increasing failure of memorable URL namespaces to map easily or uniquely to the sites/documents that users were actually seeking. Whilst it was, for example, initially easy and practical to search for *widgets* at www.widgets.com (or suitable variants), the wholesale registration of domain names by start-ups and internet opportunists quickly exhausted this mapping opportunity leaving users with the need for a tool beyond a database of bookmarks to find new information and services on new sites.

The search function requires an indiscriminate trawl of a "sea of documents" leading to the creation of huge generic indices delivering results where the vast majority of the matching references/links are never viewed and hence are effectively wasted. The nature of search may, arguably, be characterized as an individual transaction where relatively vague queries are submitted without an explicitly stated context or purpose and which are addressed by highly generalized cataloging methods delivering answers, which (through lack of a known context) are not aligned or structured according to the user's intention.

Search engines such as Google have refined this brute force method through processes of inference and analysis based on users' previous choices in order to refine (filter) the results presented in the hope that previous choices will deliver good results for current desires. This, however, arguably creates potential problems such as "filter bubbles" [1] in which new results are less likely to be returned vs. previously selected results – something not aligned with the desire to search for new knowledge in a research context.

As the data deluge worsens it will become increasingly challenging to control the search process to find relevant, good quality research data. To provide tools and approaches that support good quality research, the development of the Web Observatory has been proposed [2]. This highlights the need for new processes to address a problem not solved by existing search technologies: namely that of discovering and assembling well structured and curated results from the *web of data* for the purpose of deriving insights into Web Science research questions.

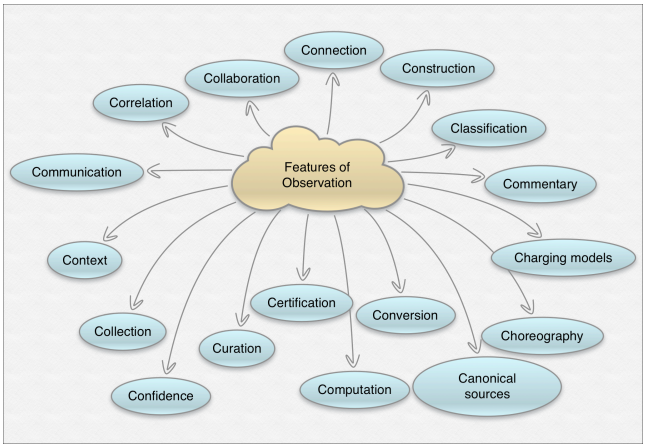
2. THE NATURE OF OBSERVATION
Over recent years a number of repositories (chiefly from academic institutions) have started to emerge as nascent observatories, which seek to implement one or more aspects of the observatory problem space. These early implementations are individual to each institution and in order to encourage interoperability and, ultimately, standardisation we present here a range of concepts, some or all of which may be present in an observation process. Whilst it not intended to suggest that all observatories must exhibit all the following features nor that the presence of *any* of these features automatically confers the status of an observatory; it could be however be argued that a system that had *none* of these features would be hard to define as an observatory.

The following may not be an exhaustive list but is intended to stimulate discussion and inform potential harmonisation in the growing area of defining and studying web observatories. The development of observation processes will likely be incremental over time with a sub-set of core features coming first followed by other features ranking in terms of importance. No arbitrary full ranking is offered here, as specific projects/observatories are likely to have different internal priorities though the development of a *minimum operating set* is likely to be a next useful step.

WWW 2013 Companion, May 13–17, 2013, Rio de Janiero, Brazil.
ACM 978-1-4503-2038-2/13/05.

In the following section we discuss processes and capabilities that we would expect to see in Observation vs. a search interaction. Italics are used to indicate a key issues for Observation, which may be required for interoperability/standardization.

2.1 FEATURES OF OBSERVATION



Observations may involve the exchange of data between two or more *collaborating* parties for the achievement of common or complementary goals. This exchange may not involve a charging structure but may nonetheless require *formal agreements and terms*. It is not anticipated that all data that will be observed will necessarily be open data and hence provided free of charge. It is anticipated that observations may involve a commercial *charging model* incl. the payment of a license fee with a mechanism to grant the permissions associated with the license vs. those without a license or with a different license. Where the observer's process requires confirmation of the source of data to be explicitly documented, a certificate format and *description of permitted use of the data* may form part of the certification. Observers may wish to base sensitive calculations/decisions on observed data and hence *confidence* in terms of *trust and provenance* will be required particularly for automated/unattended processes.

In contrast to a single request/response from a known search engine, the process of observation may be characterized as one or more *communication processes across several repositories* starting with discovery of sources, the disclosure of metadata, the negotiating/establishment of technical data exchange and the grant (either manual/technical) of licenses.

Each series of observations will typically be made in the *context of a research question* and specific linked research papers, tools and other materials, which inform the relevant curation, commentary and collaboration (see below) addressing the research question.

Once a source is identified from a repository as part of an observation service it would typically offer a *formal classification according to topics* using some knowledge classification schema.

Observers will typically need to access the raw data and linked materials from one or more repositories, which their request/search has identified – potentially using protocols/methods distinct from the query protocol itself and thus the *separate method of connection beyond the query* needs to be addressed.

Observation will often be association with longitudinal datasets from one or more sources and whilst it is not envisaged that all observatories will seek to store all data, is it anticipated that each observatory would store some data and hence a process of regular *collection*, snapshotting or processing of streaming data would be required. Given each repository may hold datasets in a variety of formats - metadata associated with the dataset will allow the observer to invoke appropriate *validation and format conversion* services.

Where more than one repository is accessed offering the same or overlapping datasets there will be the requirement to establish a de facto or *canonical source and de-duplicate* if required.

Since datasets addressing specific research questions may typically be constructed from more than one homogenous data source or heterogeneous structures for allowing for richer analysis of trends and correlations thus we must allow for data to be *complex and constructed from multiple sources*. This is analogous to the concept of variety in Big Data systems though is perhaps more correctly described as “broad data” [3] in Web Science. A composite data set comprising heterogeneous data will allow for the possibility of *correlation analysis* across disjoint topics. Indeed the data set (or sets) may form part of a larger suite of associated tools, analytics or visualizations requiring a series of one or more *computational processes*.

For data sets, which need to be refreshed, or multiple streaming services there will be the requirement to *choreograph* the timing of updates and staging of the data requiring orchestration and synchronization. Over time, Datasets, which may be generated/harvested automatically, may require various levels of on-going maintenance ranging from automated housekeeping up to full *curation processes* of selection, deletion, annotation and re-classification.

It is anticipated that meta-data, including *commentary*, by both users and curators of the data, will provide a richer environment for a qualitative understanding of data beyond stored item values.

3. CONCLUSION

What is striking is the potential richness and complexity of the fully-formed Observatory model due to multiple areas of complexity: the distributed nature of the query/discover process, the need to support disparate formats, operating models and to support complex distributed orchestration.

It is likely that observatories created for different purposes will therefore tend to evolve differently, developing the various capabilities to a greater or lesser degree based on their intended purposes.

The practice of observation in a Web Observatory context will evolve in complexity/capability over time [4] as repositories establish specific standards, linkages and interoperability methods.

We would propose that the establishment of an implementation scale showing the maturity/completeness of the offered feature set of an observatory along with a *minimum operating set of features* would be a helpful measure to establish over time as this will allow users to distinguish between types of services and potentially inform the development and refinement of existing observatories over their operating life-times.

4. ACKNOWLEDGMENTS

This work is supported under SOCIAM: The Theory and Practice of Social Machines. The SOCIAM Project is funded by the UK Engineering and Physical Sciences Research Council (EPSRC)

under grant number EP/J017728/1 and comprises the Universities of Southampton, Oxford and Edinburgh and also the EPSRC grant number EP/K503150/1.

5. REFERENCES

[1] V. Maccatrozzo, “Burst the filter bubble: using semantic web to enable serendipity,” presented at the ISWC'12: Proceedings of the 11th international conference on The Semantic Web, 2012, vol. Part II , Volume Part II.

[2] W. Hall and T. Tiropanis, “Web Evolution and Web Science,” *Computer Networks*, 2012.

[3] J. Hendler, “Increasing access to the web of ‘broad data’,” presented at the W4A '12: Proceedings of the International Cross-Disciplinary Conference on Web Accessibility, 2012.

[4] T. Tiropanis, W. Hall, D. C. DeRoure, N. Shadbolt, N. Contractor, and J. Hendler, “The Web Science Observatory,” pp. 1–4, Jan. 2013.