The Southampton University Web Observatory

Wendy Hall, Thanassis Tiropanis, Ramine Tinati,
Paul Booth, Paul Gaskell, Jonathon Hare, Leslie Carr

Electronics and Computer Science, University of Southampton, England
{wh, tt2, rt506, pmb1g11, pvg1g10, jsh2, lac}@ecs.soton.ac.uk

Keywords: Web Observatory, Web Science, Big Data, Analytics, Visualisation

1 INTRODUCTION

The Web Observatory is a project that started under the auspices of the Web Science Trust network of labs (WSTnet) to support activities on Web data collection and analysis across research institutions around the world. The objective is to build from a bottom-up fashion a distributed environment empowers Web scientists not only with access to datasets about activity on the Web but also to interoperable analytic and visualisation tools that can be combined to allow the exploration of the various aspects of activity on the Web. The harmonization of existing infrastructures is the first step to vision, which is supported by a W3C community group that discusses standardization aspects for the interoperability of Web Observatories.

In this paper we present the Web Observatory that is currently deployed at the University of Southampton. We outline its infrastructure, datasets and tools. The aim is to provide a summary of the functional aspects of the Southampton University Web Observatory (SUWO). Those aspects cover the following aspects of research about the Web:

- The use of the Web of documents.
- The use of Web search engines.
- The use of the Web of data.
- The use of social networks on the Web.
- The use of social machines on the Web.

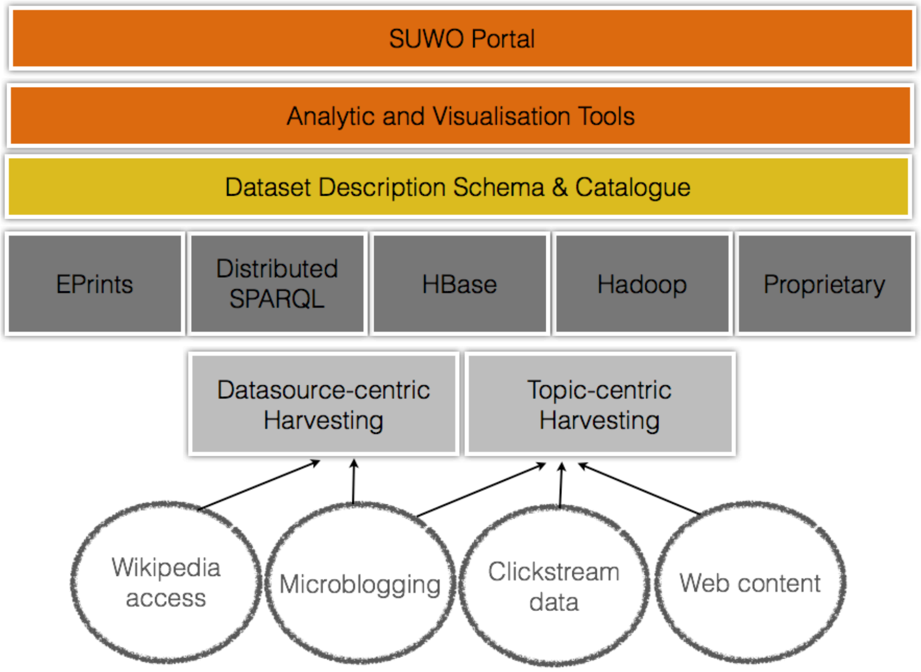
The following sections present the architecture of the infrastructure and the datasets and tools that SUWO currently incorporates.

2 THE DATA INFRASTRUCTURE

The infrastructure of SUWO includes components that support harvesting, storage and access to datasets that enable the study of the different areas of activity on the Web. Ideally, one should be able to take regular snapshots and store all datasets and content on the Web but this is a very challenging endeavor given the volume of content and data produced on the Web. To cope with this challenge, we provide for two types of harvesting activity: (i) data-source-centric harvesting, where chosen data sources are harvested or sampled for the long term, and (ii) topic-centric harvesting and archiving of data related to a specific topic from potentially different data sources. Web Scientists are able to access the repositories of datasets of the first category to study activities in the past on a variety of topics. At the same time, they are able to access topic-specific data and content harvested in the past or set up harvesters for activities that are emerging or planned in the future.

We have been evaluating a selection of different storage and formatting techniques to store and format datasets, including flat file storage, database driven solutions, and platforms such as Apache Hadoop and HBase to store and index larger datasets. Aware that this needs to be a scalable long-term solution, we are deploying an architecture that uses a mix of different storage techniques and we are working on schemas for the uniform description of Web Observatory datasets. With respect to interchange and long-term preservation of harvested content we are exploring the use of existing standards such as the ISO28500 Web ARChive (WARC) format.

Fig. 1. Harvesting, storing and accessing data in SUWO



The data that are currently harvested as part of the data-source-centric harvesting activity currently include Wikipedia access data and micro-blogging posts (Twitter). Data harvested as part of the topic-centric harvesting activity include clickstream data and Web content archives for resources on specific topics for a specific period of time. The diagram of Figure 1 outlines the components of the data harvesting infrastructure and the platforms involved, including Hadoop, HBase, SPARQL endpoints with support for distributed queries, and EPrints for storing datasets in other formats. Those platforms are currently deployed on Cloud infrastructure and on local servers.

The diversity of Web Science research activities combined with the diversity of datasources on the Web require a sophisticated infrastructure for data harvesting from different API's, RESTful interfaces, Web crawlers and raw data dumps. SUWO aims to cater for such a variety of technical differences, but ultimately offer data in standardised Open formats that can be interoperable across other Web Observatories. In addition, the potential to correlate multiple datasets is leading us to the use of properties or identifiers to link the different datasets based on time, providing a way to perform event detection and analysis across datasets. This identifier is to be part of the dataset description schema. To that end, we have proposed the *Web Observatory Time identifier*, which effectively is a shared protocol across all Web Observatories to allow cross comparison of datasets. When a dataset is created, a Web Observatory timestamp is generated, effectively providing an index, independent of data source or format.

3 ANALYTIC AND VISUALIZATION TOOLS

On top of the data harvesting and storage layer, we are building a number of analytic and visualisation tools that will be available via a dedicated portal to the Southampton University Web Observatory. The development of these tools is presenting a number of challenges. Using visual representations of abstract data to amplify cognition can be understood as the science of perception, but also semiotically as 'visual language'. Working with map visualizations for instance, information is communicated visually by encoding data variables (such as latitude and longitude) to visual properties (such as position or colour). The application of visual properties to data values has provided a range of standardised visual structures, such as scatter plots and bar charts that are commonly used and understood across many disciplines. SUWO has taken a multi-disciplinary approach to visualising data by creating working partnerships across statistics, economics, design and computer science. Such integration has uncovered wider insights, which is evident from the collaboration between Switch Concepts, an RTB ad-serving company, and the University. Working closely with Switch from the point of harvesting and refining raw data and up to the point of final graphical presentation has laid sufficient groundwork for new methodologies to develop. These take into consideration both the structure and function of visualisations in an effort to move beyond just the visual-encoding model towards a more holistic, more gestalt visualisation theory; something different to the sum of its parts.

Different technologies have been used with iterative and exploratory methods as a means to finding the most appropriate tool for a given situation; including: D3.js#, an increasingly popular Javascript library, Tableau, TileMill and Adobe Illustrator. Those stages involve the refinement of data for exploration followed by a second stage of visual refinement as illustrated in Figure 2 and in Figure 3. Technology-wise, Python coding allowed efficient aggregation and refining of the data, providing the initial visualizations using D3.js and Tableau. The latter provides rapid visualisation prototyping when working across maps, charts and graphs.

Fig. 2. Data Refinement. The focus begins on selecting, slicing, and filtering by applying statistical methods with prototype visualizations to establish useable data and data structures

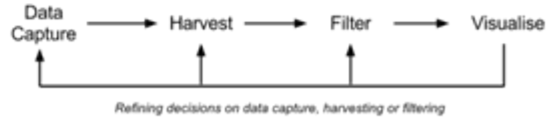
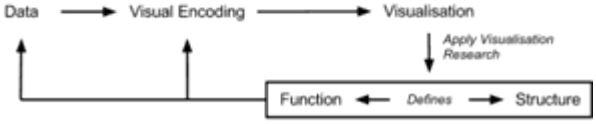


Fig. 3. Visual Refinement. Using finalized data the focus moves to encoding data properties into visual elements, applying existing research to balance the structure and function



Visualizations are commonplace within mainstream media websites and have become a feature in their own right as interactive news supplements. The New York Times and The Guardian Datablog regularly present self-contained applications that allow web users to interact with data. The structure and function of the visualisation, along with the interactive elements, determines it to be one of exploration, explanation, or presentation. Although it is possible to convey a great deal of information in a single visualisation, presenting a narrative as part of a visualised story may provide a more lasting impact.

4 FUTURE WORK

One of the first priorities for SUWO has been the deployment of the harvesting infrastructure and of the infrastructure for storage and access to large datasets in addition to the deployment of initial analytic and visualization tools. We are currently examining ways in which we can introduce a layer that will allow uniform description and discovery of datasets within SUWO or across Web Observatories in WSTnet. This involves the definition of schemas to describe those resources and portals where those descriptions will be available. Distributed queries and performance optimisation for storage and retrieval are significant challenges that will also be addressed.

The data harvested from SUWO provides a unique level of data coupled with unprecedented social context that, with appropriate questioning applied, will facilitate a new level of analytics and visualisation tools based on narratives.

Finally, we are working on the deployment of all available and emerging visualisation tools on a portal that will foster an ecosystem for analytics for Web Science in SUWO and across academic and research institutions around the world.