Developing e-Research Tools for the Analysis of Large-Scale Web Crawl Data

Robert Ackland[1], Joseph Antony[2]

[1] Australian Demographic and Social Research Institute, The Australian National University

[2] The Australian National University Supercomputer Facility

Email address of corresponding author: robert.ackland@anu.edu.au

Abstract. In this paper we describe the development of e-Research tools enabling remote access and analysis of large-scale web crawl data. The tools are being developed in the context of a planned research project titled the ".au Census", the aim of which is to gain new insights into Australian commerce and society using data from large-scale crawls of the Australian public web.

Introduction

This paper describes progress towards the development of a set of e-Research tools to facilitate new research into large-scale web crawl data. In accordance with the theme of this conference, the focus of our paper is on how large-scale web crawl data can be used to advance social science research into online networks, however we note that (of course) the crawl data presented here are of research interest in other domains (and indeed, the crawl has been collected as part of research into information retrieval).

In this paper, we hope to demonstrate how e-Research technologies can be used to facilitate collaboration between researchers from different disciplines. In our example, we have large-scale web crawl data that is being generated as part of information retrieval research at the Commonwealth Scientific and Industrial Research Organisation which a social scientist wishes to use for research. The solution that is explored is the hosting of the data in a high-performance computing environment provided by the Australian Partnership for Advanced Computing National Facility, making the data accessible for research via web services developed as part of the the Virtual Observatory for the Study of Online Networks (VOSON) e-Research tool.

The structure of the paper is as follows. In Section 1, there are some general comments about social science web research using large-scale crawls. In Section 2, we provide an overview of the VOSON e-Research tool and present an outline of the computing setup that has been established in order to facilitate analysis of large-scale crawls via VOSON. In Section 3, a more complete description of the methods and tools that have been developed is presented via a summary of the steps involved in creating a web network dataset using a 1% random sample of the large-scale crawl. Preliminary analysis of the web network data is also presented. Finally, there are some brief observations on the role of peer-produced tools and data in the advancement of social science research into online networks.

Social science research using large-scale web crawls

The emergence of the World Wide Web (WWW) has had a major impact on the way people engage in commerce, participate in the political process and socially interact. Despite the marked influence of the web on many aspects of the lives of people living in industrialized countries such as Australia, there has been very little empirical research using large-scale web crawl data into commerce and society. There are at least three major constraints that researchers from the social sciences face when contemplating using large-scale web crawl data for their research.

The first challenge relates to the applicability of web data for testing existing social science theories. While social scientists theorize on the emergence of a "network society" (e.g. Castells, 1996), and Social Network Analysis (SNA) provides a vast array of analytical approaches for analyzing networks arising from socially-generated processes (e.g. Wasserman and Faust, 1994), it is not immediately clear how such research can be operationalized with web data. To what extent can websites be seen to represent the actions of social actors? What is the exact meaning of a hyperlink? These obviously important theoretical questions are outside the purview of the present paper.

A second challenge relates to the potentially huge scale of web datasets. In computer science, the response to this challenge has been to focus on the development of automatic methods (e.g. statistical machine learning) for classifying/categorizing web pages on the basis of page content and hyperlink structure. Computer scientists also hold much hope for the Semantic Web (SW) which is an effort to build into Web pages tags or markers for data and semantic representations of the meaning of those tags (Berners-Lee, Hendler and Lassila, 2001; Shadbolt, Hall and Berners-Lee, 2006). The promise of the SW is that it will enable websites to be automatically classified, using the tags placed into the website by the site developers. However, as noted by Brent and Carnahan (2007), while the SW holds some promise for empirical social science research, there are two major problems. First, there is the issue of how existing websites can be retrofitted for the SW. Relatedly, what about archived web material, for example that contained in the Internet Archive (www.archive.org), which is now being used for social science research? Second, and more importantly, the SW assumes the existence of a single ontology (a formal representation of a set of objects and the relationships between those objects). This is not useful in the context of social science research where often the ontology governing the same set of objects will differ depending on the research question that is being asked. Also, site developers may have incentives (economic or political) to obscure what they are really about. For example, in the study of the abortion debate online, it is difficult to envisage pro-life and pro-choice groups providing website SW tags that would support a single ontology that is useful for social science research.

Adaptive sampling (Thompson, 2006; Thompson and Seber, 1996) of web data may be useful in overcoming the challenge. With adaptive sampling, a "representative" sample of websites is drawn and then coded according to the research question, and inferences are then made about the underlying population of websites (see Ackland and Phillips 2007 for an application of adaptive sampling to web data). This approach fits closely with the "paradigmatic approach" described by Brent and Carnahan (2007), which can be seen as an alternative to a research approach involving the SW. Drawing again from Brent and Carnahan (2007), the paradigmatic approach has the following advantages (over and above the approach involving the SW): (1) it recognizes that there may be multiple incompatible views of data; (2) the data structure is imposed dynamically by the researcher as part of the research process (in contrast to the SW, where a data structure is imposed at the outset by web developers)

The third challenge relates to the availability of appropriate e-Research tools to enable social scientists to work with data from large-scale web crawl data. Empirical social scientists typically use menu-driven statistical software such as SPSS, Pajek and UCINET and it is unreasonable to expect them to make use of e-Research tools that have been designed to suit the work practice of scientists who submit batch files for processing on remotely-located computers. We address this particular challenge by providing a menu-driven front-end (via the VOSON System, described below).

Virtual Observatory for the Study of Online Networks (VOSON)

The VOSON System is web-based software incorporating web mining, data visualization, and more traditional empirical social science methods such as SNA (Ackland, 2005; Ackland, O'Neil, Standish and Buchhorn, 2006) - see Figure 1. The design of VOSON is intended to integrated but not monolithic, with web services facilitating access and sharing of distributed resources such as datasets, methods and computational cycles.

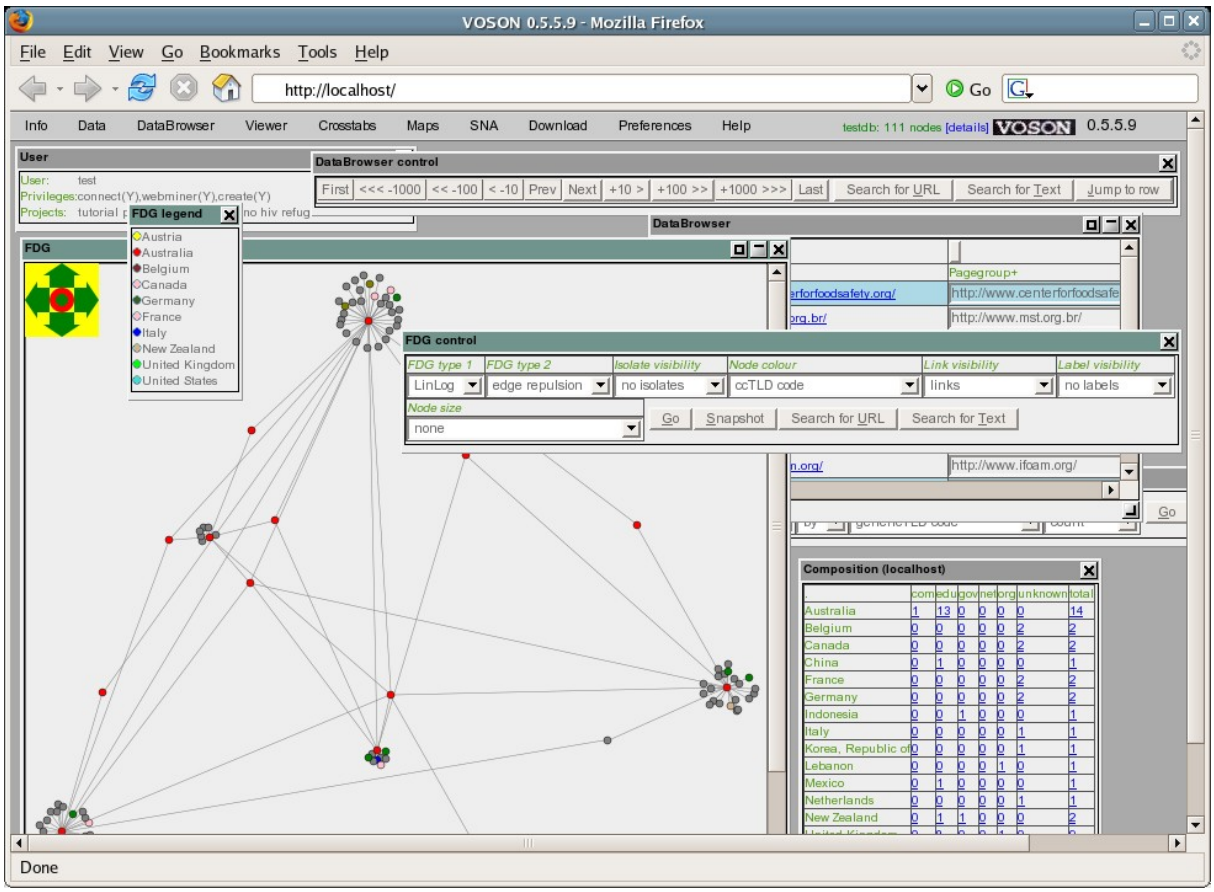

Figure 1: VOSON screenshot

VOSON features: Linux OS, PHP/javascript web interface (PEAR::HTML_AJAX[1] is used to improve interactivity), Yahoo! UI Library[2], MySQL database, Perl-based web crawler, data manipulation/analysis in Perl/C++, 3rd-party open-source statistical, text analysis and visualization tools. The current component of VOSON are:

- *Web mining and text mining.* The web crawler and Google/Yahoo APIs are used to identify hyperlinks between web pages. Page meta-data and text content are extracted and parsed.

- *Data preparation.* Data are collected at the page-level, but analysis is conducted over meaningful aggregations of pages ("pagegroups") - network nodes represent organizations, groups or people, not web pages.

- *Data visualization/analysis.* Maps showing hyperlink shortest paths between nodes are constructed using the LGL algorithm.[3] The LinLogLayout force-directed graphing algorithm[4] is used to visualize all nodes/links simultaneously. Where possible, statistical routines available in the R statistical software[5] are used, but because of the potentially large size of graphs involved, much computational work is done in C++, making use of the Boost Libraries.[6]

- *Web services.* Web services are currently implemented using PEAR::SOAP[7] and SOAP::Lite.[8] We have developed prototype C++ web services using gSOAP[9] which may be incorporated in future versions of VOSON.

Hosting VOSON services on the APAC-NF

The CSIRO web crawl dataset is much larger than web crawls previously collected and analyzed using VOSON, and it was clear that the involvement of the APAC-NF (Australian Partnership for Advanced Computing National Facility) would be key to the goal of making the CSIRO data amenable to social science research via VOSON. The APAC-NF is Australia's peak compute facility hosting both compute infrastructure (ie. supercomputers) and deep-storage facilities (in the order of petabytes). It has high bandwidth links to Australian and international networks allowing for large, efficient data movement. This in-turn allows research groups to host significant data collections as well as access software, hardware resources to mine these.

VOSON's use of web services as part of its core design permits specific data and compute functionality, which were initially performed on local machines, to be moved to facilities with larger storage and more powerful compute resources. The web services programming model permits for good software engineering practices of creating flexible systems with loosely coupled interfaces. In practice this allows VOSON to leverage both local as well as remote compute and data resources.

---

1 http://pear.php.net/package/HTML_AJAX

2 http://developer.yahoo.com/yui/menu/

3 http://apropos.icmb.utexas.edu/lgl/

4 http://www.informatik.tu-cottbus.de/~an/GD/

5 http://cran.r-project.org/

6 http://www.boost.org

7 http://pear.php.net/package/SOAP

8 http://search.cpan.org/dist/SOAP-Lite/

9 http://www.cs.fsu.edu/~engelen/soap.html

We make use of a dedicated group of machines (the Data Cluster, DC) to securely host both the CSIRO data (which are stored in XML flat files) and VOSON data processing routines which are run as web services (see Figure 2). At this early stage of our work, the DC is used for storage of the CSIRO flat-file data and the running of the Perl web crawling routines that re-crawl the CSIRO data, process the retrieved hyperlink and text data, and input these data into a VOSON MySQL database. A next step will be to host VOSON MySQL databases on the DC, and use web services to expose these data to the VOSON application server. This will enable use to take full advantage of the DC, which provides a managed, monitored point for complex database hosting. By virtue of being part of the DC, datasets automatically get backed-up for archival and reliability in case of unexpected data-loss. Further, the design of the DC allows for compute resources to sit very close to deep-storage hence allowing timely access and on-site processing of data. We intend to migrate other VOSON web services (e.g. for network mapping and sampling) onto the DC as and when the need for additional data and compute resources arises.

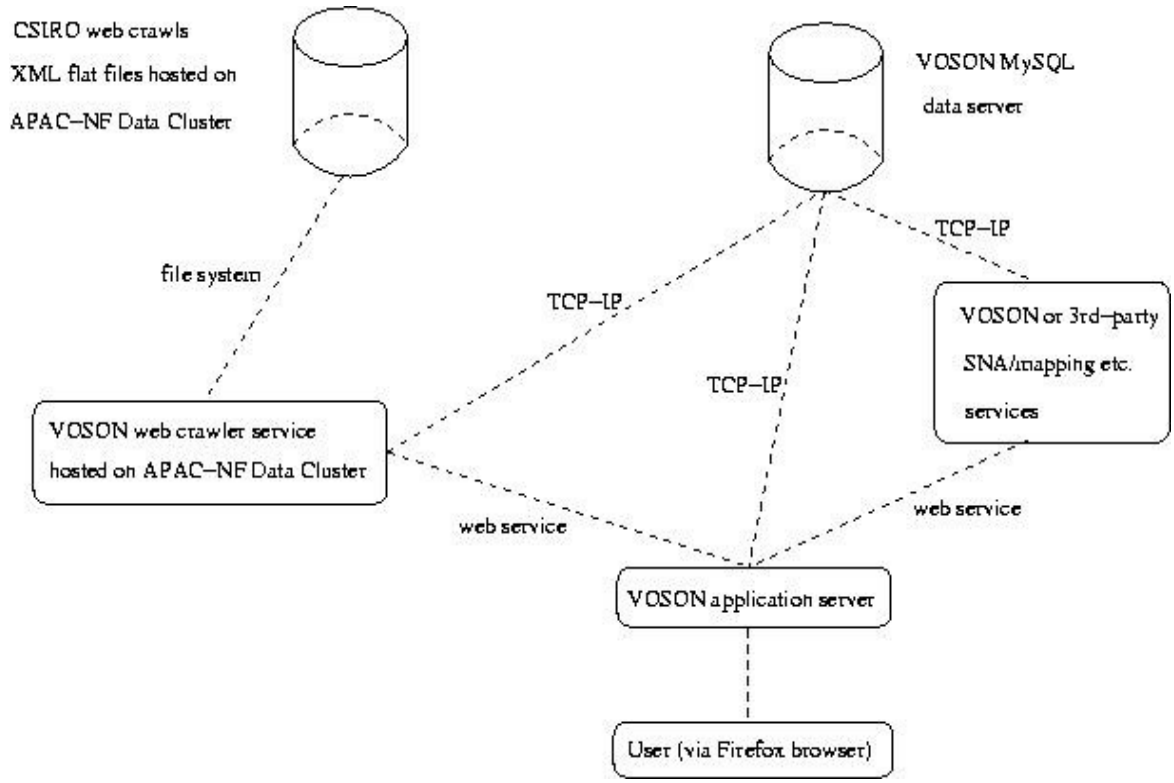

Figure 2: Architecture for analysis of CSIRO crawls via VOSON

Preparation and analysis of large-scale crawl data

In this section, we present a description of the steps involved with creating a web network dataset using a 1% sample of the CSIRO crawl data, and provide preliminary analysis of this sample. Rather than providing a thorough analysis of the data, the aim here is to give further detail on the design and performance of the various e-Research tools that are making the CSIRO crawl data amenable to social science research.

Web crawl data collected by CSIRO

As part of its information retrieval research program, the Commonwealth Scientific and Industrial Research Organisation (CSIRO) is conducting large-scale crawls of the Australian public web (defined here as publicly-available web pages with root URLs containing the .au

country-code Top-Level Domain).[10] For the present paper, we make use of a crawl conducted in late 2005 contains around 10 million web pages from approximately 200,000 websites (a website is defined here as web pages with the same root URL). Only the text information from web pages was stored in the crawl dataset (images are not collected, although the links to the images are preserved) but if a web page contains a link to a PDF or a Microsoft Office application file, a text representation of the application file was stored. An analysis of a 10 percent sample of the 2005 crawl data (Ackland, Spink and Bailey, 2007) indicated that the dataset predominantly consists of commercial websites (78 percent of sites are ".com.au") while .edu.au, .org.au and .net.au sites each account for around 5-6 percent of all sites, and .gov.au sites are approximately 2 percent.[11] For the purposes of the remainder of this paper the 200,000 websites crawled by CSIRO are referred to as "seed URLs".

Steps involved with creating a VOSON network dataset using the CSIRO crawl

The aim is to eventually create a VOSON dataset containing 100% of the CSIRO crawl, which will then be used as the basis for further analysis. Further analysis of this dataset might involve the creation of data subsets (e.g. for studying a particular aspect of the Australian public web). The dataset will also be used for extracting samples of websites using adaptive sampling procedures (where the probability of a website being sampled is related to its position in the network e.g. indegree or outdegree). In this context, the demonstration is was useful for identifying potential challenges we may face in scaling up the data collection to a 100% sample.

Creation of VOSON MySQL database containing seed URLs

The CSIRO crawl data are arranged such that each seed URL has its own directory which contains the web pages extracted by the crawler; the seed URLs were obtained by parsing the directory structure and each seed URL was inserted into the database. We also collected text content (e.g. meta keywords) from the main or entry page for each site crawled by CSIRO. In order to do this, we needed to parse the XML bundles containing the pages extracted by the CSIRO crawler and then crawl the extracted pages (by hosting them sequentially on a local web server) and this is why this step took 0.5 hours (the Perl xml parser took several seconds to parse the larger XML bundles). After this step, the VOSON database contained 1950 records.

Crawling of seed URLs

Next, each of the 1950 seed URLs was "re-crawled" by the VOSON crawler to extract hyperlinks (note: only non-intrinsic hyperlinks, or hyperlinks pointing "outside" of the site, were extracted). For this preliminary work, only the first 50 pages of each site were crawled. This step took 1.4 hours and again, much the time was taken by the Perl XML parser. After this step, the VOSON database contained 26,065 records: the 1950 seed URLs and the 24,115 web pages that the seed URLs link to.

---

10 The root URL or hostname is the portion of the URL between the "http://" and the next "/". For example, the root URL of http://www.example.com/mydir/mypage.html is "www.example.com".

11 Note that that the dataset used by Ackland, Spink and Bailey (2007) only contained text content (e.g. meta keywords) - it did not contain hyperlink data, which is the focus of the present paper.

Processing of seed URLs and URLs linked to by seed URLs

In this step, various processing was conducted in order to improve the data as a source of information on the web networking of organizations (this step took 353 seconds for the database created from the 1% sample). The database is a collection of URLs, but we ultimately want to construct web graphs where an organization represented a by a single node, rather than possibly hundreds of nodes (reflecting the 100s of pages from that organization's website that were picked up by the crawler). The steps taken were:

- URLs that did not conform with the Hypertext Transfer Protocol (HTTP), i.e. beginning with "http://" or "https://", were removed from the database. This reduced the number of records by 418.

- All URLs were reduced to their hostname or root URL - and this string is stored as the URL's "pagegroup" identifier. In a later stage (the creation of "analysis" databases), all pages from the same pagegroup are grouped into a single node in the web graph.

- Often an organization will have several hostnames (or even domain names); if this is not taken into account, then the organization will be represented by multiple nodes in the one network graph. Organizations with more than one hostname were automatically identified and placed into the same pagegroup. For example, in this step, http://voson.anu.edu.au and http://adsri.anu.edu.au would be placed in the same pagegroup.

- Depending on the research context, the above data processing step can cause problems for analysis. In some situations it would be desirable to (for example) represent all of the ANU's subsites a single node in a network, but in other contexts (e.g., research into the web presence or collaboration of particular disciplinary teams) this would be highly problematic and there would be a need to represent various ANU departments/research schools as separate nodes. For this reason, it is possible to identify (via a text file) subsites that should not be automatically grouped together (a particular use of this feature is to prevent the grouping together of sites that are hosted by the same commercial hosting service).

- Via a text file, additional sites can also be pruned from the database. The question of whether to prune a given site is dependent on the research that is being conducted. A classic example of a site that one may want to prune is http://www.adobe.com/ – many sites link to adobe to enable people to download the acrobat pdf reader and consequently, adobe is often the most central node in the network. The structure of a network is often changed markedly by pruning adobe from the database (e.g. sites that are connected via their hyperlinks to adobe and look "close together" in the force-directed graphs will become further apart once adobe is pruned).

Creation of "analysis" database

As mentioned above, the records in the VOSON database correspond to web pages, but much of the analysis we want to conduct involves representing organizations as single nodes in the network. In this step an "analysis" database is created where each record is a pagegroup or site. The size of the database was reduced by 15 percent to 21,800 records. The creation of an analysis database is fairly data intensive (because it involves adjusting link information) and it took 100 seconds for our test database.

Preliminary analysis

An analysis of the link structure of the 1% sample revealed a surprisingly amount of connectivity - 17,388 or 80% of the 21,800 nodes form a single component. Within this component, 64% of the sites are ".com", 11% are ".org", while ".edu", ".gov" and ".net" each account for 6% of the sites. As expected, the majority (52%) of the sites are Australian, while the UK the US each account for 2% of the sites, and Canada, France, Germany and New Zealand each account for 1% of the sites, and the remainder of the sites are from other countries or "unknown". The average indegree of sites in the component was 1.3, while average outdegree (calculated over the seed URLs) was 2.7. The sites were ranked according to indegree and of the top-50 sites, 14 were web technology or computer companies (e.g. adobe, google, web site traffic monitoring companies); 14 were government sites; 7 were commercial web hosting sites (e.g. users.bigpond.com.au, www.geocities.com); 6 were media sites (e.g. newspapers); 5 were company web sites, and 4 were university web sites. The above indicates the need for the data pre-processing mentioned earlier in order to control for the presence of companies that are being linked to simply because they provide software that is used in either the development of websites or by people browsing the web (e.g. adobe) and also to ensure that sites that are being hosted on commercial hosting sites are not grouped together.

Peer-produced tools and data for research into online networks

A major aim of the VOSON project is to provide a platform that will enable collaborative and seamless access to the wide range of tools and data sources that are necessary for conducting research into online networks. Web services can facilitate access to network research tools developed by 3rd parties (that is, while the network research tools that are currently accessible within VOSON have all been developed "in-house", our goal is to attract the participation of other developers of network research tools who are interested in hosting their own tools services, or at least contributing source code for services that can be hosted elsewhere).

With regards to collaborative access to and sharing of data, the VOSON system incorporates the concept of a "data common" - this is a web network database that a group of researchers jointly access and which they are encouraged to improve. Via a wiki, researchers are able to collaboratively edit configuration files that feed into the data processing steps that were outlined above. For example, via the wiki, researchers can: identify which sites to prune from the database, select sites for merging (e.g. when an organization has several domain names); flagged sites that should not be automatically grouped with other sites (e.g. because they are hosted by a common commercial hosting service), identify new seed sites for crawling, classify sites according particular research domains, modify classification schemes.

Drawing from the work of Benkler (2006), the VOSON project can therefore be seen as an example of peer-produced tools and data for the advancement of social science research into online networks. VOSON also has something in common with the emerging phenomenon of internet-enabled collaboration which Tapscott and Williams (2007) term "wikinomics" and "user innovation" studied by von Hippel (2005). Social science research into online networks involves the use of a number of tools (e.g. web crawler, text mining, social network analysis) and it is not feasible for a single tool developer to be able to "keep up" with the modifications to the tools that are needed to keep pace with developments in research methodology. Web

services can enable network researchers to make their tools available to other researchers, thus spurring further innovation in network research.

References

Ackland, R. (2005), "Virtual Observatory for the Study of Online Networks (VOSON) - Progress and Plans," *Proceedings of 1st International Conference on e-Social Science, 22-24 June 2005, Manchester.*

Ackland, R., O'Neil, M., Standish, R. and M. Buchhorn (2006), "VOSON: A Web Services Approach for Facilitating Research into Online Networks," *Proceedings of 2nd International e-Social Science Conference, 28-30 June, Manchester.*

Ackland R. and T. Pillips (2007), "Adaptive Sampling of Hyperlink Networks: An Application to Researching Social Inclusion in Australia," presentation to International Sunbelt Social Network Conference, Corfu Island, Greece, May 1-6 2007.

Ackland, R, A. Spink and P. Bailey (2007), "Characteristics of .au Websites: An Analysis of Large-Scale Web Crawl Data from 2005," *Proceedings of 13th Australasian World Wide Web Conference, 30 June - 4 July 2007, Brisbane.*

Benkler, Y. (2006), *The Wealth of Networks,* New Haven: Yale University Press.

Berners-Lee, T., J. Hendler and O. Lassila (2001), "The Semantic Web," *Scientific American,* May 2001.

Brent, E. and T. Carnahan (2007), "Artifical Intelligence and the Internet," presentation to ESRC e-Society Programme and SAGE Handbook of Online Research Methods Colloquium 28,29 March 2007. http://www.york.ac.uk/res/e-society/events/onlineresearch/AI.ppt (accessed 23rd April 2007).

Castells, Manuel (1996), *The rise of the network society. The information age: Economy, society and culture Vol. I.* London: Blackwell.

Shadbolt, N., T. Berners-Lee and W. Hall (2006), The Semantic Web Revisited. *IEEE Intelligent Systems* 21(3) pp. 96-101.

Tapscott, D. and A. Williams (2007), *Wikinomics: How Mass Collaboration Changes Everything,* Atlantic Books.

Thompson, S. (2006), "Adaptive Web Sampling," *Biometrics,* 62, 1224-1234.

Thompson, S. and G. Seber (1996), *Adaptive Sampling.* John Wiley & Sons, New York.

von Hippel, E. (2005), *Democratizing Innovation,* Creative Commons. License

Wasserman, S. and K. Faust (1994), *Social Network Analysis.* Cambridge University Press.