# Submission to the National Research Infrastructure Taskforce

29th August, 2003

Dr Robert Ackland
Research Fellow
Centre for Social Research
Research School of Social Sciences (Coombs Building, 9)
The Australian National University
CANBERRA ACT 0200
AUSTRALIA

Robert.Ackland@anu.edu.au
http://acsr.anu.edu.au/staff/ackland
ph : +61 2 6125 0312
fax : +61 2 6125 2992
mob.: 0438 833 525

### *Introduction*

The Internet is a maturing technology that is increasingly facilitating human and organisational interaction. Social scientists are turning to the Internet as a rich source of data for research on interactions and the formation of networks. The identification of online (or Web) communities can provide important insights into the networking behaviour of individuals and organisations. For example, it is of interest to see whether the ideological associations that exist between different groups (for example, political parties) carry over to the online environment, and how different groups are adapting to the Web as medium for expanding and creating networks (thus giving a measure of the relative "prominence" of different groups on the Web).

However social scientists currently do not have appropriate research infrastructure for the efficient collection and analysis of Web data. Web data is vastly different both in format and volume to the (survey-based) data usually studied by empirical social scientists, and social scientists need to look to the research of computer scientists and engineers (i.e. the people who built the Internet and WWW) for insights into the appropriate tools and methodologies.

The Virtual Observatory for the Study of Online Communities (VOSOC) is being established in the Centre for Social Research at the Research School of Social Sciences, The Australian National University. VOSOC will provide the infrastructure for quantitative research into the formation of Web communities and is being developed in the context of a political science research project which focuses on the online networking behaviour of political parties. VOSOC will be able to facilitate research into the formation of *any* type of online community, where such communities are identified by hypertext linkages between Websites. For more details, see the VOSOC homepage - http://acsr.anu.edu.au/staff/ackland/VOSOC.html.

I would now like to address the following items in the context of my experience in establishing VOSOC.

### *What significant infrastructure will you need to support research in your discipline over the coming five to 10 years?*

There are several key infrastructure requirements for VOSOC to be a successful project:

- *Information Access (Data Grid).* VOSOC will contain large-scale connectivity data sets. Since the data is collected via the Internet (using web crawlers), it is clear that it will be more efficient (from the point of view of data collection) for datasets to be located in particular locations. For example, connectivity datasets could be located in Europe, North America and Australia, and the software would ensure that Web mining processes were scheduled to run out of the server closest to the target sites. It is therefore apparent that VOSOC could potential involve large-scale data sets located at multiple sites and researchers would need to be able access these datasets in a coordinated way.
- *Cooperative Working and Visualisation.* A key element of VOSOC is the formation of a collaborative work environment. With regards to the political science research project that is currently being facilitated by VOSOC, it is envisaged that VOSOC would enable teams of political scientists, located in multiple sites, to simultaneously access and manipulate connectivity data sets. Data visualisation is also an essential aspect of VOSOC. While there has already been substantial initial research to find and incorporate appropriate visualisation technology, there may be other visualisation technologies that are better suited for VOSOC. Also, there will be substantial technical challenges involved with the implementation of visualisation technology currently employed within VOSOC to a remote computing environment; to overcome these challenges we will need access to the technical resources of a grid infrastructure program.
- *Distributed Computing.* In the software underlying VOSOC, cybermaps are constructed via the application of shortest-path algorithms. In the future, algorithms will be used to empirically identify Web communities from connectivity databases. These algorithms will become computationally more difficult as the size of the connectivity data sets increases and it will be necessary to have access to grid infrastructure that will allow distributed computing solutions.
- In the software underlying VOSOC, information on linkages between Web sites is obtained via crawlers extracting hyperlinks from html code. Characteristic information is obtained by analysts looking at the Web sites in the connectivity database. There is a lot of potential for other data mining technologies to be used to improve and augment the data on linkages and site characteristics. For example, Bayesian classification tools could be used to classify sites based on the analysis of content of Web pages.

***How will developments in technology influence research in your discipline and your infrastructure needs?***

The software underlying VOSOC is a combination of existing information retrieval and data visualisation technology. As new technology is developed, there will be scope to incorporate this technology into the VOSOC project.

***In your experience, are current funding programmes able to provide adequately for infrastructure? If not, what are the major obstacles?***

I am currently waiting the final decision from the Australian Research Council on a 2003 Discovery Grant application I made relating to VOSOC. I would, however, like to offer my initial impressions regarding the availability of funding for this type of research. This project involves the adaptation of existing technologies developed in computer science relating to data visualisation and information retrieval to a social science research project. The project is aiming to provide much-needed research infrastructure for social scientists who are interested in studying networking on the WWW. In funding programmes, this type of "cross-disciplinary" research project will be assessed by either computer scientists or social scientists (or both). There are two major obstacles that must be overcome before a project of this nature would be funded:

- *Perceptions of computer scientists.* While the computer scientists I have spoken with about this project have generally been encouraging, I have noted the following perceptions held by some individuals in this field that could potentially hinder its success in funding applications:
    - ❑ Lack of awareness of the extent to which social scientists are interested in the Internet as a source of data for empirical research. Since human behaviour is increasingly being facilitated by the Internet, then it is natural that social scientists will wish to study this behaviour. There currently do not exist appropriate tools for this type of social science research.
    - ❑ It is generally thought that social scientists only deal with relatively small quantities of data and hence do not need access to advanced computing solutions for information access, visualisation and computation. The quantities of data generated from the Web are potentially vast and advanced computing solutions are necessary for their collection, storage and analysis.
    - ❑ My impression is that computer scientists are wary of allocating funding to research which aims to create a new research tool using existing and available technologies. In fact, I had one discussion with a senior computer scientist who disagreed with my use of the term "new tool" to describe the software that underlies VOSOC, precisely because this software is comprised of existing technologies (and does not involve the development of new technology). I believe that the combining and adaptation of existing technologies into software that supports new and innovative research is a valid use of research funds.
- *Perceptions of social scientists.* This type of project also faces obstacles relating to the perceptions of social scientists:
    - ❑ Lack of awareness of the extent to which social scientists are interested in the Internet as a source of data for empirical research. In political science, for example, there are many who see the Internet as "just another medium" for political parties to get their message out, and hence should be studied using traditional tools of analysis (e.g. using techniques of "content analysis" applied to the study of political newspaper advertisements, for example, to study a party website). The Internet is in fact a medium that is entirely different to the traditional media used by political parties (exemplified by the networking function that can be facilitated by the Web site hyperlinks) and political scientists require new tools and methodologies if they are to fully utilise the data that is available from the Web.
    - ❑ Data collection activities are generally not given high value. In the social sciences, more kudos is given to those who *analyse* data rather than those who *create* data sets for analysis by others. Much of computer science research, by contrast, is devoted to the development of technologies for the efficient creation and management of data that is to be analysed by others. The differences between the two disciplines is largely because the data sets traditionally studied by social scientists (e.g. survey data) are not large – generally less than 10 megabytes – and hence social scientists have not had to pay much attention to data collection and management issues.
    - ❑ Negative perceptions towards the term "data mining". VOSOC is being established using Web mining techniques (data mining applied to data from the Web). In the social sciences, the term "data mining" is a pejorative, referring to the process of changing sample sizes or variable definitions in order to get a desired result in a test of a particular behavioural model. In order to get VOSOC

accepted as important research infrastructure it will be necessary to convince some social scientists that (a) there are new types of data that are needed to study online behaviour, and (b) data mining techniques, as developed in computer science, are essential to the construction and analysis of these new data sets.

***In your experience, what limits are imposed on research by lack of access to research infrastructure?***

In the area of research into the online networking behaviour of political groups, it is apparent that lack of access to appropriate research infrastructure has greatly limited research activities. Political scientists are generally unaware of the technical aspects of the Internet and the technologies that have been developed in computer science to collect and analyse Web data. To date, the majority of empirical political science research on political party use of the Internet for networking functions has involved researchers using Web browsers to look at party sites and manually code outbound links (and inbound links via the Google browser interface). The need for manual techniques for collecting Web data has clearly limited the research of political scientists both in terms of the quantity of data that can be collected and the study of changes in networking behaviour over time.

Computer scientists, on the other hand, are well aware of Internet-related technologies and tools (having built them), but are not aware of the research interests of political and other social scientists. What VOSOC aims to do is to bridge this gap by providing appropriate research infrastructure for quantitative research into the formation of Web communities and networking behaviour.

***In your view, what is the potential for collaborative use of infrastructure, including with overseas facilities?***

There is great potential for collaborative use of grid computing infrastructure.