Digital Methods for Internet Research

Part I

Prelude

L. Rudner & W. Schafer (1999). “How to write a scholarly research report.” *Practical Assessment, Research & Evaluation*, 6(13), http://pareonline.net/getvn.asp?v=6&n=13 (accessed 10 November 2009).

Modern Language Association (2009). *MLA Handbook for Writers of Research Papers.* 7th ed. New York: MLA. (UvA Media Studies version)

1. Virtual methods & digital methods

B. Wellman (2010). “Studying the Internet Through the Ages,” in M. Consalvo and C. Ess (eds.), *The Handbook of Internet Studies.* Wiley-Blackwell: 17-23.

D. Lazer et al. (2009). “Life in the network: the coming age of computational social science,” *Science*. 323(5915): 721-723.

C. Borgman (2010). “The Digital Future is Now: A Call to Action for the Humanities.” *Digital Humanities Quarterly*. Fall, 3(4).

R. Rogers (2009). *The end of the virtual - Digital methods*. Amsterdam: Amsterdam University Press.

Manovich, L. (2009), “How to Follow Global Digital Cultures, or Cultural Analytics for Beginners,” in K. Becker and F. Stalder (eds.) *Deep Search: The Politics of Search beyond Google*. Innsbruck: Studienverlag, 198-212.

2. Internet censorship research. History and analysis

N. Villeneuve (2007). "Evasion tactics: Global online censorship is growing, but so are the means to challenge it and protect privacy." *Index on Censorship*. 36(4): 71-85.

J. Zittrain, and B. Edelman (2002). "Documentation of Internet filtering in Saudi Arabia." Working Paper, Berkman Center for Internet & Society, Harvard Law School.

R. Faris and N. Villeneuve (2008). "Measuring Global Internet Filtering." in R. Deibert et al. (eds.), *Access denied: The practice and policy of global Internet filtering.* Cambridge, MA: MIT Press: 5-27.

R. Rogers (2009). "The Internet treats censorship as a malfunction and routes around it? A new media approach to the study of state Internet censorship," in J. Parikka and T. Sampson (eds.), *The spam book: On viruses, porn, and other anomalies from the dark side of digital culture*. Cresskill, NJ: Hampton Press, 229-247.

R. Deibert and R. Rohozinski (2010). “Control and Subversion in Russian Cyberspace,” *Access Controlled*. Cambridge, MA: MIT Press, 15:34.

J. Wright, T. de Souza and I. Brown (2011). “Fine-Grained Censorship Mapping: Information Sources, Legality and Ethics.” FOCI'11 (USENIX Security Symposium), San Francisco, 8 August.

3. Website histories and historiographies

M. Dougherty, E.T. Meyer, C. Madsen, C. van den Heuvel, A. Thomas and S. Wyatt (2010). *Researcher Engagement with Web Archives: State of the Art.* London: JISC

B. A. Howell (2006). "Proving web history: How to use the Internet archive." *Journal of Internet Law*. 9(8): 3-9.

J. Murphy, N.H. Hashim & P. O'Connor (2007). "Take me back: Validating the Wayback Machine." *Journal of Computer-Mediated Communication*. 13(1).

R. Rogers (in press). *Digital Methods*. Cambridge, MA: MIT Press (excerpt: The Website as Archived Object).

S. Schneider and K. Foot (2004). “The Web as an object of study.” *New Media and Society*. 6(1): 114-122.

0. Prelude

L. Rudner & W. Schafer (1999). “How to write a scholarly research report.” *Practical Assessment, Research & Evaluation*, 6(13).

Rudner, Lawrence M. & William D. Schafer (1999). "How to write a scholarly research report." *Practical Assessment, Research & Evaluation*, 6(13), *http://pareonline.net/getvn.asp?v=6&n=13 (accessed 10 November 2009)*.

How to Write a Scholarly Research Report

Lawrence M. Rudner, ERIC & University of Maryland
William D. Schafer, University of Maryland

Researchers communicate their results and help accumulate knowledge through conference papers, reports, on-line journals and print journals. While there are many rewards for having research disseminated in a scholarly outlet, the preparation of a good research report is not a trivial task.

This article discusses the common sections of a research report along with frequently made mistakes. While the emphasis here is on reports prepared for scholarly, peer-reviewed publication, these points are applicable to other forms of research reports. Dissertations and theses, for example, provide more detail than scholarly publications yet they adhere to the same basic scientific writing principles. Since all scientific research involves observation, description and analysis, points made in this article are applicable to historical and descriptive, as well as to experimental, research.

[...]

FIRST STEPS IN WRITING A RESEARCH REPORT

You should constantly think about writing your report at every stage of your research activities. [...]

Plan your report to focus on a single important finding or highly related group of findings. In the process of analyzing your data, you probably uncovered many relationships and gained numerous insights into the problem. Your journal article submission, however, should contain only one key point. The point should be so fundamental that you should be able express it in one sentence or, at most, in a paragraph. If you have several key points, consider writing multiple manuscripts.

When writing your manuscript, keep in mind that the purpose is to inform the readers of what you investigated, why and how you conducted your investigation, the results and your conclusions. As the investigator and writer, your job is simply to report, not to convince and usually not to advocate. You must provide enough detail so readers can reach their own conclusions about the quality of your research and the veracity of your conclusions.

SECTIONS OF YOUR REPORT

Title - It is important that the title be both brief and descriptive of your research. Search engines will use the title to help locate your article. Readers make quick

decisions as to whether they are going to invest the time to read your article largely based on the title. Thus, the title should not contain jargon or vernacular. Rather, the title should be short (generally 15 words or less) and clearly indicate what the study is about. If in doubt, try to specify the cause and effect relationship in your key point. Avoid trite and wasteful phrases such as "A study of ..." or "An investigation to determine ..."

{Personal data - For the sake of the course Digital Methods for Internet Research, include your name, student number, the course, the assignment number, your supervisor, your team members, the date, and your email address.}

[…]

Introduction - You will usually start your report with a paragraph or two presenting the investigated problem, the importance of the study, and an overview of your research strategy. You do not need to label this section. Its position within the paper makes that obvious.

The introductory paragraphs are usually followed by a review of the literature. Show how your research builds on prior knowledge by presenting and evaluating what is already known about your research problem. Assume that the readers possess a broad knowledge of the field, but not the cited articles, books and papers. Discuss the findings of works that are pertinent to your specific issue. You usually will not need to elaborate on methods *here*.

The goal of the introduction and literature review is to demonstrate “the logical continuity between previous and present work” (APA, 1994, p. 11). This does not mean you need to provide an exhaustive historical review. Analyze the relationships among the related studies instead of presenting a series of seemingly unrelated abstracts or annotations. The introduction should motivate the study. The reader should understand why the problem was researched and why the study represents a contribution to existing knowledge. Unless the study is an evaluation of a program, it is generally inappropriate to attempt to motivate the study based on its social importance.

Method - The method section includes separate descriptions of the sample, the materials, and the procedures. These are subtitled and may be augmented by further sections, if needed.

Describe your sample with sufficient detail so that it is clear what the sample represents. A discussion of how the sample was formed is needed for replicability and understanding your study. [...]

A description of your instruments, including all surveys, tests, questionnaires, interview forms, searches, and other tools used to provide data, should appear in the materials subsection. Evidence of reliability and validity should be presented. Since reliability is a property of scores from a specific use of a specific instrument for a specific population, you should provide reliability estimates based on your data.

The design of the study, whether it is a case study, a controlled experiment, a meta- analysis, or some other type of research, is conveyed through the procedures subsection. It is here that the activities of the researcher are described, such as what was said to the participants, how groups were formed, what control mechanisms were employed, etc. The description is sufficient if enough detail is present for the reader to replicate the essential elements of the study. [...]

Results - Present a summary of what you found in the results section. Here you should describe the techniques that you used, each analysis and the results of each analysis.

Start with a description of any complications, such as [...] missing data that may have occurred. Examine your data for anomalies, such as outliers, points of high influence, miscoded data, and illogical responses. Use your common sense to evaluate the quality of your data and make adjustments if need be. Describe the process that you used in order to assure your readers that your editing was appropriate and purified rather than skewed your results. [...]

For most research reports, the results should provide the summary details about what you found rather than an exhaustive listing of every possible analysis and every data point. Use carefully planned tables and graphs. While tables and graphs should be self- explanatory, do not include a table or graph unless it is discussed in the report. Limit them to those that help the reader understand your data as they relate to the investigated problem.

Discussion - At this point, you are the expert on your data set and an authority on the problem you addressed. In this section, discuss and interpret your data for the reader, tell the reader of the implications of your findings and make recommendations. Do not be afraid to state your opinions.

Many authors chose to begin the discussion section by highlighting key results. Return to the specific problem you investigated and tell the reader what you now think and why. Relate your findings to those of previous studies, by explaining relationships and supporting or disagreeing with what others have found. Describe your logic and draw your conclusions. Be careful, however, not to over generalize your results. Your conclusions should be warranted by your study and your data.

Be sure to recognize the limitations of your study. Try to anticipate the questions a reader will have and suggest what problems should be researched next in order to extend your findings into new areas.

References - There should be a one-to-one match between the references cited in the report and the references listed in the reference section.

{*For footnotes/endnotes as well as the bibliographic references, please follow the MLA style guide: http://owl.english.purdue.edu/owl/resource/747/01/}*

*{**Appendix** - Attach your complete data sets to your paper or, preferably, make them available on-line and reference them in your paper. Researchers would like to be able to verify results.}*

Editor's note: The reference to APA in the text refers to the American Psychology Association's style guide. In Media Studies at the University of Amsterdam, please use the MLA style (Modern Language Association).

Modern Language Association (2009). *MLA Handbook for Writers of Research Papers*. 7th ed. New York: MLA. (UvA Media Studies version)

MLA

REFERENCING GUIDE

October 2011

Contents

Section 1 – Introduction

Section 2 – Formats for Printed Material

Section 3 – Formats for Electronic Material

Section 4 – Formats for Other Material Types

Section 5 – FAQ

Section 1 – Introduction

The aim of this guide is to offer an introduction to the practice of referencing to students who are preparing written assignments for academic credit at the Department of Media Studies of the University of Amsterdam. When writing an academic essay, students are required to refer to the work of other authors. Each time they do so, it is necessary to identify their work by making reference of it – both in the text of your essay and in a list at the end of your essay (in the reference list or bibliography). This practice of acknowledging authors is known as referencing.

There are many academic referencing systems used in academic writing. This guide explains the **MLA system**, which is one of the most used systems of citation (particularly in the humanities) and the system that we use at the Department of Media Studies. The MLA style of citation has been developed by the Modern Language Association (http://www.mla.org/) and provides an in-text method of referencing sources. Within this system, each reference consists of two parts: the *parenthetical reference*, which only provides brief identifying information within the text (author's surname and page numbers), and the *Reference List* (or *Works Cited*) which provides full bibliographic information.

The two-part references must be provided whenever you use – i.e. quote or paraphrase – someone else's opinions, theories, data or organisation of material. You need to reference information from books, articles, websites, videos, other print or electronic sources, and personal communications. All these different types of material need specific referencing. In other words, each type has an accepted 'format' for presentation within the Reference List (or Works Cited).

The following is a set of guidelines for formatting references in your Reference List as well for referencing sources in the body paragraphs of your assignment (in-text referencing). The coming three sections provide the format style (followed by an example) of all sorts of reference list entries. They are broadly separated into 'Printed Material' (Section 2), 'Electronic Material' (Section 3) and 'Other Material' (Section 4). Section 5, entitled 'FAQ', explains the format style of your in-text references and discusses particular issues you may encounter while formatting your references, both in your text and your Reference List. The 'question & answer' format is used so that you can check areas of specific concern easily.

After reading this guide, you should be able to:

- understand how to use the MLA referencing system
- indicate others writers' ideas in your own work using an accepted citation style
- format appropriate references correctly from these citations
- deal with a range of bibliographic and electronically formatted material

Before you start reading, please keep in mind that one golden rule applies:

Be consistent in everything you do!

This consistency applies to format, layout, type-face and punctuation.

Section 2 – Formats for Printed Material

Nb.

- Always remember to use correct source information for all your references and the same punctuation consistently in each kind of format
- Note the consistency of use of *italics* for titles. Italics are the preferred format but it is acceptable to underline
- The place of publication is the city (normally the first stated), *not* the country
- Authors should appear in the order that they are presented on the title page of the source; only the first author's name is reversed

2.1 Book

Author Surname, First Name. *Title.* Place of publication: Publisher, Year of publication.

Eg.

Fraser, Matthew. *Weapons of Mass Distraction: Soft Power and American Empire*. New York: St. Martin's Press, 2003.

Nb.

If you refer to a republished book, add the original publication year after the title.

Eg.

Klein, Naomi. *No Logo*. 2000. New York: Picador, 2002.

Nb.

With titles in **Dutch**, **French**, **Spanish**, and most other non-English languages, only the first word is capitalized. With titles in German and Luxembourgish, all nouns are capitalized according to their writing system. The same holds for titles of chapters and articles. The title of a journal is always capitalized according to the title case capitalization (see also page 6, 18 and 20).

Eg.

Hermes, Joke, and Maarten Reesink. *Inleiding televisiestudies*. Amsterdam: Boom, 2003.

Kracauer, Siegfried. *Von Caligari zu Hitler: eine psychologische Geschichte des deutschen Films*. Frankfurt am Main: Suhrkamp, 1984.

2.2 Book Chapter

To refer to a specific chapter of a book by one and the same author, add the chapter title and page numbers.

Author Surname, First Name. "Title Article." *Title Book.* Edition. Place of publication: Publisher, Year of publication. Page numbers.

Eg.

Atton, Chris. “Approaching Alternative Media: Theory and Methodology.” *Alternative Media*. London, Thousand Oaks and New Delhi: Sage Publications, 2001. 7-32.

2.3 Anthology or Edited Book

To refer to the edited book as a whole, quote the editor(s) in the text. In the reference list you then indicate editorship by using either ed. for a single editor or eds. for more than one editor.

Eg.

Yeager, Patricia, ed. *The Geography of Identity*. Michigan: University of Michigan Press, 1998.

2.4 Chapter in Anthology or Edited Book

An edited book will often have a number of authors for different chapters (on different topics). To refer to a specific author’s ideas (from a chapter), quote them in the text – not the editors. Then in your reference list indicate the chapter details *and* the book details from which it was published.

Author Surname, First Name. “Title Article.” *Title Book.* Ed./Eds. First Name Surname editor(s). Place of publication: Publisher, Year of publication. Page numbers.

Eg.

Fornäs, Johan. “Media Passages in Urban Spaces of Consumption.” *Geographies of Communication: The Spatial Turn in Media Studies*. Eds. Jesper Falkheimer and André Jansson. Göteborg: Nordicom, 2006. 205-20.

2.5 Journal Article

Author Surname, First Name. “Title of Article.” *Journal Title* Volume.Part number (Year of publication): page numbers.

Do not worry about omitting the part number if not available.

Nb.

The month of publication may be added prior to the year of publication, especially if the part number is not known. If you do, be consistent and include it in all your references to journal articles.

Eg.

Gates, Philippa. “Always a Partner in Crime: Black Masculinity in the Hollywood Detective Film.” *Journal of Popular Culture* 32.1 (Spring 2004): 20-9.

Scheijen, Suzanne van. “Financiële crisis in de media: framingsonderzoek naar berichtgeving over financiële crises in Nederlandse dagbladen.” *Tijdschrift voor Mediageschiedenis* 1 (June 2011): 45-63.

2.6 Corporate Author

Sometimes it is impossible to find a named individual as an author. What has usually happened is that there has been a shared or ‘corporate’ responsibility for the production of the material. Therefore the ‘corporate name’ becomes the author (often called the ‘corporate author’). Corporate authors can be government bodies, companies, professional bodies, clubs or societies, and international organizations.

Format is the same as for a book, but uses the ‘corporate’ (company, business, organisation) author in place of a named author.

Eg.

Institute of Waste Management. *Ways to Improve Recycling.* Northampton: Institute of Waste Management, 1995.

Nb.

For journal articles without authors the journal title becomes both author and cited journal title.

2.7 Conference Proceeding

Treat published proceedings of a conference like an edited book, but add information about the conference.

Editor Surname, First Name, ed./eds. *Title of Proceedings.* Conference Proceedings Title, Date, Place. Place of Publication: Publisher, Year of publication.

Eg.

Freed, Barbara, ed. *Foreign Language Acquisition Research and the Classroom*. Conference Proceedings of Consortium for Language Teaching Conference, October 1989, University of Pennsylvania. Lexington: Heath, 1991.

Cite a paper in the proceedings like a work in a collection of pieces by different authors.

Author Surname, First Name. “Title of Paper.” *Title of Proceedings*, *date, place.* Ed. Place of Publication: Publisher, Year of Publication. Page numbers.

Eg.

Mann, Jill. “Chaucer and the Woman Question.” *This Noble Craft: Proceedings of the Tenth Research Symposium of the Dutch and Belgian University Teachers of Old English and Historical Linguistics, Utrecht, 19-20 January 1989*. Ed. Erik Kooper. Amsterdam: Rodopi, 1991. 173-88.

2.8 Newspaper Article

Journalist Surname, First Name. "Title of News Item." *Name of Newspaper*. Date of publication, Page number.

Eg.

Peters, Roger. "Picking up Maxwell's Bills." *Independent.* 4 June 1992, 28.

Nb.

If the page number is not marked or otherwise unavailable, leave out this information. If it is a news article and does not attribute an author, begin the entry with the title of the article.

Eg.

"Lottery for Breast Cancer Helps." *The Guardian.* 21 March 1995.

2.9 Article from Reference Book

Author Surname, First Name (if given). "Title of Article." *Name of Encyclopedia*. Edition. Year of publication.

Eg.

Avery, Jennie. "Poland." *Encyclopaedia Britannica*. 2nd ed. 1994.
"Accord." *The Oxford English Dictionary*. 2nd ed. 1989.

If the reference book does not arrange its articles alphabetically, try including the volume and page numbers:

"Cold War." *Columbia Encyclopedia*. 5th ed. 5th vol. 1998. 12-15.

If the reference book is not well known, provide full publication information:
"Euthanasia." *Encyclopedia of World Ethics*. 2nd ed. 7th vol. New York: Simon Press, 2001. 54-68.

2.10 Dissertation & Thesis

Cite a **published** dissertation like a book adding useful dissertation information before the publication facts.

Author Surname, First Name. *Title.* Diss. (Level of dissertation). Awarding Institution, Publisher: Place, Date.

Eg.

Valentine, Mary-Blair Truesdell. *An Investigation of Gender-based Leadership Styles of Male and Female Officers in the United States Army*. Diss. (Ph.D Thesis). George Mason University, 1993. Ann Arbor: UMI, 1993.

An **unpublished** dissertation (or thesis) should have the title details enclosed in quotation marks, with the added descriptive label Unpublished Diss.*,* and then add the level of the dissertation and the awarding institution followed by a comma and the year of completion.

Author Surname, First Name. "Title." Unpublished Diss. (Level of dissertation). Awarding Institution, Year of completion.

Eg.

Kirkland, John. "Lay Pressure Groups in the Education System: A Study of Two English Boroughs." Unpublished Diss. (Ph.D. Thesis). Brunel University, 1988.

2.11 Government or Legal Documents

Available data may vary for these, but where possible include the following:

Government Department/Institute. Subdivision of Department/Institute (if known). *Title of Document.* (Name of chairperson if it is a committee). Place of publication: Publisher, Year of publication.

Eg.

Department of Health and Social Services*. Inequalities in Health: Report of a Research Group.* (Chairman: Sir Douglas Black). London: DHSS, 1980.

Ministry of Education, Youth and Culture. Culture Division. *The National Cultural Policy of Jamaica: Towards Jamaica the Cultural Superstate*. Kingston: Culture Division, 2003.

2.12 Map

Creator's Surname, First Name. (may be mapmaker, cartographer compiler etc.) *Title.* Scale (normally given as a ratio). Place of publication: Publisher, Year of publication.

Eg.

Jones, Harold. *East Anglia: North.* 1:10,000. Peterborough: Grove, 1953.

Nb.

If the name of the creator/originator is not known, use the title of the map in its place.

2.13 Work of Art

Artist Surname, First Name. *Title.* Material type, measurements. Place: Gallery, Date of creation.

Eg.

Renoir, Pierrre-August. *The Skiff.* Oil on canvas, 71 x 92 cm. London: The National Gallery, 1875.

Section 3 – Formats for Electronic Material

Nb.

- The principles for referencing electronic materials are in general the same as for other types of materials.
- The nature of web publications can often mean that author names and publication dates are unavailable. The solution to this problem is to decide who is responsible for producing the source and they will then become the 'author'.
- It is often easier to find information if you look at the Home Page link for the site you are in or at the 'About Us' or 'Contact Us' type of links.

3.1 General Web Page

Name of website. Editor(s) of the website (if given). Year of publication. Associated institution. Date of access. <URL>.

Do not worry about omitting the editor(s) of the website if not available.

Eg.

BBC on the Internet. 2005. British Broadcasting Company. 12 April 2005. <http://www.bbc.com>.

Nb.

The date of access is the date which you viewed or downloaded the document. It may be subject to changes or updating and including this date in your reference allows for this possibility.

3.2 Specific Web Article

Author Surname, First Name. "Title." *Name of Website*. Editor(s) of website (if given). Year of publication. Associated institution (if known). Date of access. <URL>.

Do not worry about omitting the editor(s) of the website and associated institution if not available.

Eg.

Smith, Fred. "New Football Recruits." *Northwestern Football*. Ed. Alex Shokey. 2004. Northwestern University. 6 June 2004. <http://www.football.northwestern.edu/recruits>.

Nb.

If a web article does not contain page numbers use n. pag. (no pagination) in place of page numbers.

3.3 E-book

Referencing an e-book, first include the same information as a regular book. After citing the original publication information, add the electronic publication information. The format is then as follows:

Author Surname, First Name. *Title*. Place of publication: Publisher, Year of publication. *Name of website*. Editor of the website (if given). Date of electronic publication (if known). Associated institution (if known). Date of access. <URL>.

Do not worry about omitting the editor(s) of the website, the date of electronic publication, and associated institution if they are not available.

Eg.

Hutcheon, Leonell. *Politics of Postmodernism*. London: Routledge, 2002. *Eblib*. 3 August 2009. <http://reader.eblib.com/Reader.aspx?p=181639&o>.

3.4 Article in Electronic Journal (WWW)

Some journals are published **freely** and **solely** on the internet, and therefore it is advised to add information about its online presence when citing an article from such a journal. The format for this is:

Author Surname, First Name. "Title." *Journal Title* Volume number.Issue number (Year of Publication): Page numbers. Date of access. <URL>.

Eg.

Hillis, Ken. "Los Angeles as Moving Picture." *Aether: The Journal of Media Geography* 6.A (2010): 1-9. 17 August 2011. <http://130.166.124.2/~aether/pdf/volume_06/hillis.pdf>.

Nb.

- The month of publication may be included before the year of publication, especially is the part number is not known. If you do, be consistent and add it in all your references of journal articles.
- If a journal exists in both print and electronic form it is often simpler to use the print journal format for referencing the item, regardless of which item you have viewed.

3.5 Wiki Article

Wiki name. "Title of Article." *Associated Institution*. Year of publication. Date of access. <URL>.

Eg.

Wikipedia. "William Shakespeare." *Wikimedia Foundation*. 2008. 3 July 2010. <http://en.wikipedia.org/wiki/William_shakespeare>.

3.6 Blog

Author Surname, First name. "Title of Blog Entry." *Title of Blog.* Associated institution. Date of posting. Date of access. <URL>.

Eg.

Shaw, Andrea. "Tia Dalma's Portrayal in *Pirates of the Caribbean*." *Ordinary Anointments*. 10 June 2011. 11 August 2011. <http://blogs.jamaicans.com/ordinarya/>.

3.7 Online Video (e.g. YouTube)

For online videos, provide the author only if you are sure that person created the video. Do not list the person who posted the video online as the author. If you are unsure, treat the citation as having no author.

Creator (if available). "Title of Post." *Title of Website.* Date of creation/upload. Date of access. <URL>.

Eg.

Takayma-Ogawa, Joan, and Jeanne Willette. "What is Information Literacy?" *YouTube*. 14 March 2007. 20 April 2010. <http://www.youtube.com/watch?v=yeopJX5jJV8>.

"Slingshot Fun." *YouTube*. 29 January 2007. 30 April 2010. <http://www.youtube.com/watch?v=CCmZYce0J2E>.

3.8 E-mail

Senders Surname, First Name. (Senders e-mail address), "Subject of Message." E-mail to: First Name Surname (Recipients email address). Date sent (Day month year).

Eg.

Halmond, Kirsty. (Khalmond@imaginary.co.uk), "Changes to Report Style Format." E-mail to: Carl Brown (Carl-brown234@daylight.com). 12 July 2008.

Nb.

E-mail messages are usually only cited in the running text ("In an email to the author on July 12, 2010, Kirsty Halmond revealed . . .") and are rarely listed in the reference list. In parenthetical citations, the term personal communication (or pers. comm.) can be used.

Section 4 – Formats for Other Material Types

Nb.

- It is advised to create a separate ***Film List*** or ***Media List*** when you have used more than two films or other media resources (including online videos) respectively.

4.1 Film

Title. Dir. Name Director. Distributor, Year of release.

If appropriate you can include the names of writers, performers and producer – between the title and the distributor.

Eg.

The Apartment. Dir. Billy Wilder. United Artists, 1960.

4.2 TV/Radio Programme

"Episode Title." Episode number. *Programme/Series Title*. Network. Transmission date.

Do not worry about omitting the episode title and number if they are not available.

Eg.

"The Empty Child." Episode 9. *Doctor Who.* BBC1. 21 May 2005.
The Voice of Holland. RTL4. 15 October 2010.
Women's Hour. BBC Radio 4. 29 July 2004.

4.3 Video Game

Title. Version number (if available). Designed by First Name Surname Designer (if available). Publisher, Release Year.

Eg.

Donkey Kong. Designed by Shigeru Miyamoto. Nintendo, 1991.
The Sims. 2. Electronic Arts, 2004.

4.4 Musical Score

Composer Surname, First Name. *Title of Work*. Ed./Eds. Name Editor(s). You could also add other arrangers, for example Scored by or Arranged by (note that name is not written surname first). Place of publication: Publisher, Year of publication.

Eg.

Mozart, Wolfgang Amadeus. *Flute Concertos: Concerto no. 2 in D, K. 314 and Andante in C, K. 315.* Ed. Tony Wye. Sevenoaks: Novello, 1983.

4.5 Online Image

Originator. *Title of Image*. Year of creation/upload. Date of Access. <URL>.

Eg.

Daisy Chains. *Victoria Butterfly Gardens.* 2009. 3 August 2009. <http://www.flickr.com/photos/69561650@N00/3784458656/>.

4.6 Personal Communication; Conversations, Interviews and Telephone Calls

As this data has not been published anywhere (and is therefore not recoverable), details should only be recorded within the text.

Surname, First name. Type of communication (e.g. interview or Personal communication), Date of communication.

Eg.

…we need to "invest more money in student accommodation" (Jones, Sally. Interview, 25 August 2005) and until we do…

4.7 Lecture Notes

Lecturer's Surname, First name. "Title of Lecture." *Course/Series Title*. Institution. City, Date.

Eg.

Martens, Emiel. "Introduction: Towards a Theory of Media Activism." *Media Activism*. University of Amsterdam. Amsterdam, 9 September 2011.

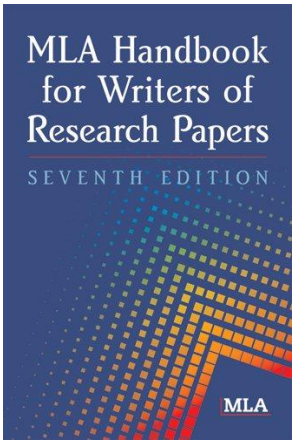
If you use your own (unpublished) notes taken at the lecture, details should only be recorded within the text.

Eg.

During the first lecture of the course *Media Activism*, which was entitled "Introduction: Towards a Theory of Media Activism" (9 September 2011), Emiel Martens showed that…

Section 5 – FAQ

1. What is referencing?



When preparing a piece of written work you will inevitably come across other peoples' ideas, theories or data which you will want to make reference to in your own work. Making reference to others is called **'citing'**, and the list of these authors' works are given at the end of a piece of written work in the form of a **'reference list'** or **'works cited'**. The process of citing authors (and the associated reference list) can be done in a number of styles. This guide presents the **MLA Style** (http://www.mla.org/) as described in the *MLA Handbook for Writers of Research Papers,* 7th ed. New York: The Modern Language Association of America, 2009. This is the style we use at the Department of Media Studies.

2. Why should I reference?

- To show evidence of the breadth of your research
- To strengthen your argument
- To acknowledge other peoples' ideas correctly
- To allow the reader of your work to locate the cited references easily, and so evaluate your interpretation of those ideas
- To avoid plagiarism
- To avoid losing marks!

3. What is a Reference List or Works Cited?

At the end of your essay under the heading **'Reference List'** or **'Works Cited'** you list all the items you have made reference to in your essay. This list of books, journals, newspaper articles (or whatever) is organised alphabetically by the surnames of the authors (or originators) of the work.

As a student in Media Studies, you will often refer in your essay to films, television programmes, websites and other media works. To maintain a clear arrangement, it is advised to include a separate **'Film List'** or **'Media List'** when you have used these media. List all these items alphabetically by title.

4. Where do I find the information that I need for my list of references?

Usually from the title page (or reverse title page) of the book or document you are citing. Remember though that:

- The *order* of authors' names should be retained
- Cite the first named *place* of publication

- Edition dates are *not* reprint dates (new editions will have new text and must be cited as such). The copyright sign © will often indicate the date of production

If your material has not originated from a commercial publisher and lacks obvious title page data, then the appropriate information should be gleaned from any part of the publication, if you can say with some certainty that it fulfils the required criteria for your reference list.

5. How do I present referred material in my essay?

You present material in two main ways:

- **Quoting** material directly from its source – word for word as it was in its original form. Your essay should not be a 'cut and paste' exercise using other peoples' words. Use quotations only when you have to use the text in its original form or for presenting a longer quote which you use to highlight and expand on ideas or issues in your essay.

- **Paraphrasing (or summarizing)** text that you have read. Putting the ideas into your own words (in the context of answering the question) and then stating where that information came from (see next section). Paraphrasing and summarizing is a skill that needs to be practiced and developed.

6. How do I format parenthetical references?

Each source in the reference list at the end of your essay corresponds to a reference in the text. In MLA style, in-text references are called parenthetical references. When you quote or paraphrase someone else's work, you give the author's surname followed by the page number(s) in parentheses, generally at the end of the sentence.

There is no punctuation between the name and the page number. When you mention the author in the sentence itself, you need only give the page number. You do not need to cite page numbers if you are referring to an entire work, or if the work is only one page long.

Eg.
(Beeton 23)

If the author's name is mentioned in the text, only the page number(s) need(s) to appear in the parentheses.

Beeton argues that "film-induced tourism was not the sole driver for international tourism growth of the 1980s" (23).

Citations when you are using more than one work by the same author. If you are referring to more than one of a particular person's works in your essay, you add an abbreviated title in parentheses, with a comma between the surname of the author and the title of the work.

(Klein, *No Logo* 177)
(Klein, *Shock Doctrine* 235)

Again, if you mention the author's name in the sentence, you leave it out of the parentheses. If you also mention the title of the work in your sentence, you leave that out as well.

Naomi Klein, in her *No Logo*, states that "when we try to communicate with each other by using the language of brands and logos, we run the very real risk of getting sued" (177).

Nb.

You also include the abbreviated title of the works when you use two different authors with the same surname.

Citing work by two or three authors. Use the last names of each.

(Falkheimer and Jansson 15)
(Ashcroft, Griffiths and Tiffin 8)

Citing work by more than three authors. Give all the authors' last names or just use the first and 'et al.' (meaning 'and others') for the rest. *In any case, use the same form as the entry in your Reference List.*

(Bia, Pedreno, Small, Finch and Patterson 161)
(Bia et al. 161)

Citing work by groups or corporate authors. Use full name of group or a shortened form.

(Modern Language Association 115) (MLA 115)

Citing work by an unknown author. Use a few words of the title.

(*Recent Innovations* 231)

Citing more than one work. Use semicolons to separate the citations.

(Leane 54; Johnston 80-3)

Nb.

For exact quotations from sources without page numbers, use paragraph numbers, if available. If the work does not have page numbers or paragraph numbers, you leave out this information.

Citations taken from a secondary source should generally be avoided; consult the original work whenever possible. If only an indirect source is available, put the abbreviation 'qtd. in' (quoted in) before the indirect source in the parenthetical reference and include the indirect source in the Reference List.

In a May 1800 letter to Watt, Creighton wrote, "The excellent Satanism reflects immortal honor on the Club" (qtd. in Hunt and Jacob 493).

If the reference is a film, radio/TV programme or video game, you only refer to the title of the film, programme (episode or series) or game. The **first time** you mention the film, programme or game in the text, you also include the year of publication (release/transmission) in parenthesis following the title.

The Apartment (1960) "The Empty Child" (2005) *Women's Hour* (2004)
Donkey Kong (1991) *The Sims 2* (2004)

7. How do I incorporate long quotations in my essay?

Any quotation that is three lines or less is considered a short quotation and should be incorporated into your sentence.

Longer quotations of **four typed lines** or more should be:

- preceded by a colon
- indented from your main text
- *not* have quotation marks
- typed single space
- cite author (if not mentioned in the text) and page numbers
- The final punctuation comes before the parentheses

Eg.

Certain passages are remarkable for their poetic quality:

> It was just a fragment, no more than 30 seconds: The Euston Road, hansoms, horse drawn trams, passers-by glancing at the camera but hurrying by without the fascination or recognition that came later. It looked like a still photograph, and had the superb picture quality found in expert work of the period, but this photograph moved. (Walkley 83)

8. What are the MLA conventions regarding punctuation?

In MLA, the following conventions regarding punctuation apply:

- Double quotation marks are used for quotations from other texts
- Commas and periods that come directly after a quotation go inside, not outside the quotations marks. However, if the parenthetical reference comes directly after the quotation, then the comma or period should be placed after the reference.

Eg.

While Beth Fowkes Tobin focuses on "the representation of cultural encounters that occurred in British colonies during the late eighteenth century," she specifically addresses "paintings of colonial officials and colonized places, plants, and peoples" (1).

9. What will my Reference List/Works Cited look like?

All works that you have mentioned throughout your essay must be listed **alphabetically** by surname of author (or originator). They should have **hanging indents**, that is, the first line of an entry should be flush left, and the second and subsequent lines should be indented ½ (or five spaces). The MLA style specifies using **title case capitalization**, i.e. capitalize the first words and all principal words, including those that follow hyphens in compound terms. Separate author, title, and publication information with a period (.) followed by one space. Use a colon (:) and a space to separate a title from a subtitle.

Some other important points to remember:

- Only include works to which you have actually referred in the essay.
- The main title of the document should be distinguishable.
- The date is the year of publication *not* printing.
- For a book the edition is only mentioned if other than the first.
- The place of publication is the city *not* the country.
- Journal titles should be given in full.
- Volume and part numbers should be written: 25.2.
- Page numbers should be written: 33-9, 44-67.
- Capitalize all words, except articles (“the” and “a”), (short) prepositions (e.g. “of” “on”, “in”, “into”, “at”, “up”), and other “small” words (“and”, “if”, “it”) when they are not at the beginning of the title (or subtitle).

Nb.

With book, chapter and article titles in **Dutch**, **French**, **Spanish**, and most other non-English languages, only the first word is capitalized. With titles in **German** and **Luxembourgish**, all nouns in these titles are capitalized according to their writing system. The title of a journal is always capitalized according to the title case capitalization.

Eg.

Ashcroft, Bill, Gareth Griffiths and Helen Tiffin, eds. *The Post-Colonial Studies Reader*. London and New York: Routledge, 1995.

Benshoff, Harry, and Sean Griffin. *America on Film: Representing Race, Class, Gender, and Sexuality at the Movies*. Malden, Oxford and Carlton: Blackwell Publishing, 2006.

Gates, Philippa. “Always a Partner in Crime: Black Masculinity in the Hollywood Detective Film.” *Journal of Popular Culture* 32.1 (Spring 2004): 20-9.

Lott, Tommy. “Hollywood and Independent Black Cinema.” Eds. Steve Neale and Murray Smith. *Contemporary Hollywood Cinema*. London and New York: Routledge. 1998. 211-28.

Miller, Toby, et al. *Global Hollywood*. Berkeley: University of California Press, 2002.

Oostindie, Gert. “Caraïbische dilemma’s.” *Het paradijs overzee: de ‘Nederlandse’ Caraïben en Nederland*. Amsterdam: Uitgeverij Bert Bakker, 1997. 277-304.

Richardson, Michael. *Otherness in Hollywood Cinema*. New York and London: Continuum, 2010.

Shaw, Andrea. “Tia Dalma’s Portrayal in *Pirates of the Caribbean*.” *Ordinary Anointments*. 10 June 2011. 11 August 2011. <http://blogs.jamaicans.com/ordinarya/>.

10. What do I do if publication details are not given?

Occasionally you will come across documents that lack basic publication details. In these cases it is necessary to indicate to your reader that these are not available. A series of abbreviations can be used:

- (corporate) author not given	use the title of the work
- no page numbers	use n. pag in place of the page numbers
- no date	use n.d.
- no place of publication	use n.p. before the colon
- no publisher	use n.p. after the colon
- not known	use n.k.

Eg.

n.p: University of Gotham, 1993.	no place of publication
New York: n.p., 1993.	no publisher

11. What do I do if a material type is not covered in this guide?

When you want to make reference to a material type that is not covered in this guide, you can always search online to try to find the way in which you have to include it in your reference list.

The official website of the MLA style can be found at http://www.mla.org. In addition, you could simply go to an online search engine (e.g. Google) and type ‘MLA’ followed by the type of material or format exception that you are looking for, e.g. ‘MLA painting’. You will probably find different ways to list the material in your reference list – just keep in mind: be consistent in everything you do!

12. Why is this guide written in English?

This guide is written in English for two main reasons. First of all, the English language is the common language of science, and MLA is a style rendered in English. The *MLA Handbook for Writers of Research Papers* is originally published in English and when you search online for a material type that is not covered in this guide (see question 11) you will mainly find examples in English. Secondly, this guide is used for all our courses at the Department of Media Studies and these courses are increasingly offered in the English language and followed by non-Dutch students. By having the guide in English, all courses and students are able to use it.

13. What do I do if I am writing my essay in Dutch?

When you write your essay in Dutch, you maintain the MLA style and only translate the necessary details of the references, particularly the abbreviations:

• Editor = Ed.	Redacteur = Red.
• Editors = Eds.	Redacteurs = Reds.
• no date = n.d.	zonder datum = z.d.
• no place = n.p.	zonder plaats = z.p.
• no publisher = n.p.	zonder uitgeverij = z.u.
• no pagination = n. pag.	zonder paginering = z. pag.
• not known = n.k.	niet bekend = n.b.
• quoted in = qtd. in	Geciteerd in = gecit. In
• Diss. (Ph.D Thesis)	Diss. (proefschrift)
• Unpublished Diss. (Ph.D. Thesis)	Ongepubliceerde diss. (proefschrift)
• Plaatsnamen = London	Londen

- 'en' instead of 'and' when listing the authors, editors or places of publication.

Nb.

With book, chapter and article **titles in Dutch**, only the first word is capitalized. The title of a journal is always capitalized according to the title case capitalization.

Eg.

Hermes, Joke, en Maarten Reesink. *Inleiding televisiestudies*. Amsterdam: Boom, 2003.

Oostindie, Gert. "Caraïbische dilemma's." *Het paradijs overzee: de 'Nederlandse' Caraïben en Nederland*. Amsterdam: Uitgeverij Bert Bakker, 1997. 277-304.

This MLA Referencing Guide is condensed and adapted from:
MLA. *MLA Handbook for Writers of Research Papers.* 7th ed. New York: MLA 2009.

1. Virtual methods & digital methods

B. Wellman (2010). “Studying the Internet Through the Ages,” in M. Consalvo and C. Ess (eds.), *The Handbook of Internet Studies.* Wiley-Blackwell: 17-23.

1

Studying the Internet Through the Ages

Barry Wellman

Pre-History

As a tribal elder, I often think back to the state of Internet and society scholarship before the dawning of the Internet. Although sociologist Roxanne Hiltz and computer scientist Murray Turoff had published their prophetic *Network Nation* in 1978, linking social science with computerized communication, the word "Internet" hadn't been invented.

As one of the first social scientists to be involved in research studying how people communicate online, I started going in 1990 to biannual gatherings of the then-tribe: CSCW (computer supported cooperative work) conferences that were dominated by computer scientists writing "groupware" applications. Lotus Notes applications were in vogue. Lab studies were the predominant research method of choice, summarized in Lee Sproull and Sara Kiesler's *Connections* (1991).

But all that people wanted to deal with were small closed groups. I remember standing lonely and forlorn at the microphone during a comments period at the CSCW 1992 conference. Feeling extremely frustrated (and now prophetic), I exclaimed:

> You don't understand! The future is not in writing stand-alone applications for small groups. It is in understanding that computer networks support the kinds of social networks in which people usually live and often work. These social networks are not the densely-knit, isolated small groups that groupware tries to support. They are sparsely-knit, far-reaching networks, in which people relate to shifting relationships and communities. Moreover, people don't just relate to each other online, they incorporate their computer mediated communication into their full range of interaction: in-person, phone, fax, and even writing.

I pleaded for paying more attention to how people actually communicate in real life. But this approach was disparagingly referred to as "user studies," much less exciting to computer geeks than writing new applications. Conference participants listened politely and went back to developing applications for small

groups. I helped develop one too, for it was exciting and fun to collaborate with computer scientists and be one of the few sociologists who actually built stuff. Maybe, we'd get rich and famous. Our Cavecat/Telepresence desktop videoconferencing systems were stand-alone groupware at their then-finest (Mantei et al., 1991; Buxton, 1992). But, they never got out of the laboratory as our grant ran out and they were expensive to hardwire in those pre-Internet days. Little did we realize that Cisco would appropriate our Telepresence name as a trademark 15 years later, without so much as a hand-wave.

The First Age of Internet Studies: Punditry Rides Rampant

Economic forces were already fueling the turn away from stand-alone groupware towards applications that supported social networks. This was the proliferation of the Internet as it became more than an academic chat room. Unlike groupware, the Internet was open-ended, far-flung, and seemingly infinite in scope. The Internet became dot.com'ed, and the boom was on by the mid-1990s.

The Internet was seen as a bright light shining above everyday concerns. It was a technological marvel, thought to be bringing a new Enlightenment to transform the world. Communication dominated the Internet, by asynchronous email and discussion lists and by synchronous instant messaging and chat groups. All were supposedly connected to all, without boundaries of time and space. As John Perry Barlow, a leader of the Electric Frontier Foundation, wrote in 1995:

> With the development of the Internet, and with the increasing pervasiveness of communication between networked computers, we are in the middle of the most transforming technological event since the capture of fire. I used to think that it was just the biggest thing since Gutenberg, but now I think you have to go back farther (p. 56).

In their euphoria, many analysts lost their perspective and succumbed to presentism and parochialism. Like Barlow, they thought that the world had started anew with the Internet (*presentism*). They had gone beyond groupware, and realized that computer-mediated communication – in the guise of the Internet – fostered widespread connectivity. But like the groupware folks, they looked at online phenomena in isolation (*parochialism*). They assumed that only things that happened on the Internet were relevant to understanding the Internet. Their initial analyses of the impact of the Internet were often unsullied by data and informed only by conjecture and anecdotal evidence: travelers' tales from Internet *incognita*. The analyses were often utopian: extolling the Internet as egalitarian and globe-spanning, and ignoring how differences in power and status might affect interactions on and offline. The dystopians had their say too, worrying that "while all this razzle-dazzle connects us electronically, it disconnects us from each other,

having us 'interfacing' more with computers and TV screens than looking in the face of our fellow human beings" (Texas broadcaster Jim Hightower, quoted in Fox, 1995, p. 12).

Pundits and computer scientists alike were still trying to get a handle on what was happening without taking much account of social science knowledge. In my frustration, I began to issue manifestos in the guise of scholarly articles. Two presented my case, based on my 30-plus years of experience as a social network analyst and community analyst. "An Electronic Group is Virtually a Social Network" (1997) contrasted groups and groupware with social networks and social networkware. It asserted that the Internet was best seen as a computer-supported social network, in fact the world's largest component (a network in which all points are ultimately connected, directly or indirectly). The second paper, "Net Surfers Don't Ride Alone" (with Milena Gulia, 1999) took aim at the vogue for calling every interaction online a "community." It argued that the Internet was not the coming of the new millennium, despite the gospel of *Wired* magazine (then the *Vogue* magazine of the Internet), but was a new technology following the path of other promoters of transportation and communication connectivity, such as the telegraph, railroad, telephone, automobile, and airplane. It showed how community dynamics continued to operate on the Internet – this was not a totally new world – and how intertwined offline relationships were with online relationships.

The Second Age of Internet Studies: Systematic Documentation of Users and Uses

The second age of Internet studies began about 1998 when government policy-makers, commercial interests, and academics started to want systematic accounts of the Internet. They realized that if the Internet boom were to continue, it would be good to describe it rather than just to praise it and coast on it. But the flames of Internet euphoria dimmed with the collapse of the dot.com boom early in 2000. The pages of *Wired* magazine shrank 25 percent from 240 pages in September 1996 to 180 pages in September 2001, and then shrank another 17 percent to 148 pages in September 2003: a decline of 38 percent since 1996.

Moreover, the uses of the Internet kept expanding and democratizing. The initial killer applications of communication – variants of email and instant messaging – were joined by information, via the Netscape/Internet Explorer enabled World Wide Web. Search engines, such as Alta Vista and then Google moved web exploring beyond a cognoscenti's game of memorizing arcane URLs and IP addresses. What exactly was going on, besides the hype of Internet promotion by the mass media, governments, NGOs, entrepreneurs, and academics going for suddenly available grants?

The Internet opened our field up way beyond small-group studies. The second age of Internet studies was devoted to documenting this proliferation of Internet

users and uses. It was based on large-scale surveys, originally done by marketing-oriented firms (and with some bias towards hyping use), but increasingly done by governments, academics, and long-term enterprises such as the Pew Internet and American Life Study (www.pewinternet.org) and the World Internet Project (www.worldinternetproject.net). These studies counted the number of Internet users, compared demographic differences, and learned what basic things people have been doing on the Internet. For example, we came to know that a majority of adults in many developed countries have used the Internet, and women were rapidly increasing their presence. However, we discovered that the socioeconomic gap persists in most countries even with increasing use, because poorer folks are not increasing their rate of use as much as wealthier, better-educated ones (Chen & Wellman, 2005).

Neither the utopian hopes of Barlow nor the dystopian fears of Hightower have been borne out. Despite Barlow's hopes, the Internet has not brought a utopia of widespread global communication and democracy. Despite Hightower's fears, high levels of Internet use have not lured people away from in-person contact. To the contrary, it seems as if the more people use the Internet, the more they see each other in person (distance permitting) and talk on the telephone (see the studies in Wellman & Haythornthwaite, 2002). This may be because the Internet helps arrange in-person meetings and helps maintain relationships in between meetings (Haythornthwaite & Wellman, 1998). It may also mean that gregarious, extroverted people will seize on all media available to communicate (Kraut et al., 2002).

To the surprise of some, the purportedly global village of the Internet has not even destroyed in-person neighboring. In "Netville," a suburb near Toronto, the two-thirds of the residents who had always-on, super-fast Internet access knew the names of three times as many neighbors as their unwired counterparts, spoke with twice as many, and visited in the homes of 1.5 as many (Hampton & Wellman, 2003). Given opportunities to organize, people will often connect with those who live nearby, online as well as offline (Hampton, 2007).

Yet, the globe-spanning properties of the Internet are obviously real, nowhere more so than in the electronic diasporas that connect émigrés to their homeland. In so doing, they enable diasporas to aggregate and transmit reliable, informal news back to often-censored countries (Miller & Slater, 2000; Mitra, 2003; Mok, Wellman, & Carrasco, 2009).

The Third Age: From Documentation to Analysis

The use of the Internet has kept growing. But, its proliferation has meant that it no longer stands alone, if it ever did. It has become embedded in everyday life. The ethereal light that dazzled from above has become part of everyday things. We have moved from a world of Internet wizards to a world of ordinary people routinely using the Internet. The Internet has become an important thing, but it is not a special thing. It has become the utility of the masses, rather than the

plaything of computer scientists. Rather than explosive growth, the number of Internet users has become steady state in North America, although the types of Internet use have proliferated. Yet, the burgeoning of diverse Web 2.0 applications, from Facebook social-networking software to YouTube home videos, has increased desires to know about which applications to use. Reflecting the routinization of the Internet, *Wired* has moved from its *Vogue*-ish origins to become more of a how-to-do-it magazine. Its length of 160 pages in September 2008 is an 8 percent increase from September 2003, although I wonder how it will withstand the new global recession.

How do scholars engage with the Internet in this third age? The first two ages of Internet studies were easy. In the first age, little large-scale data were used, just eloquent euphoria or despair. In the second age, researchers grabbed low-hanging fruit using standard social scientific methods – surveys and fieldwork – to document the nature of the Internet.

Two opposing – but complementary – trends are now apparent in the third age. One trend is the development of "Internet studies" as a field in its own right, bringing together scholars from the social sciences, humanities, and computer sciences. The annual conference of the Association of Internet Researchers (AoIR) started in 2000, and has become institutionalized in the last few years, so much so that many participants do not realize what a shoestring, hope-filled gathering the first meeting was at the University of Kansas. AoIR quickly became international, with conferences in the Netherlands, Australia, Canada, and Denmark attracting many hundreds. Its AIR list serve is even bigger. For vacation-minded researchers, the Hawaii International Conference on System Science offers a congenial venue. Many journals, often backed by major publishers, focus on the Internet and society, including *Computers in Human Behavior*, *Information, Communication and Society* (which puts out an annual AoIR conference issue), *The Information Society*, the online-only *Journal of Computer Mediated Communication*, *New Media and Society*, and the *Social Science Computing Review*.

The second trend is the incorporation of Internet research into the mainstream conferences and journals of their disciplines, with projects driven by ongoing issues. This brings the more developed theories, methods, and substantive lore of the disciplines into play, although sometimes at the cost of the adventurous innovativeness of interdisciplinary Internet research. I take two examples from my own discipline of sociology.

One phenomenon is the incorporation of the longstanding concern about the "digital divide" into the study of stratification. Moving beyond the second-age counting of which kinds of people are on – or off – line, Eszter Hargittai (2004) has shown the differential distribution of skills – and not just access – in the American population. It is not just getting connected; it is getting usably connected. Put another way, there are non-economic factors of social inequality – linked to skill and cultural capital – that strongly affect the structure of increasingly computerized societies and the life chances of their members (DiMaggio et al., 2004).

A second continuing debate has been about the loss of community first discussed more than a century ago, by Ferdinand Tönnies in 1887. Instead of the

former debate about whether industrialization and urbanization had withered community, research now turned to television (Putnam, 2000) and the Internet (Kraut et al., 1998, 2002). Systematic field research showed that community ties were thriving, with online connectivity intertwined with offline relationships (Wellman & Haythornthwaite, 2002; Boase et al., 2006; Wellman et al., 2006; Wang & Wellman, 2010). For example, our NetLab is currently looking at what kinds of relationships the Internet does (and does not) foster. As an overarching thought, our NetLab believes that the evolving personalization, portability, ubiquitous connectivity, and wireless mobility of the Internet are facilitating a move towards individualized networks (Kennedy et al., 2008). The Internet is helping each person to become a communication and information switchboard, between persons, networks, and institutions.

What of groupware, where I started nearly 20 years ago? It has been transmuted from supporting small closed groups into social-network software that connects dispersed, complex networks of friends and colleagues and helps to connect the hitherto unconnected.

I am not standing alone any more. Groups have clearly become networked individuals: on the Internet and off it (Wellman, 2001, 2002). The person has become the portal.

Note

This is a revised version of an article originally published in *New Media & Society*, 6 (2004), 108–14.

Acknowledgements: My thanks to Cavecat/Telepresence colleagues who first involved me in this area: Ronald Baecker, Bill Buxton, Janet Salaff, and Marilyn Mantei Tremaine. Bernie Hogan and Phuoc Tran are among the NetLab members who have provided useful comments along the way. Intel Research's People and Practices, the Social Science and Humanities Research Council of Canada, and Bell Canada have been the principal supporters of our NetLab research.

References

Barlow, J. P. (1995). Is there a there in cyberspace? *Utne Reader*, March–April, 50–56.

Boase, J., Horrigan, J., Wellman, B., & Rainie, L. (2006). *The Strength of Internet Ties.* Washington: Pew Internet and American Life Project. www.pewinternet.org.

Buxton, B. (1992). Telepresence: Integrating shared task and person spaces. Paper presented at the Proceedings of Graphics Interface, May, Vancouver.

Chen, W., & Wellman, B. (2005). Charting digital divides within and between countries. In W. Dutton, B, Kahin, R. O'Callaghan, & A. Wyckoff (eds.), *Transforming Enterprise* (pp. 467–97). Cambridge, MA: MIT Press.

DiMaggio, P., Hargittai, E., Celeste, C., & Shafter, S. (2004). From unequal access to differentiated use: A literature review and agenda for research on digital inequality. In K. Neckerman (ed.), *Social Inequality.* New York: Russell Sage Foundation.

Fox, R. (1995). Newstrack. *Communications of the ACM*, 38 (8), 11–12.

Hampton, K. (2007). Neighbors in the network society: The e-neighbors study. *Information, Communication, and Society*, 10 (5), 714–48.

Hampton, K., & Wellman, B. (2003). Neighboring in Netville: How the Internet supports community and social capital in a wired suburb. *City & Community*, 2 (3), 277–311.

Hargittai, E. (2004). Internet use and access in context. *New Media & Society*, 6, 137–43.

Haythornthwaite, C., & Wellman, B. (1998). Work, friendship and media use for information exchange in a networked organization. *Journal of the American Society for Information Science*, 49 (12), 1101–14.

Hiltz, S. R., & Turoff, M. (1978). *The Network Nation*. Reading, MA: Addison-Wesley.

Kennedy, T., Smith, A., Wells, A. T., & Wellman, B. (2008). Networked families. Pew Internet and American Life Project, October. http://www.pewinternet.org/PPF/r/266/report_display.asp.

Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V., & Crawford, A. (2002). Internet paradox revisited. *Journal of Social Issues*, 58 (1), 49–74.

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukhopadhyay, T., & Scherlis, W. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist*, 53 (9), 1017–31.

Mantei, M., Baecker, R., Sellen, A., Buxton, W., Milligan, T., & Wellman, B. (1991). Experiences in the use of a media space: Reaching through technology. *Proceedings of the CHI '91 Conference* (pp. 203–8). Reading, MA: Addison-Wesley.

Miller, D., & Slater, D. (2000). *The Internet: An Ethnographic Approach*. Oxford: Berg.

Mitra, A. (2003). Online communities, diasporic. *Encyclopedia of Community*, vol. 3 (pp. 1019–20). Thousand Oaks, CA: Sage.

Mok, D., Wellman, B., & Carrasco, J-A. (2009). Does distance still matter in the age of the Internet? *Urban Studies*, forthcoming.

Putnam, R. (2000). *Bowling Alone*. New York: Simon & Schuster.

Sproull, L., & Kiesler, S. (1991). *Connections*. Cambridge, MA: MIT Press.

Wang, H. H., & Wellman, B. (2010). Social connectivity in America. *American Behavioral Scientist*, Forthcoming, 53.

Wellman, B. (1997). An electronic group is virtually a social network. In S. Kiesler (ed.), *Culture of the Internet* (pp. 179–205). Mahwah, NJ: Lawrence Erlbaum.

Wellman, B. (2001). Physical place and cyberspace: The rise of personalized networks. *International Urban and Regional Research*, 25 (2), 227–52.

Wellman, B. (2002). Little boxes, glocalization, and networked individualism. In M. Tanabe, Peter van den Besselaar, & T. Ishida (eds.), *Digital Cities II: Computational and Sociological Approaches* (pp. 10–25). Berlin: Springer.

Wellman, B., & Gulia, M. (1999). Net surfers don't ride alone: Virtual communities as communities. In B. Wellman (ed.), *Networks in the Global Village* (pp. 331–66). Boulder, CO: Westview.

Wellman, B., & Haythornthwaite, C. (eds.) (2002). *The Internet in Everyday Life*. Oxford: Blackwell.

Wellman, B., & Hogan, B., with Berg, K., Boase, B., Carrasco, J-A., Côté, R., Kayahara, J., Kennedy, T., and Tran, P. (2006). Connected lives: The project. In P. Purcell (ed.), *Networked Neighbourhoods: The Online Community in Context* (pp. 157–211). Guildford, UK: Springer.

D. Lazer et al. (2009). “Life in the network: the coming age of computational social science,” *Science*. 323(5915): 721-723.

Published in final edited form as:
Science. 2009 February 6; 323(5915): 721–723. doi:10.1126/science.1167742.

Life in the network: the coming age of computational social science

David Lazer,
Harvard University

Alex (Sandy) Pentland,
MIT

Lada Adamic,
University of Michigan

Sinan Aral,
NYU

Albert Laszlo Barabasi,
Northeastern University

Devon Brewer,
Interdisciplinary Scientific Research

Nicholas Christakis,
Harvard University

Noshir Contractor,
Northwestern University

James Fowler,
UCSD

Myron Gutmann,
University of Michigan

Tony Jebara,
Columbia University

Gary King,
Harvard University

Michael Macy,
Cornell University

Deb Roy, and
MIT

Marshall Van Alstyne
Boston University

We live life in the network. When we wake up in the morning, we check our e-mail, make a quick phone call, walk outside (our movements captured by a high definition video camera), get on the bus (swiping our RFID mass transit cards) or drive (using a transponder to zip through the tolls). We arrive at the airport, making sure to purchase a sandwich with a credit card before boarding the plane, and check our BlackBerries shortly before takeoff. Or we visit the doctor or the car mechanic, generating digital records of what our medical or automotive problems are. We post blog entries confiding to the world our thoughts and feelings, or maintain personal

social network profiles revealing our friendships and our tastes. Each of these transactions leaves digital breadcrumbs which, when pulled together, offer increasingly comprehensive pictures of both individuals and groups, with the potential of transforming our understanding of our lives, organizations, and societies in a fashion that was barely conceivable just a few years ago.

The capacity to collect and analyze massive amounts of data has unambiguously transformed such fields as biology and physics. The emergence of such a data-driven "computational social science" has been much slower, largely spearheaded by a few intrepid computer scientists, physicists, and social scientists. If one were to look at the leading disciplinary journals in economics, sociology, and political science, there would be minimal evidence of an emerging computational social science engaged in quantitative modeling of these new kinds of digital traces. However, computational social science is occurring, and on a large scale, in places like Google, Yahoo, and the National Security Agency. Computational social science could easily become the almost exclusive domain of private companies and government agencies. Alternatively, there might emerge a "Dead Sea Scrolls" model, with a privileged set of academic researchers sitting on private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest in the accumulation, verification, and dissemination of knowledge.

What potential value might a computational social science, based in an open academic environment, offer society, through an enhanced understanding of individuals and collectives? What are the obstacles that stand in the way of a computational social science?

From individuals to societies

To date the vast majority of existing research on human interactions has relied on one-shot self-reported data on relationships. New technologies, such as video surveillance, e-mail, and 'smart' name badges offer a remarkable, second-by-second picture of interactions over extended periods of time, providing information about both the structure and content of relationships. Consider examples of data collection in this area and of the questions they might address:

Video recording and analysis of the first two years of a child's life (1)

Precisely what kind of interactions with others underlies the development of language? What might be early indicators of autism?

Examination of group interactions through e-mail data

What are the temporal dynamics of human communications—that is, do work groups reach a stasis with little change, or do they dramatically change over time (2,3)? What interaction patterns predict highly productive groups and individuals? Can the diversity of news and content we receive predict our power or performance (4)?

Examination of face-to-face group interactions over time using sociometers

Small electronics packages ('sociometers') worn like a standard ID badge can capture physical proximity, location, movement, and other facets of individual behavior and collective interactions. What are patterns of proximity and communication within an organization, and what flow patterns are associated with high performance at the individual and group levels (5)?

Science. Author manuscript; available in PMC 2009 September 16.

Macro communication patterns

Phone companies have records of call patterns among their customers extending over multiple years, and e-Commerce portals such as Google and Yahoo collect instant messaging data on global communication. Do these data paint a comprehensive picture of societal-level communication patterns? What does the "macro" social network of society look like (6), and how does it evolve over time? In what ways do these interactions affect economic productivity or public health?

Tracking movement

With GPS and related technologies, it is increasingly easy to track the movements of people (7,8). Mobile phones, in particular, allow the large scale tracing of people's movements and physical proximities over time (9), where it may be possible to infer even cognitive relationships, such as friendship, from observed behavior (10). How might a pathogen, such as influenza, driven by physical proximity, spread through a population (11)?

Internet

The Internet offers an entirely different channel for understanding what people are saying, and how they are connecting (12). Consider, for example, in this political season, tracing the spread of arguments/rumors/positions in the blogosphere (13), as well as the behavior of individuals surfing the Internet (14), where the concerns of an electorate become visible in the searches they conduct. Virtual worlds, by their nature capturing a complete record of individual behavior, offer ample opportunities for research, for example, experimentation that would be impossible or unacceptable (15). Similarly, social network websites offer an unprecedented opportunity to understand the impact of a person's structural position on everything from their tastes to their moods to their health (16), while Natural Language Processing offers increased capacity to organize and analyze the vast amounts of text from the Internet and other sources (17).

In short, a computational social science is emerging that leverages the capacity to collect and analyze data with an unprecedented breadth and depth and scale. Substantial barriers, however, might limit progress. Existing ways of conceiving human behavior were developed without access to terabytes of data describing their minute-by-minute interactions and locations of entire populations of individuals. For example, what does existing sociological network theory, built mostly on a foundation of one-time 'snapshot' data, typically with only dozens of people, tell us about massively longitudinal datasets of millions of people, including location, financial transactions, and communications? The answer is clearly "something," but, as with the blind men feeling parts of the elephant, limited perspectives provide only limited insights. These emerging data sets surely must offer some qualitatively new perspectives on collective human behavior.

There are significant barriers to the advancement of a computational social science both in approach and in infrastructure. In terms of approach, the subjects of inquiry in physics and biology present different challenges to observation and intervention. Quarks and cells neither mind when we discover their secrets nor protest if we alter their environments during the discovery process (although, as discussed below, biological research involving humans offers some similar concerns regarding privacy). In terms of infrastructure, the leap from social science to a computational social science is larger than from, say, biology to a computational biology, in large part due to the requirements of distributed monitoring, permission seeking, and encryption. The resources available in the social sciences are significantly smaller, and even the physical (and administrative) distance between social science departments and engineering or computer science departments tends to be greater than for the other sciences. The availability of easy-to-use programs and techniques would greatly magnify the presence

of a computational social science. Just as mass-market CAD software revolutionized the engineering world decades ago, common computational social science analysis tools and the sharing of data will lead to significant advances. The development of these tools can, in part, piggyback on those developed in biology, physics and other fields, but also requires substantial investments in applications customized to social science needs.

Perhaps the thorniest challenges exist on the data side, with respect to access and privacy. Many, though not all, of these data are proprietary (e.g., mobile phone and financial transactional data). The debacle following AOL's public release of "anonymized" search records of many of its customers highlights the potential risk to individuals and corporations in the sharing of personal data by private companies (18). Robust models of collaboration and data sharing between industry and the academy need to be developed that safeguard the privacy of consumers and provide liability protection for corporations.

More generally, properly managing privacy issues is essential. As the recent NRC report on GIS data highlights, it is often possible to pull individual profiles out of even carefully anonymized data (19). To take a non-social science example: this past Summer NIH and the Wellcome Trust abruptly removed a number of genetic databases from online access (20). These databases were seemingly anonymized, simply reporting the aggregate frequency of particular genetic markers. However, research revealed the potential for de-anonymization, based on the statistical power of the sheer quantity of data collected from each individual in the database (21).

A single dramatic incident involving a breach of privacy could produce a set of statutes, rules, and prohibitions that could strangle the nascent field of computational social science in its crib. What is necessary, now, is to produce a self-regulatory regime of procedures, technologies, and rules that reduce this risk but preserve most of the research potential. As a cornerstone of such a self-regulatory regime, Institutional Review Boards (IRBs) must increase their technical knowledge enormously to understand the potential for intrusion and individual harm because new possibilities do not fit their current paradigms for harm. For example, many IRBs today would be poorly equipped to evaluate the possibility that complex data could be de-anonymized. Further, it may be necessary for IRBs to oversee the creation of a secure, centralized data infrastructure. Certainly, the status quo is a recipe for disaster, where existing data sets are scattered among many different groups, with uneven skills and understanding of data security, with widely varying protocols.

Researchers themselves must tackle the privacy issue head on by developing technologies that protect privacy while preserving data essential for research (22). These systems, in turn, may prove useful for industry in managing privacy of customers and security of their proprietary data.

Finally, the emergence of a computational social science shares with other nascent interdisciplinary fields (e.g., sustainability science) the need to develop a paradigm for training new scholars. A key requirement for the emergence of an interdisciplinary area of study is the development of complementary and synergistic explanations spanning different fields and scales. Tenure committees and editorial boards need to understand and reward the effort to publish across disciplines (23). Certainly, in the short run, computational social science needs to be the work of teams of social and computer scientists. In the longer run, the question will be: should academia be building computational social scientists, or teams of computationally literate social scientists and socially literate computer scientists?

The emergence of cognitive science in the 1960s and 1970s offers a powerful model for the development of a computational social science. Cognitive science emerged out of the power of the computational metaphor of the human mind. It has involved fields ranging from

neurobiology to philosophy to computer science. It attracted the investment of substantial resources to establish a common field, and it has created enormous progress for public good in the last generation. We would argue that a computational social science has a similar potential, and is worthy of similar investments.

References

1. Roy, D.; Patel, R.; DeCamp, P.; Kubat, R.; Fleischman, M.; Roy, B.; Mavridis, N.; Tellex, S.; Salata, A.; Guiness, J.; Levit, M.; Gorniak, P. The Human Speechome Project. Twenty-eighth Annual Meeting of the Cognitive Science Society; 2006.
2. Eckmann JP, Moses E, SergI D. Entropy of dialogues creates coherent structures in e-mail traffic. Proceedings of the National Academy of Sciences of the United States of America 2004;101:14333–14337. [PubMed: 15448210]
3. Kossinets G, Watts D. Empirical Analysis of an Evolving Social Network. Science 2006;311(5757): 88–90. [PubMed: 16400149]
4. Aral, S.; Van Alstyne, M. Network Structure & Information Advantage. Proceedings of the Academy of Management Conference; Philadelphia, PA. 2007.
5. Pentland, A. Honest Signals: how they shape our world. MIT Press; Cambridge, MA: 2008.
6. Onnela, J-P.; Saramäki, J.; Hyvönen, J.; Szabó, G.; Lazer, D.; Kaskil, K.; Kertész, J.; Barabási, A-L. Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences of the United States of America; 2007.
7. Shaw, B.; Jebara, T. Minimum Volume Embedding. Proceedings of the Conference on Artificial Intelligence and Statistics; 2007.
8. Jebara, T.; Song, Y.; Thadani, K. Spectral Clustering and Embedding with Hidden Markov Models. Proceedings of the European Conference on Machine Learning; 2007.
9. González MC, Hidalgo CA, Barabási AL. Understanding individual human mobility patterns. Nature 2008;453:779–782. [PubMed: 18528393]
10. Eagle, N.; Pentland, A.; Lazer, D. Inferring friendships from behavioral data. HKS working paper; 2008.
11. Colizza V, Barrat A, Barthelemy M, Vespignani A. Prediction and predictability of global epidemics: the role of the airline transportation network. Proceedings of the National Academy of Sciences of the United States of America 2006;103:2015–2020. [PubMed: 16461461]
12. Watts D. Connections A twenty-first century science. Nature 445:489. [PubMed: 17268455]
13. Adamic, L.; Glance, N. The Political Blogosphere and the 2004 U.S. Election Divided They Blog. LinkKDD-2005; Chicago, IL: 2005.
14. J. Teevan. 2008. “How People Recall, Recognize and Re-Use Search Results,” To appear in ACM Transactions on Information Systems (TOIS) special issue on Keeping, Re-finding, and Sharing Personal Information.
15. Bainbridge W. The scientific research potential of virtual worlds. Science 2007;317(5837):472–476. [PubMed: 17656715]
16. Lewis K, Kaufman J, Gonzalez M, Wimmer A, Christakis N. Tastes, Ties, and Time: A New (Cultural, Multiplex, and Longitudinal) Social Network Dataset Using Facebook.com. Social Networks. 2009in press
17. Gardie C, Wilkerson J. Text annotation for political science research. Journal of Information Technology and Politics 2008;5:1–6.
18. Barbarao M, Zeller T Jr. A Face Is Exposed for AOL Searcher No. 4417749. New York Times. 2006 August 9;
19. National Research Council. Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data. In: Myron, P., editor. Gutmann and Paul Stern. Washington: National Academy Press; 2007.
20. Felch, J. DNA databases blocked from the public. LA Times; August 29. 2008

21. Homer N, Szelinger S, Redman M, Duggan D, Tembe W. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. PLoS Genetics 2008;4(8):e1000167.10.1371/journal.pgen.1000167 [PubMed: 18769715]

22. Backstrom, L.; Dwork, C.; Kleinberg, J. Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography. Proc. 16th Intl. World Wide Web Conference; 2007.

23. Van Alstyne M, Brynjolfsson E. Could the Internet Balkanize Science? Science 1996;274:1479–1480.

24. Image courtesy of Sense Networks.

25. We will supply animation in supporting online materials.

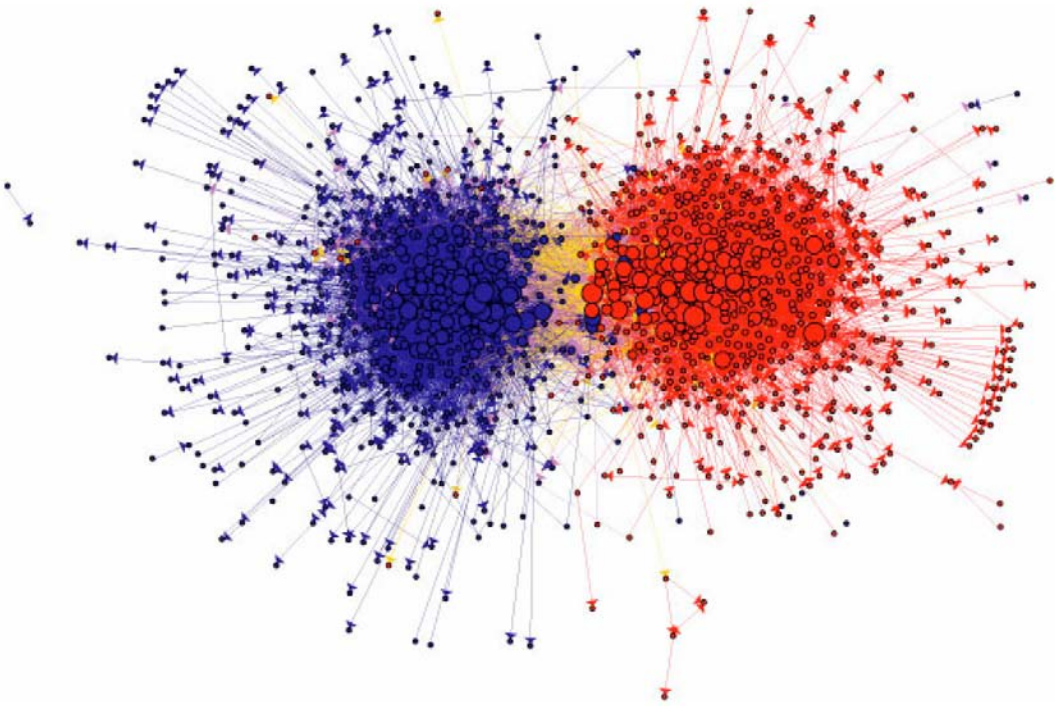


Figure 1.
This figure summarizes the link structure within a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it (10).

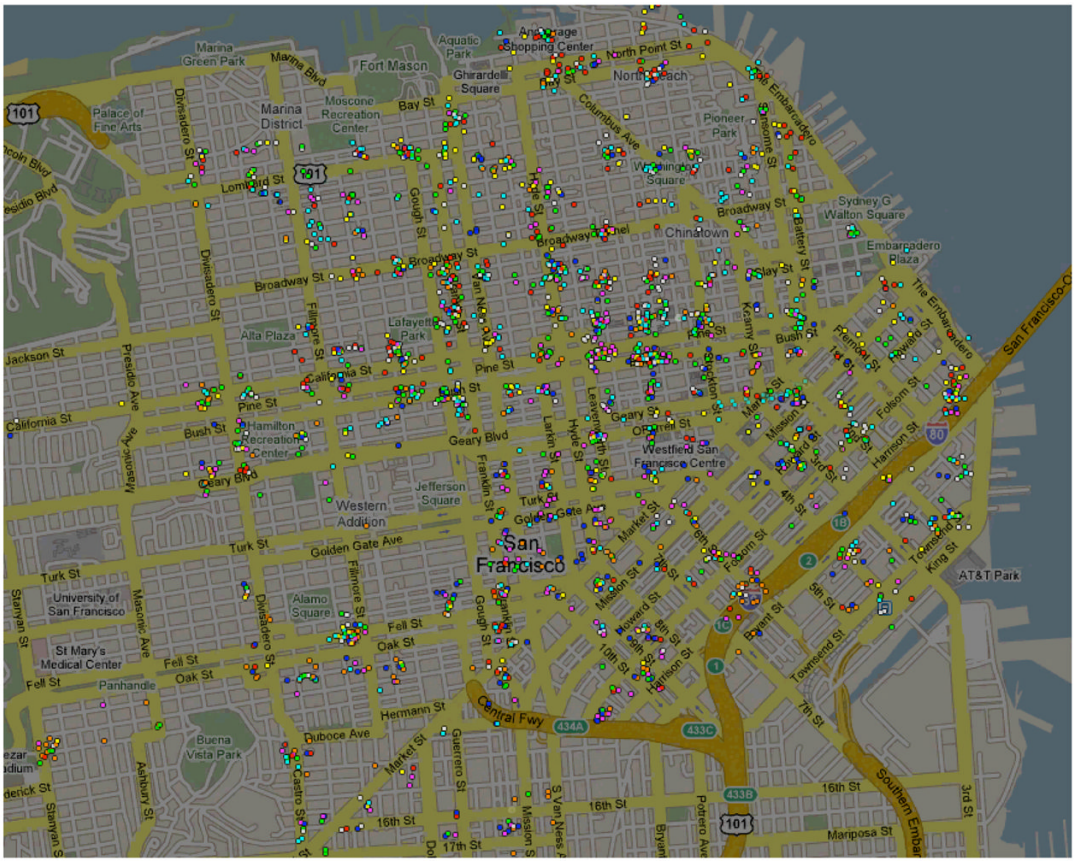


Figure 2.

The location (after adding randomized synthetic noise) of several hundred mobile devices in the city of San Francisco. Each location is color coded to indicate which of 8 “tribes” (or social clusters) each user belongs to. Tribes are computed by clustering (otherwise anonymized) users according to how similar their movement patterns are over a few weeks. The movement analysis is performed using the Minimum Volume Embedding algorithm (7,8,24)

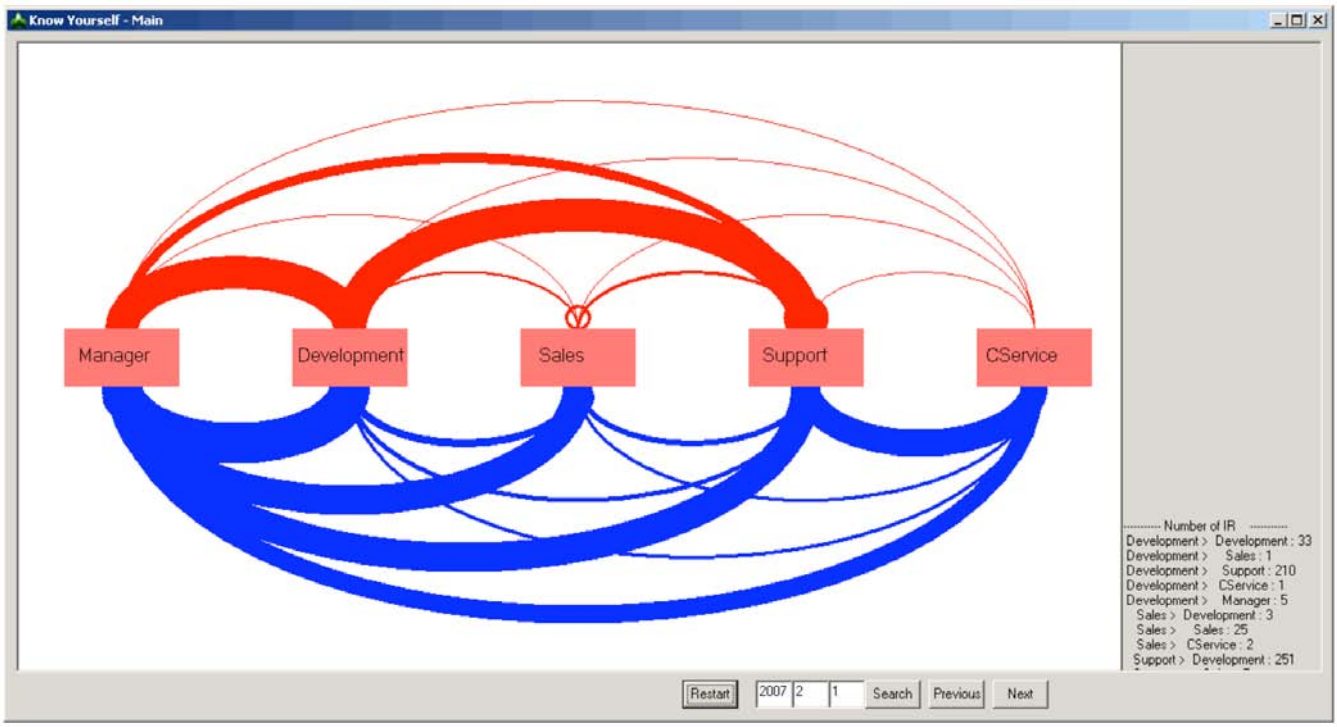


Figure 3.
Patterns of email (blue) and face-to-face communication (read) within a German bank over a period of one month. Productivity and information overload is correlated with the sum of both types of communication, but not with either alone (25)

C. Borgman (2010). “The Digital Future is Now: A Call to Action for the Humanities.” *Digital Humanities Quarterly*. Fall, 3(4).

The Digital Future is Now:

A Call to Action for the Humanities[1]

Christine L. Borgman
Professor & Presidential Chair in Information Studies UCLA
Borgman@gseis.ucla.edu

Revised Final Accepted Version for *Digital Humanities Quarterly*

Table of Contents

[1] Based on Keynote Presentation to Digital Humanities '09 Conference, College Park, MD, June 23, 2009

ABSTRACT

The digital humanities are at a critical moment in the transition from a specialty area to a full-fledged community with a common set of methods, sources of evidence, and infrastructure – all of which are necessary for achieving academic recognition. As budgets are slashed and marginal programs are eliminated in the current economic crisis, only the most articulate and productive will survive. Digital collections are proliferating, but most remain difficult to use, and digital scholarship remains a backwater in most humanities departments with respect to hiring, promotion, and teaching practices. Only the scholars themselves are in a position to move the field forward. Experiences of the sciences in their initiatives for cyberinfrastructure and eScience offer valuable lessons. Information- and data-intensive, distributed, collaborative, and multi-disciplinary research is now the norm in the sciences, while remaining experimental in the humanities. Discussed here are six factors for comparison, selected for their implications for the future of digital scholarship in the humanities: publication practices, data, research methods, collaboration, incentives, and learning. Drawing upon lessons gleaned from these comparisons, humanities scholars are "called to action" with five questions to address as a community: What are data? What are the infrastructure requirements? Where are the social studies of digital humanities? What is the humanities laboratory of the 21st century? What is the value proposition for digital humanities in an era of declining budgets?

INTRODUCTION

This is a pivotal moment for the digital humanities. The community has laid a foundation of research methods, theory, practice, and scholarly conferences and journals. Can we seize this moment to make digital scholarship a leading force in humanities research? Or will the community fall behind, not-quite-there, among the many victims of the massive restructuring of higher education in the current economic crisis? Much is at stake in the community's ability to argue for the value of digital humanities scholarship and to assemble the necessary resources for the field to move from "emergent" to "established."

The sciences, arts, and humanities have converged and diverged in various ways over the centuries. In the area of digital scholarship, many interests are in common across the disciplines. It is the pace of adoption that is divergent. The sciences, and to a lesser extent the social sciences, have been successful in developing the technical, social, and political infrastructure for digital scholarship under the rubrics of *cyberinfrastucture* – the term used in the U.S., and *eScience* – the term more widely used in the U.K. and elsewhere (U.K. Research Council e-Science Programme, 2009; Atkins et al., 2003). Digital scholarship remains emergent in the humanities, while eScience has become the norm in the sciences. The humanities need not emulate the sciences, but can learn useful

lessons by studying the successes (and limitations) of cyberinfrastructure and eScience initiatives.

While leaving definitions of "the humanities" to the reader, two complementary definitions of "digital humanities" provide a useful scope statement. Frischer's definition (2009, p. 15) is "the application of information technology as an aid to fulfill the humanities' basic tasks of preserving, reconstructing, transmitting, and interpreting the human record." One resulting from the UCLA Mellon seminar claims that "Digital humanities is not a unified field but an array of convergent practices that explore a universe in which print is no longer the exclusive or the normative medium in which knowledge is produced and/or disseminated" (Digital Humanities Manifesto, 2009). Taken together, the digital humanities is a new set of practices, using new sets of technologies, to address research problems of the discipline.

PROBLEM STATEMENT

Interest in the digital humanities has grown steadily for several decades. The Digital Humanities Conferences have occurred annually since 1989, sponsored by the Alliance of Digital Humanities Organizations. Constituent organizations of the Alliance have held conferences since 1973 (Alliance of Digital Humanities Organizations, 2009). MITH (Maryland Institute for Technology in the Humanities, 2009) celebrated its tenth anniversary, and IATH (Institute for Advanced Technology in the Humanities, 2009) at the University of Virginia its 17th anniversary. Academic research in the digital humanities at UCLA, Duke, Stanford, King's College London, and elsewhere also appears to be thriving. Funding continues apace, with the Mellon Foundation, Council on Library and Information Resources, National Endowment for the Humanities, U.K. Arts and Humanities Research Council, and others focusing on infrastructure, tools, and services to support humanities scholarship in digital environments. Yet digital scholarship remains a backwater in much of the humanities. Concerns about publishing, tenure, and promotion for digital humanities scholars are a continuing theme in the conferences and in the literature of the field (Friedlander, 2008; 2009; Unsworth et al., 2006).

Despite many investments and years of development, basic infrastructure for the digital humanities is still lacking. Those who wish to gather and analyze digital data for humanities problems often find the overhead daunting, as exemplified by this emailed complaint from a history student in my scholarly communication course, who is pursuing a doctoral dissertation about the German enlightenment:

> I'm finding that something as simple as constructing my maps of related concepts are not easily applied to primary sources in digital libraries. *So what use are the digital libraries, if all they do is put digitally unusable information on the web?* The digital libraries don't offer a platform for traditional note taking, much less for larger scale analysis, either quantitative or qualitative. (emphasis added; quoted with permission)

"Digital libraries," the term used by my student, usually implies the existence of tools, services, and a library imprimatur of cataloging and curation. Her complaint is more about digital collections, which often lack basic capabilities for retrieval or analysis. This distinction is particularly relevant to the digital humanities. Content in digital collections may be "relatively raw," as Lynch (2002) puts it; others can add layers of interpretation, presentation, tools, and services, but these layers may be maintained separately from the content (Borgman, 1999; 2000; Lynch, 2002). The invisibility of essential infrastructure for digital scholarship in the humanities is but one of many challenges to be addressed in growing the field. Until analytical tools and services are more sophisticated, robust, transparent, and easy to use for the motivated humanities researcher, it will be difficult to attract a broad base of interest within the humanities community.

Whose problem is it to improve the situation; that is, to design, develop, and deploy the scholarly infrastructure for digital humanities? As my UCLA colleague, Johanna Drucker, put it so well, "Leaving it to 'them' is unfair, wrongheaded, and irresponsible. Them is us." (Drucker, 2009, p. B8). She believes that the digital humanities are at a "critical juncture," and is concerned that her fellow scholars are deferring responsibility for action to librarians, computer scientists, technology developers, publishers, and others.

The operant terms in "digital humanities scholarship" are the latter two. Scholarly methods are as deeply seated in the humanities as they are in the sciences (Borgman, 2007). Only those who do the work and who require the infrastructure are in a position to take the field forward. Librarians and technology developers are essential partners, but those who conduct the research must take the lead.

This article, based on a keynote presentation to the most recent Digital Humanities Conference, reviews and reflects upon the differences between the approaches of the sciences and the humanities to digital scholarship (Borgman, 2009). First, I frame the notion of scholarly information infrastructure, then compare the approaches to digital scholarship of the sciences and the humanities. My analysis concludes with a call to action for the humanities community.

SCHOLARLY INFORMATION INFRASTRUCTURE

The term "scholarly information infrastructure" encompasses the technology, services, practices, and policy that support research in all disciplines. Cyberinfrastructure and eScience – both coined initially in reference to the sciences and technology, and both now used more broadly – refer to an infrastructure that enables forms of scholarship that are information- and data-intensive, distributed, collaborative, and multi-disciplinary. eResearch has become the collective term for variants such as eScience, eSocial Science, and eHumanities (Borgman, 2007). The report of the *Commission on Cyberinfrastructure for the Humanities and Social Sciences* (Unsworth et al., 2006) was modeled on the

strategy for science and technology (Atkins et al., 2003), while diverging to emphasize the humanities' motivations to make cultural heritage more widely available for teaching, research, and outreach. A similar argument is made by Todd Presner and Chris Johanson (2009) that digital humanities offers the opportunity to reconceptualize society as our cultural heritage migrates to digital formats, thus altering our relationship to knowledge and culture.

The technical and policy infrastructure for scholarship is being built rapidly, particularly for the sciences (Cyberinfrastructure Vision for 21st Century Discovery, 2007; Hey, Tansley & Tolle, 2009). Rare are the encompassing visions for scholarly infrastructure that originate in the humanities. Amy Friedlander (2009) provides a notable exception. She identified four research areas in digital scholarship where the interests of humanists, technology researchers, and others converge. These are scale, language and communication, space and time, and social networking. Issues of scale are of general interest because methods and problems must be approached much differently when one has, for example, the full text of a million books rather than a handful. Inspection is no longer feasible; only computational methods can examine corpora on that scale. Issues of language and communication, which are central to the humanities, are of broader interest for problems such as pattern detection and cross-language indexing and retrieval. Space and time encompass the new research methods possible with geographic information systems, geo-tagged documents and images, and the increased ability to make temporal comparisons. Social network analysis, long popular in sociology and bibliometrics, has become generalized to include patterns of social relationships in older texts or in online communication. Cross-cutting agendas such as these can be very influential in the design of an encompassing infrastructure. Humanities researchers need to be at the table as fundamental infrastructure decisions are being made.

SCIENCE [AND, OR, VERSUS] THE HUMANITIES

The humanities and the sciences each encompass broad swaths of scholarship, with much internal diversity. These two communities have significant commonalities, while differing in important ways. Identified here are six factors for comparison, selected for their implications for the future of digital scholarship in the humanities: publication practices, data, research methods, collaboration, incentives, and learning. The first five of these are drawn from longer analyses published elsewhere (Borgman, 2007); the last is drawn from the *NSF Task Force on Cyberlearning* (Borgman et al., 2008). The sequence of topics is cumulative to reflect how the boundaries are blurring between the sciences and the humanities.

Publication practices

Scholarly journal publication is shifting rapidly toward electronic formats, especially in the sciences. Some journals are dropping print publication altogether, others are declaring the online version (usually released several weeks to several months prior to the printed edition) to be the edition of record. Under pressure from authors, the majority of

scholarly journals now appear to allow online posting of some form of pre-print or post-print (SHERPA/RoMEO: Publisher copyright policies & self-archiving, 2009).

For physics and related areas of computer science and mathematics, arXiv is the locus of scholarly communication. Monthly deposits of new papers now number more than 5,000; the site, which contains over 500,000 papers, typically receives 50,000 visits per hour (ArXiv.org e-Print archive, 2009). At least three iPhone applications are available for arXiv searching and retrieval. ArXiv, similar repositories in fields such as economics, and institutional repositories such as ePrints, employ standard data structures that make their contents readily discoverable by search engines (Open Archives Initiative Protocol for Metadata Harvesting, 2009; Research Papers in Economics, 2009; EPrints, 2010). It is little wonder that our science colleagues claim they never go to their campus libraries any more; their libraries come to them.

In the humanities, neither journal nor book publishing has moved rapidly toward online publication, despite pioneering efforts such as the 1990 launch of the *Journal of Post Modern Culture* as an electronic-only journal and the 2005 launch of *Vectors* as an online-only multi-media journal (Journal of Post Modern Culture, 2000; Vectors: Journal of Culture and Technology in a Dynamic Vernacular, 2009; Hamma, 2009; King et al., 2006; Whalen, 2009). A few of the established humanities journals have begun online versions that take advantage of digital technologies. Beginning in March, 2010, for example, the *JSAH* (Journal of the Society of Architectural Historians, 2009) will publish an online version that will support "zoomable images, video, GIS map integration, Adobe Flash VR, 3-D models, and online reference linking" – while continuing to publish its static print version.

The reasons for the slow adoption of digital publishing in the humanities are many, from not trusting online dissemination to a general reluctance to experiment with new technologies, even those well proven – "professionally indisposed to change" as Ken Hamma (2009) puts it. Monographic publishing, which is core to humanities scholarship, has begun a seismic shift toward digital publishing (Jaschik, 2008; 2009; Poe, 2001; Willinsky, 2006; 2009). A growing number of university presses are offering online access to monographs they publish in print, whether or not they also offer digital-only or print-on-demand formats. Other university presses are reinventing themselves in digital form (Rice University Press Mission Statement, 2008). The University of California Press recently announced a partnership with the California Digital Library, which hosts the university's institutional repository, to offer "a suite of publishing services robust and flexible enough to support the complexities of content, format, and dissemination that increasingly define scholarly communications" (University of California Publishing Services, 2009).

The "love affair with print" (Whalen, 2009) of art historians and other humanities scholars places not only "traditional" humanities scholarship at risk but also that of digital humanities. The distinction between print and digital publication is as much about epistemology as genre. Digital publishing is not simply repackaging a book or article as a computer file, although even a searchable pdf has advantages over paper. By

incorporating dynamic multi-media or hypermedia, digital publishing offers different ways of expressing ideas and of presenting evidence for those ideas (Lynch, 2002; Presner, 2010, forthcoming; Presner & Johanson, 2009). When digital scholarship is published in print venues, much of its sophistication is lost.

Digital publishing differs from print publishing in several ways. One is the shorter time from submission to publication. While speed of publication is a much greater concern in the sciences than in the humanities, much of that time delay involves the physical production of the journal or book. Reviewing time varies little between print and digital formats. The humanities could benefit from faster turnaround, reaching audiences much sooner.

A second advantage of digital publishing – even more critical – is the larger audience for online publications. Anyone with an online connection and a subscription (in the case of fee-paid content), anywhere in the world, can read digital publications. Only those with access to a physical copy can read print-only publications. The number of titles and the number of copies of scholarly books and journals published in print form are decreasing rapidly, thus limiting both publishing outlets and readership. Maureen Whalen's (Whalen, 2009) concern for art history, with its continuing reliance on print publishing, is that "the voices of authority … will be talking amongst themselves."

Two other consequences of the inexorable shift toward digital publication should be of concern to the humanities. One is that print material – including older material – becomes "widowed" as students and scholars alike search only online. The widowing problem was recognized early in the days of online catalogs, and was a major impetus for research libraries to digitize their entire back catalogs rather than only records of new material (Borgman, 2000; Lynch, 2003; Lynch & Garcia-Molina, 1995).

The other consequence is that easier access to online material frequently increases its rate of citation. Articles published in open access journals, open repositories, or dual-published by providing preprints or postprints online, tend to have a citation advantage over articles published only in closed-access journals, whether print or online. The degree of advantage varies by field and by a number of other factors, including how "open access" is defined (The Facts about Open Access: A Study of the Financial and Non-Financial Effects of Alternative Business Models on Scholarly Journals, 2005; Directory of Open Access Repositories, 2008; Directory of Open Access Journals, 2009; Open Content Alliance, 2009; The effect of open access and downloads ('hits') on citation impact: a bibliography of studies, 2009; Bailey, 2005).

While the details of these studies are much contested between authors, editors, librarians, and publishers, the simple tautology that easier discovery is associated with higher citation is difficult to dispute. As do authors in other fields, scholars in the humanities desire recognition in the form of citations to their work. Universities consider citation metrics in hiring and promotion decisions, despite known problems in their use for evaluating scholarly productivity (Bollen & Van de Sompel, 2008; Kurtz & Bollen, 2010; Monastersky, 2005; Reedijk & Moed, 2008).

In sum, the sciences have benefited from online publication in ways that the humanities have not (yet). Digital publication is faster, reaches a wider audience, and tends to increase the citation rate over print-only publication. As the proportion of print-only publication continues to decrease, those for whom it is their only venue risk reaching an ever smaller and more closed community with their scholarship. Curation of digital objects is a concern in all fields, and is a topic that has the attention of management in libraries and archives. Nonetheless, digital publication has become the norm, and those who cling to print publication as the only acceptable format for promotion and tenure may be left out of the academic mainstream.

Data in digital scholarship

Central to the notion of cyberinfrastructure and eScience is that "data" have become essential scholarly objects to be captured, mined, used, and reused. This trend has been under way in science for many years, to varying degrees by field. As the technical and communications infrastructure became sufficiently robust to support large-scale data analysis and exchange, data became more valuable commodities. The availability of large volumes of data has enabled scientists to ask new questions, in new ways. Environmental scientists can conduct longitudinal analyses and make comparisons between locales using datasets compiled from multiple sources. Similarly, genome data offer analytical power at much finer granularity, and at larger scales.

While "data" is less familiar terminology in the humanities, the availability of large text, image, audio, and multi-media corpora has a similar result, enabling scholars in multiple fields to interrogate sources in new ways (Crane, Babeu & Bamman, 2007). Judging by presentations at the 2009 Digital Humanities Conference, *data* is becoming a popular term, whether framed in terms of "mining" or "cultural analytics." Data mining "is the process of identifying patterns in large sets of data . . . to uncover previously unknown, useful knowledge." (National Centre for Text Mining, 2009). Cultural analytics is a term that arose in the humanities as an analog to "visual analytics," "business analytics," and "web analytics," and includes the use of "computer-based techniques for quantitative analysis and interactive visualization" to identify patterns in large cultural data sets (Manovich, 2009).

What are data?

The increasing value of data begs the question of "what are data?" Definitions associated with archival information systems offer a useful starting point: "A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen" (*Reference Model for an Open Archival Information System*, 2002, 1-9).

Another way to think about data is by origin. In the context of cyberinfrastructure, the four categories of data identified in an influential U.S. policy report (*Long-Lived Digital Data Collections*, 2005), and incorporated in National Science Foundation strategy (*Cyberinfrastructure Vision for 21st Century Discovery*, 2007), are now widely accepted. *Observational* data include weather measurements and attitude surveys, either of which may be associated with specific places and times or may involve multiple places and times (e.g., cross-sectional, longitudinal studies). *Computational* data result from executing a computer model or simulation, whether for physics or cultural virtual reality. Replicating the model or simulation in the future may require extensive documentation of the hardware, software, and input data. In some cases, only the output of the model might be preserved. *Experimental* data include results from laboratory studies such as measurements of chemical reactions or from field experiments such as controlled behavioral studies. Whether sufficient data and documentation to reproduce the experiment are kept varies by the cost and reproducibility of the experiment. *Records* of government, business, and public and private life also yield useful data for scientific, social scientific, and humanistic research.

Data as evidence

The need to address categories and levels of data is a pragmatic concern for managing information. Yet data are often in the eye of the beholder. In Buckland's terms, data are "alleged evidence" (Buckland, 1991; Edwards, Jackson, Bowker & Knobel, 2007). What counts as good data varies widely, as one person's noise is often another person's signal. Similarly, the choices of data depend heavily on the questions being asked (Scheiner, 2004).

Whether any given set of observation or records can be considered data depends on context, even in the sciences. In our research on science and technology researchers in the environmental sciences, we found differing views of data on concepts as basic as temperature. Some of the computer science and engineering researchers interviewed said roughly, *"temperature is temperature,"* whereas biologists gave much more nuanced descriptions of how temperature was measured: *"There are hundreds of ways to measure temperature. 'The temperature is 98' is low-value compared to, 'the temperature of the surface, measured by the infrared thermopile, model number XYZ, is 98.' That means it is measuring a proxy for a temperature, rather than being in contact with a probe, and it is measuring from a distance. The accuracy is plus or minus .05 of a degree. I [also] want to know that it was taken outside versus inside a controlled environment, how long it had been in place, and the last time it was calibrated, which might tell me whether it has drifted…"* (Borgman, Wallis, Mayernik & Pepe, 2007). Thus these two groups of researchers, often working side-by-side in the field as collaborators, had very different perspectives on what were acceptable data for their evidentiary purposes.

Studies of scientific practice, such as our work in embedded sensor networks, is providing insights for the design of cyberinfrastructure and eScience. The social studies of science and technology is a large and burgeoning field, with multiple journals and book series, and a scholarly society established more than 40 years ago (Van House,

2004). No comparable body of research on scholarly practices in the humanities exists, with the exception of research on information-seeking behavior (Anderson, 2004; Bates, 1996a; b; Bates, Wilde & Siegfried, 1993; 1995; Case, 2006; Siegfried, Bates & Wilde, 1993; Stone, 1982; Tibbo, 2003; Wiberley, 2003; Wiberley & Jones, 1994). Lacking an external perspective, humanities scholars need to be particularly attentive to unstated assumptions about their data, sources of evidence, and epistemology. We are only beginning to understand what constitute data in the humanities, let alone how data differ from scholar to scholar and from author to reader. As Allen Renear remarked, "in the humanities, one person's data is another's theory" (personal communication, June 22, 2009).

Data sources

The sciences and humanities differ greatly in their sources of data and the degree of control they have over those data (Borgman, 2007). Scientific data sources vary by discipline, as seen in these few examples:

- Medicine: x-rays
- Chemistry: protein structures
- Astronomy: spectral surveys
- Biology: specimens
- Physics: events, objects
- Ecology: weather, ground water, sensor readings, historical records

Scientists, generally speaking, use data that were created by and for scientific purposes. They usually generate their own data, as in field observations or laboratory studies, or may acquire data from collaborators or other scientists. They may also acquire data from repositories in their field or from government sites, such as records of rainfall or river flow. Scientific documentation such as laboratory and field notebooks is sometimes considered to be data and sometimes metadata.

The social sciences occupy the middle position between the sciences and humanities on a continuum of data sources and control. Those at the scientific end of the scale gather their own observations, whether opinion polls, surveys, interviews, or field studies; build models of human behavior; and conduct experiments in the laboratory or field. Other social scientists rely on records collected by others, such as economic indicators or demographic data from the census. Government and corporate records are often of interest, as are the mass media. A number of important data repositories exist, especially for large social surveys (e.g., Survey Research Center, Institute for Social Research, 2009; Survey Research Center, UC-Berkeley, 2009; UK Data Archive, 2009).

The humanities and arts are the least likely of the disciplines to generate their own data in the forms of observations, models, or experiments. Humanities scholars rely most heavily on records, whether newspapers, photographs, letters, diaries, books, articles; records of birth, death, marriage; records found in churches, courts, schools, and colleges; or maps. Any record of human experience can be a data source to a humanities scholar. Many of those sources are public while others are private. Cultural records may be found

in libraries, archives, museums, or government agencies, under a complex mix of access rules. Some records are embargoed for a century or more. Some may be viewable only on site, whether in print or digital form. Data sources for humanities scholarship are growing in number and in variety, especially as more records are digitized and made available to the public.

Lynch's (2002) dichotomy of raw material vs. interpretation has a number of implications for the digital humanities. Two are of concern here. One is that raw materials are more likely to be curated for the long term than are scholars' interpretations of those materials. It is the nature of the humanities that sources are reinterpreted continually; what is new is the necessity of making explicit decisions about what survives for migration to new systems and formats. Second is the implication for control of intellectual property. Generally speaking, humanities scholars have far less control over the intellectual property rights of their sources – these raw materials – than do scientists, whose data usually are original observations or specimens. Typically, scholars can read, view, and cite cultural records, but often need explicit permission to reproduce them – and frequently need to pay a fee, especially in the case of images, to include them in reports of their research.

Intellectual property constraints on publishing of digital humanities scholarship are much different than those that usually apply in other disciplines. Rights to reproduce material remain closely tied to a print model, specified by number of copies printed and by temporal rules on sale that are irrelevant to online publication. Even cultural institutions as sophisticated as the Getty Trust encounter structural barriers to online publication of humanities scholarship (Whalen, 2009). The policy shift toward data sharing, well under way in the sciences, generally presumes that those who produce the data have the authority to release or deposit them for reuse (OECD Principles and Guidelines for Access to Research Data from Public Funding, 2007; Arzberger et al., 2004).

In sum, "what are data?" is an important question for the humanities. The answer will determine what data are produced, how they are captured, and how they are curated for reuse. Data sharing in the humanities is a complex set of issues – not that they are simple in the sciences – that must be addressed. The humanities community needs a critical mass of digital resources and needs common tools, services, and repositories if they are to move beyond "boutique projects" (Friedlander, 2009) to a solid foundation of theory and method.

Research methods

Questions of "what are data?" are inextricable from the choice of research method. Many of the sciences, especially those "big science" areas that require large scale instrumentation and produce vast volumes of data, are in transition to a data-driven paradigm (Bell, Hey & Szalay, 2009; Foster, 2009). As the analysis, modeling, and merging of data become more central to scientific research, partnerships between scientists and computer scientists are becoming the norm.

An important case example of the changing role of data in science is the Sloan Digital Sky Survey (*Sloan Digital Sky Survey*, 2006), begun in 1992 by Jim Gray, Alex Szalay, and others (Gray et al., 2005; Gray & Szalay, 2002; Szalay, 2008). It was the first major astronomical survey founded on the premise that the resulting data would be openly and freely available, both to the astronomy community and to the public at large. Not only did astronomers mine the Sloan datasets for research purposes – more than 1700 scholarly papers were published – but manifold more users of these data were students and amateur astronomers. Amateurs, whose backyard telescopes could never yield data of such quality, also made important discoveries.

The Sloan Digital Sky Survey is significant for its openness, research productivity, and community engagement, and because it instantiates the "value chain" of scholarship (Borgman, 2007). On the SDSS site, papers are linked to the datasets on which they are based and datasets are linked to papers about them. One can enter the chain from either point and follow the relationships. While the project has ceased collecting new observations, the Sloan data remain available for use and are a canonical experiment in curation of large-scale datasets (Choudhury et al., 2008; Choudhury & Stinson, 2007). Astronomers and computer scientists are now engaged in the next generation project, Panoramic Survey Telescope and Rapid Response System, which is yielding about twenty times as much data as Sloan (PAN-STARRS, 2009).

Humanities scholars are more likely to find their data sources in the library – their traditional laboratory – than in the skies. While the library continues to be more central to scholarship in the humanities than it is to other fields, the characteristics of that relationship are changing. The use of physical space and of library staff has changed radically in the last two decades, largely in response to flat or declining university library budgets. Campus libraries have been consolidated in efforts to minimize the number of public service points to be staffed. Books, journals, and other physical materials have been moved to remote facilities, paged from the stacks upon request. Professional librarians, while a smaller proportion of library staffs, are turning their attention away from collection building – given the budget crises – and toward making the best use of the materials they have. The sciences are placing less demand on the physical library, allowing university libraries to reconfigure their spaces to benefit faculty and students in the humanities. Prime floor space previously devoted to card catalogs, journals, and book stacks is now available for groups to work together with physical and digital resources. More librarians have backgrounds in the humanities than in the sciences, and many are eager to partner with humanities scholars in building better tools and services for discovering, interpreting, and using scholarly content.

At most universities today, humanities scholars and students are the primary constituency for physical books, journals, and records. This community also makes the finest distinctions among editions, printings, and other variants – distinctions that are sometimes overlooked in the transition from print to digital form. For general reading, any edition may suffice, and some degradation in image quality may be an acceptable tradeoff for access to large corpora of books and journals. Scholars are much more

dependent on metadata to identify and compare variants, and may require physical copies to examine characteristics of printing and paper, annotations, and other details.

Differences in the methods of using print and digital objects are being thrown into sharp relief by mass digitization projects, most recently by the intense public debate over Google's book-scanning project. Concerns include not only the quality of scanning and of metadata, but the possibility that libraries will discard physical copies of books for which scans are available (UC and the Google Book Settlement: Frequently Asked Questions, 2009; Duguid, 2007; Nunberg, 2009; Samuelson, 2009). Also lost in most of these discussions is the distinction between scanning for search and access purposes (the Google approach) and scanning for preservation purposes, which has higher standards for image quality and for metadata (NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials, 2002; Mass Digitization: Implications for Information Policy, 2006; Greenstein, Ivey, Kenney, Lavoie & Smith, 2004).

Digital humanities projects have yet to achieve the scale of data, audience, or participation as the Sloan Digital Sky Survey. However, several long-lived digital humanities projects have made important contributions to research methods and data quality. Perseus is usually considered the first digital library in the humanities, with planning begun in 1985 and services available by 1987 (Perseus Digital Library, 2009; Crane et al., 2001; Marchionini & Crane, 1994). The initial collections of Perseus cover the history, literature, and culture of the Greco-Roman world. They have since expanded into other areas, and conducted significant research on the classification, management, and use of visual and textual materials (Crane, 2006; Mahoney, 2002; Smith, Mahoney & Crane, 2002). Rome Reborn, begun in 1997 at UCLA, was first concerned with digital library problems such as metadata, organization of historical and architectural periods, and representing relationships between textual sources and visual models (Rome Reborn, 2009; Frischer, 2004; 2009). Now the system exists in multiple manifestations, supports three-dimensional "fly-throughs," audio typical of the time period (including spoken Latin), and gladiator fights in the amphitheater using the latest computer graphics technology. Perseus, Rome Reborn, and newer projects such as HyperCities integrate map layers from Google Earth and other sources, which broadens their scope, audience, and interoperability with other components of the scholarly information infrastructure (HyperCities, 2009; Presner, 2010, forthcoming).

In sum, choices of data sources, research methods, and research problems are inextricably linked. Research methods in the sciences and in the humanities are becoming more data-driven. The key to "better" data – that is, data suitable for curation, reuse, and sharing – is capturing data as cleanly as possible and as early as possible in its life cycle. Agreements about data sources, structures, and formats will further the development of information infrastructure for digital humanities scholarship.

Collaboration

The size of collaborations is increasing in all fields, as measured by the number of co-authors on papers, and at the fastest rate in the sciences (Cronin, 2005). In sciences that rely heavily on instrumentation, such as astronomical observatories and particle accelerators, collaborations are large, diverse, and essential. Sciences that are more inductive and are conducted in field settings, such as habitat biology, tend to work in smaller groups. Sciences of all sizes are grappling with data management issues, as data are the glue – and often the product – of collaboration.

As noted above, the new forms of scholarship characterized by eResearch are information- and data-intensive, distributed, collaborative, and multi-disciplinary. Collaborations, when effective, produce new knowledge that is greater than the sum of what the participating individuals could accomplish alone. In fields where collaboration is the norm, graduate students learn teamwork, whether in the laboratory, the field, or in group work on data collection and analysis. Science dissertations frequently are carved out of larger group projects, with the student identifying a research problem worthy of sustained investigation. Funding agencies in the sciences consider dissertations to be important products of awards to faculty investigators. Dissertations and theses are listed explicitly in National Science Foundation annual reports, for example.

While the digital humanities are increasingly collaborative, elsewhere in the humanities the image of the "lone scholar" spending months or years alone in dusty archives, followed years later by the completion of a dissertation or monograph, still obtains. Students often are discouraged from conducting dissertation research under a faculty grant. Instead, they are expected to spend yet more time identifying funding for solo research. When one is groomed to work alone and does so for the years required to complete the doctorate, collaborative practices do not come easily.

Friedlander (2009, p. 6) argues that for digital humanities to thrive, "one component must be a set of organizational topics and questions that do not bind research into legacy categories and do invite interesting collaborations that will allow for creative cross-fertilization of ideas and techniques and then spur new questions to be pursued by colleagues and students." As she suggests, the digital humanities need to move beyond large numbers of small, uncoordinated projects. Collaborative projects attract more resources and more attention. If properly designed, they also may be more sustainable, creating platforms on which new projects can be constructed. The plethora of boutique digital humanities projects risks the same fate as most digital learning objects. While intended for general use, they lack a common technical platform, common data structures, and means to aggregate or decompose modules to a useful level of granularity (Borgman et al., 2008).

An indicator of collaboration in the digital humanities community is the shift over the last two decades from a focus on the audience – those who might learn or appreciate the cultural content presented – to a focus on participation, in which scholars, students, and the public can contribute content or conduct their own investigations (Electronic Cultural

Atlas Initiative, 2009; Ivanhoe: a game of critical interpretation, 2009; Tibetan and Himalayan Library, 2009; Presner, 2010, forthcoming). The latter approach is more readily sustainable, as more people have vested interests in its capabilities and availability, and because it reflects current technical practices for Internet architecture (Architecture of the World Wide Web, 2004; Semantic Web Activity: W3C, 2009).

Scholarly collaboration is much studied but little understood. Among the predictors of success are the ability to achieve a common vocabulary and shared knowledge (Kanfer et al., 2000; Olson & Olson, 2000). The more disciplines involved, the more effort is required to achieve common ground. Investments must be made in learning enough about each other's disciplines that at least a pidgin language is established (Galison, 1997). Relationships take time, and must be nurtured. One important measure of success, and a worthwhile goal in eResearch, is that papers suitable for publication in each of the participating disciplines arise from a joint project. The recent multi-national, multi-disciplinary, multi-year funding awards for innovative uses of data included several humanities-computer science partnerships (Digging into Data, 2009). Virtual Vellum, for example, applies advanced computational methods to explore authorship of 15th century manuscripts (Ainsworth, 2009). The results are likely to advance the state of optical character recognition and other computing techniques with broad application.

In sum, the digital humanities community could benefit from more collaborative partnerships within the field and between the humanities and disciplines such as computer science. Collaboration requires investment in listening skills, always being alert to nuanced differences in assumptions, theories, definitions, and methods. Lessons and skills learned from these partnerships can enhance the scholarship of all participants. Common technology platforms also are important to achieve interoperability and sustainability, and can be leveraged as investments across projects.

Incentives to participate

Constructing a critical mass of data sources for scholarship in any field presumes that people will share the products of their research. Because data and collaboration are so central to the methods of digital scholarship, data sharing is an important indicator of success for eResearch, although practices are somewhat different in the sciences and in the humanities.

The public nature of scholarship has deep roots. Notions of "open science" date back at least to Francis Bacon, with scientific findings being accepted only after peer review. Scholars' incentives to share their results include recognition and acceptance of their work, which in turn drives hiring and promotion. In the sciences, authors may be required to release data as a condition of publishing the papers on which they are based. Funding agencies also are becoming more assertive about the release of data that result from grants. However, publishing data is a far less mature practice than is publishing books and articles. Releasing a major dataset rarely brings as much recognition as releasing a major paper or book, but that balance is shifting, at least in the sciences (Borgman, 2007; Hey et al., 2009).

Scholars compete as well as collaborate, and thus have reasons *not* to share their data sources. The following are disincentives that apply to all disciplines, albeit to varying degrees (Borgman, 2007): (1) Faculty get more rewards for publishing papers and books than for releasing data; (2) the effort of individuals to document their data for use by others is much greater than the effort required to document them only for use by themselves and their research team; (3) data and sources offer a competitive advantage and are essential to establishing the priority of claims; and (4) data are often viewed as one's own intellectual property to be controlled, whether or not the data (or their sources) are legally owned. Means exist to address each of these concerns, but all are complex responses to a complex environment.

The first disincentive is the most universal across disciplines. The sciences and medicine are under the greatest pressure to release their data. In these disciplines the reward structure is adapting, and repositories and data structures exist. While humanities scholars are under less pressure to release their data and sources, they are contributing models, modules, and tools to participatory projects and shared collections.

Data documentation is an issue in all fields, but as the volume of data increases, consistent documentation becomes progressively more necessary. Once data are captured cleanly, sharing them later becomes less of a problem. Humanities scholars are acutely aware of the importance of metadata and finding aids in discovering sources. Metadata are equally important for data curation. Scholars understand the roles that documentation must play, while librarians and archivists have the expertise in documentation standards, practices, and technologies. Data documentation is thus an obvious area of partnership for humanities scholars and information professionals, together addressing the requirements for sustainability of research products.

The third disincentive – competitive advantage – is often addressed in the sciences through embargoes, whereby the investigators have a set period of time (from a few months to a few years, depending on the field) after the end of a grant before being required to share their data. Embargoes serve two complementary purposes: they protect the scholars' control over data, and they ensure that others will have access to the data within a reasonable time period. In the humanities, scholars are similarly concerned about controlling access to the sources of their data, whether the Dead Sea Scrolls or a set of manuscripts in a university archive, until they have published their research. As data sources such as manuscripts and out-of-print books are digitized and made publicly available, individual scholars will be less able to hoard their sources. This effect of digitization on humanities scholarship has been little explored, but could be profound. Open access to sources promotes participation and collaboration, while the privacy rules of libraries and archives ensure that the identity of individuals using specific sources is not revealed. Libraries and archives endeavor to maintain privacy in the use of digital as well as print sources. However, when digital content is controlled by commercial entities, protecting the privacy of users is a greater concern (Mass Digitization: Implications for Information Policy, 2006; Hoofnagle, 2009).

Intellectual property, the fourth disincentive to share data and sources, is the most intractable. The need to establish data sharing agreements in collaborative projects arose early in eScience initiatives and is far from resolved (David, 2003; David & Spence, 2003). In the case of the sciences, ownership – or at least control – usually can be clarified through negotiation. If the research depends upon material acquired from others, such as cell lines, rules on data release will be governed by contract. The reliance of humanities scholarship on cultural records, as discussed above, creates particularly complex intellectual property challenges in the sharing of data. For example, an art historian usually can publish his or her notes, but not the paintings on which the research is based. In the case of cultural models such as digital cities, it can be difficult to distinguish between data that represent an individual city and the model in which those data are incorporated. Difficulties in separating data from models (a problem in the sciences and in the humanities) plague both curation and data release efforts (HyperCities, 2009; Rome Reborn, 2009; Serving and Archiving Virtual Environments, 2009).

In sum, the digital humanities encounter most of the same incentives and disincentives for sharing data and sources faced by the sciences and by other disciplines. The details play out somewhat differently, of course. The need to build critical masses of cultural sources and interoperable technology platforms affirms the need to broker agreements about data. If the infrastructure for the digital humanities errs toward openness, as is the norm in much of the sciences, the field will advance more quickly.

Learning

The last comparison between the sciences and humanities, but by no means the least, is the role of information technology in learning. "Cyberlearning," as argued by the National Science Foundation's Task Force, can leverage the nation's investment in cyberinfrastructure to benefit learning at all ages – "from K to grey" (Borgman et al., 2008). This argument was made earlier in the humanities, claiming that cyberinfrastructure could serve the humanities both for scholarship and for making cultural material more readily accessible for learning and outreach (Unsworth et al., 2006). Cyberlearning is defined as the use of networked computing and communications technologies to support learning. The scope of cyberlearning concerns in the Task Force report was necessarily constrained to the U.S. and to the domains funded by the NSF, which do not include the arts and humanities. However, the Task Force noted explicitly that the value of cyberlearning encompasses the sciences, social sciences, humanities, and arts, and is an important international initiative.

Several of the recommendations for advancing the state of cyberlearning have analogies for advancing the state of digital humanities. One is the need to build a vibrant field by promoting cross-disciplinary communities, publishing best practices, and recruiting diverse talents. The Cyberlearning Task Force made a careful distinction between

cyberlearning as learning *with* distributed computing technologies and workforce development as teaching people *about* cyberinfrastructure. The latter is also a concern of the National Science Foundation (Cyberinfrastructure Vision for 21st Century Discovery, 2007). In the humanities as in the sciences, people need to learn *about* cyberinfrastructure before they can learn *with* it – or can use it for their research and teaching.

Another analogous recommendation from the cyberlearning report is the need to instill a "platform perspective." As noted earlier, the takeup rate of digital learning modules has been limited by reliance on unique tools, proprietary software, and general lack of interoperability. Unless products are easily adapted to new uses, others have little incentive to invest in them. Both cyberinfrastructure and cyberlearning initiatives are constructing common technical platforms that will improve the sustainability and reuse of tools, services, and content. Some of these technical platforms can be leveraged for digital humanities scholarship. Where capabilities are lacking, the community can work in concert to construct them. Common platforms and standards are among the goals of the Mellon-funded Bamboo project, for example (Project Bamboo, 2009).

The Cyberlearning Task Force also recommended initiatives to enable students to use data. By embedding data skills early in the science curriculum – in the primary grades where feasible – students can learn to "think like scientists" early on. Hands-on science approaches endeavor to engage students in "real" science, making it more interesting and exciting than purely textbook approaches (Pea, Wulf, Elliott & Darling, 2003). Projects like the Sloan Digital Sky Survey and eBird encourage individuals to contribute their observations – whether about the sky or about birds in their backyards – for use in scientific investigations (Sloan Digital Sky Survey, 2006; eBird, 2009). The same promise applies to the humanities. If students can explore cultural records from the early grades and learn to construct their own narratives, they may find the study of humanities more lively. By the time they are college students, they will have learned methods of collaborative work and the use of distributed tools, sources, and services. Projects such as Perseus, HyperCities, and the Valley of the Shadow already enable students in humanities courses to engage in new forms of collaborative discourse (Perseus Digital Library, 2009; Ayres, 2004; Presner, 2010, forthcoming).

Lastly, the Task Force made a strong recommendation to the NSF to promote open educational resources. Educational content resulting from cyberlearning grants should be made available online with permission for unrestricted use and recombination. New proposals for research and development in cyberlearning should include plans to make their materials available and sustainable. These recommendations are relevant to all disciplines. Open educational resources are growing rapidly in variety and number (Atkins, Brown & Hammond, 2007; Baker, 2009; Thierstein, 2009). Licensing models such as Creative Commons (Creative Commons, 2009) now include specific capabilities for licensing learning materials (ccLearn, 2009) and scientific data (Science Commons, 2009). Digital humanities projects, whether or not they include a learning component, also can benefit from Creative Commons licenses. The owners of intellectual property retain their copyright; they simply license it for reuse under publicly stated conditions.

Intellectual property owned by others must not be appropriated, of course, but usually it can be linked if not specifically licensed.

Openness matters for the digital humanities for reasons of interoperability, discovery, usability, and reusability. Open resources – that is, those that can be used under license or are in the public domain – are more malleable for research and for learning. They can be mixed up and mashed up, and others can add value to them. Resources that are available via open repositories also are more readily discovered than those posted on local websites (OER Commons, 2009; Open Education, 2009; The Case of the Textbook: Open or Closed?, 2009; Atkins et al., 2007).

In sum, cyberlearning is important for the digital humanities for a number of reasons. One is the need to learn how to use and how to evaluate digital cultural materials early; graduate school is rather late. Second is the need to build common technology platforms for digital humanities scholarship, which will advance the field by leveraging efforts and resources and by increasing interoperability. Third is the value of open access to resources, which then become more malleable for research and for learning. Last is the need to build a strong community of digital humanities scholars, one that represents a much larger portion of the humanities than is the case today.

SUMMARY

My student's complaint, "So what use are the digital libraries, if all they do is put digitally unusable information on the web?" nicely captures the challenges facing the humanities today. Digital content, tools, and services all exist, but they are not necessarily useful or usable. Much work remains to build the scholarly infrastructure necessary for digital scholarship to become mainstream in the humanities. Humanities scholars must lead the effort, because only they understand the goals and requirements for what *should* be built. Librarians, archivists, programmers, and computer scientists will be essential collaborators, each bringing complementary skills.

A number of developments in cyberinfrastructure, eScience, and eResearch offer guidance to the digital humanities community in the quest to become a more established field with a broader base of infrastructure. One is in the area of publication practices. The humanities lag in digital publication of journals and books. Digital publishing, while far from a panacea, offers a number of advantages in the speed, scope, and format of communication. Scholarly print publishing is on the decline, and those who publish only in print form risk being isolated, talking only to each other. More digital-only venues are needed, where dynamic and visual work can be published in its vernacular form.

Another area is the dissemination and use of data. The humanities community should continue to clarify their choices of data and data sources, for these will drive what content is produced, captured, managed, and available for reuse. Questions of data are closely related to research methods, which also are evolving. Data-driven research methods are most valuable when they enable scholars to ask new questions in new ways.

Collaboration is essential in digital humanities projects. Few individuals have the range of expertise required to execute these projects alone. Humanists should continue to seek out complementary partners and encourage people to listen and learn from each other. Working together is also more likely to lead to common platforms and other means of reducing the overhead of technical projects.

In both the sciences and the humanities, incentives to share one's writing are more obvious than are incentives to share one's data and sources. In the sciences, data release is being encouraged (or required) by journals and funding agencies, and data-driven research methods can draw upon large corpora that grow as new observations are contributed. In the humanities, data release is less of an issue, but the availability of common technical platforms, tools, and services will promote the sharing of data and sources. The disincentives to share are complex in both the sciences and the humanities, but are being addressed. As the sciences learn how to share data and to share credit for their findings, the humanities can build upon their best practices. Intellectual property constraints remain a major stumbling block, and the considerations vary between the sciences and the humanities.

Opportunities for using cyberinfrastructure for learning exist in all disciplines. Distributed access to scholarly content, common technical platforms, and open resources will advance the humanities as well as the sciences.

A CALL TO ACTION

In the process of developing the keynote presentation for the 2009 Digital Humanities Conference and in writing this paper, I consulted many individuals in the digital humanities community for their thoughts on the issues facing the field. From these discussions and my analyses above, five pressing problems emerged.

What are data?

What constitute data in the humanities? What are data sources? How are they made, shared, valued, used, and reused? Answering these questions will enable the digital humanities community to be more articulate about its scope and its goals, and better positioned to identify their requirements for infrastructure.

What are the infrastructure requirements?

The sciences have struggled with this question for a decade or two already. They have convened workshops and study panels, and launched funding initiatives addressed specifically at defining, designing, and deploying the necessary infrastructure for eScience. The humanities have tackled this question on a much smaller scale, leaving them in the position of building upon the infrastructure constructed by and for other

disciplines. As Johanna Drucker (2009) put it so well, "them is us." It is time for the community to articulate its own requirements and to act upon them.

Where are the social studies of digital humanities?

Why is no one following digital humanities scholars around to understand their practices, in the way that scientists have been studied for the last several decades? This body of research has informed the design of scholarly infrastructure for the sciences, and is a central component of cyberinfrastructure and eScience initiatives. Given how rapidly scholarship in the humanities is evolving, it is fertile ground for behavioral research. The humanities community should invite more social scientists as research partners and should make themselves available as objects of study. In doing so, the community can learn more about itself and apply the lessons to the design of tools, services, policies, and infrastructure.

What is the humanities laboratory of the 21st century?

This is a question of great concern to research libraries as well as to humanities scholars. The library continues to be a laboratory for the humanities, but not the only laboratory. Humanities scholars run computing laboratories and may work in distributed virtual environments for research and for learning. Humanists need to partner both with librarians and with the information technology planning and policy groups on their campuses. These communities urgently need to "think together" about the common challenges faced in a time of shrinking budgets for collections, physical space, staffing, and technology services.

What is the value proposition for digital humanities in an era of declining budgets?

For universities, the current economic recession is like no other. Public and private universities alike are re-examining core principles as budgets are slashed by 10% to 30% from one year to the next. Nothing is sacred, and "because it's beautiful" is not a viable economic argument. The sciences have been remarkably effective at making the argument for their value in economic and political terms, whether to university administrations, legislatures, funding agencies, or the general public. While the humanities will have difficulty making parallel arguments in terms of economic competitiveness and medical advances, they have plenty to offer in terms of cultural understanding, writing and design skills, and critical thinking. Digital scholarship also promotes technical skills, which can be highlighted.

Digital projects require resources in the form of computers, software, staff, and content. Non-digital scholarship also costs money, of course, but more often in the form of travel and subsistence expenses for research in remote archives. Tradeoffs in travel and digitization can be made more explicit. The number of people who will use and benefit from any given project also can be made clearer. Investments in common technical platforms and standards that leverage resources across larger numbers of people and projects are easier to justify.

The digital humanities community has produced some beautiful work and made many advances in technology, design, and standards. Now is the moment to consolidate that knowledge and to articulate the community's requirements and goals. Go forth and do great things…

ACKNOWLEDGEMENTS

I am grateful to the colleagues who provided thoughtful commentary on an earlier draft of this paper, including Murtha Baca, Gregory Britton, and Maureen Whalen of the Getty Trust; Johanna Drucker, Alberto Pepe, Todd Presner, and Katie Shilton of UCLA; Amy Friedlander, Council on Library and Information Resources; Bernard Frischer, University of Virginia; Alexander Parker, Harvard University; and two anonymous reviewers.

Many other people were very generous with their time in response to my inquiries about the past, present, and future of the digital humanities, including (in alphabetical order) William Dutton, Oxford Internet Institute; Neil Fraistat, University of Maryland; Richard Furuta, Texas A&M; Kimberly Garmoe, Anne Gilliland, UCLA; Charles Henry, Council on Library and Information Resources; Jason Hewitt, UCLA; Jieh Hsiang, National Taiwan University; Marina Jirotka, Oxford University; Matthew Kirschenbaum, University of Maryland; Clifford Lynch, Coalition for Networked Information; Lev Manovich, University of California, San Diego; Ann O'Brien, Loughborough University; Susan Parker, UCLA; Allen Renear, University of Illinois; David Robey, Oxford University; Ben Shneiderman, University of Maryland; Harold Short and Paul Spence, King's College, London; Joshua Sternfeld, UCLA; Sarah Thomas, Bodleian Library; Sharon Traweek, UCLA; Anne Trefethen, University of Oxford; John Unsworth, University of Illinois; Sarah Watstein and Robert Winter, UCLA.

REFERENCES

Ainsworth, P. (2009). Virtual Vellum. Retrieved from http://www.shef.ac.uk/hri/projects/projectpages/virtualvellum.html on 31 December 2009.

Alliance of Digital Humanities Organizations. (2009). Retrieved from http://www.digitalhumanities.org/ on 16 August 2009.

Anderson, I. (2004). Are you being served? Historians and the search for primary sources. Archivaria, 58: 81-129.

Architecture of the World Wide Web. (2004). Retrieved from http://www.w3.org/TR/webarch/ on 2 June 2009.

ArXiv.org e-Print archive. (2009). Retrieved from http://arxiv.org/ on 12 August 2009.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G. C., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. F. & Wouters, P. (2004). An International Framework to Promote Access to Data. Science, 303(5665): 1777-1778.

Atkins, D. E., Brown, J. S. & Hammond, A. L. (2007). A Review of the Open Educational Resources (OER) Movement: Achievements, Challenges, and New Opportunities. William and Flora Hewlett Foundation. Retrieved from http://www.hewlett.org/oer on 9 September 2009.

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messina, P., Messerschmitt, D. G., Ostriker, J. P. & Wright, M. H. (2003). Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon panel on Cyberinfrastructure. National Science Foundation. Retrieved from http://www.nsf.gov/cise/sci/reports/atkins.pdf on 18 September 2006.

Ayres, E. L. (2004). The Valley of the Shadow. University of Virginia. Retrieved from http://valley.vcdh.virginia.edu/ on 28 September 2005.

Bailey, C. (2005). Open Access Bibliography: Liberating Scholarly Literature with E-Prints and Open Access Journals. Washington, DC: Association of Research Libraries. Retrieved from http://info.lib.uh.edu/cwb/oab.pdf on 28 September 2006.

Baker, J. (2009). It Takes a Consortium to Support Open Textbooks. EDUCAUSE Review, 44(1): 30-33.

Bates, M. J. (1996a). Document familiarity, relevance, and Bradford's law: The Getty online searching project report no 5. Information Processing & Management, 32(6): 697-707.

Bates, M. J. (1996b). The Getty end-user online searching project in the humanities: Report No 6: Overview and conclusions. College & Research Libraries, 57(6): 514-523.

Bates, M. J., Wilde, D. N. & Siegfried, S. L. (1993). An analysis of search terminology used by humanities scholars -- The Getty Online Searching Project Report No.1. Library Quarterly, 63(1): 1-39.

Bates, M. J., Wilde, D. N. & Siegfried, S. L. (1995). Research practices of humanities scholars in an online environment - The Getty Online Searching Project Report No. 3. Library & Information Science Research, 17(1): 5-40.

Bell, G., Hey, T. & Szalay, A. (2009). Beyond the data deluge. Science, 323: 1297-1298.

Bollen, J. & Van de Sompel, H. (2008). Usage Impact Factor: the effects of sample characteristics on usage-based impact metrics. Journal of the American Society for Information Science and Technology, 59(1): 136-149.

Borgman, C. L. (1999). What are digital libraries? Competing visions. Information Processing & Management, 35(3): 227-243.

Borgman, C. L. (2000). From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World. Cambridge, MA: The MIT Press.

Borgman, C. L. (2007). Scholarship in the Digital Age: Information, Infrastructure, and the Internet. Cambridge, MA: MIT Press.

Borgman, C. L. (2009). Scholarship in the Digital Age: Blurring the Boundaries between the Sciences and the Humanities. Digital Humanities '09, College Park, MD, Maryland Institute for Technology in the Humanities. Retrieved from http://works.bepress.com/borgman/216/ on 12 August 2009.

Borgman, C. L., Abelson, H., Dirks, L., Johnson, R., Koedinger, K. R., Linn, M. C., Lynch, C. A., Oblinger, D. G., Pea, R. D., Salen, K., Smith, M. S. & Szalay, A.

(2008). Fostering Learning in the Networked World: The Cyberlearning Opportunity and Challenge. A 21st Century Agenda for the National Science Foundation. Report of the NSF Task Force on Cyberlearning. Office of Cyberinfrastructure and Directorate for Education and Human Resources. National Science Foundation. Retrieved from http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf08204 on 12 August 2008.

Borgman, C. L., Wallis, J. C., Mayernik, M. S. & Pepe, A. (2007). Drowning in Data: Digital Library Architecture to Support Scientists' Use of Embedded Sensor Networks. Proceedings of the 7th Joint Conference on Digital Libraries, Vancouver, BC, Association for Computing Machinery. 269 - 277.

Buckland, M. K. (1991). Information as thing. Journal of the American Society for Information Science, 42(5): 351-360.

Case, D. O. (2006). Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior (2nd ed.). San Diego: Academic Press.

ccLearn. (2009). A Project of Creative Commons. Retrieved from http://learn.creativecommons.org/ on 6 August 2009.

Choudhury, S., DiLauro, T., Szalay, A., Vishniac, E., Hanisch, R., Steffen, J., Milkey, R., Ehling, T. & Plante, R. (2008). Digital data preservation for scholarly publications in astronomy. International Journal of Digital Curation, 2(2): 20-30. Retrieved from http://www.ijdc.net/index.php/ijdc/issue/view/3 on 17 August 2009.

Choudhury, S. & Stinson, T. (2007). The Virtual Observatory and the Roman de la Rose: Unexpected Relationships and the Collaborative Imperative. Academic Commons. Retrieved from http://www.academiccommons.org/commons/essay/VO-and-roman-de-la-rose-collaborative-imperative on 22 July 2008.

Crane, G. R. (2006). What do you do with a million books? D-Lib Magazine, 12(3). Retrieved from http://www.dlib.org/dlib/march06/crane/03crane.html on 17 August 2006.

Crane, G. R., Babeu, A. & Bamman, D. (2007). eScience and the humanities. International Journal on Digital Libraries, 7(1-2): 117-122.

Crane, G. R., Chavez, R. F., Mahoney, A., Milbank, T. L., Rydberg-Cox, J. A., Smith, D. A. & Wulfman, C. E. (2001). Drudgery and deep thought: Designing a digital library for the humanities. Communications of the Association for Computing Machinery, 44(5): 35-4018 April 2006.

Creative Commons. (2009). Retrieved from http://www.creativecommons.org on 18 April 2009.

Cronin, B. (2005). The Hand of Science: Academic Writing and its Rewards. Lanham, MD: Scarecrow Press.

Cyberinfrastructure Vision for 21st Century Discovery (2007). National Science Foundation. Retrieved from http://www.nsf.gov/pubs/2007/nsf0728/ on 17 July 2007.

David, P. A. (2003). The economic logic of 'Open Science' and the balance between private property rights and the public domain in scientific data and information: A primer. In. The Role of the Public Domain in Scientific Data and Information.

Washington, D.C., National Academy Press: 19-34. Retrieved from http://siepr.stanford.edu/papers/pdf/02-30.html on 30 September 2006.

David, P. A. & Spence, M. (2003). Towards Institutional Infrastructures for E-Science: The Scope of the Challenge. Oxford Internet Institute Research Reports: University of Oxford. 92 Retrieved from http://129.3.20.41/eps/le/papers/0502/0502002.pdf on 30 September 2006.

Digging into Data. (2009). Retrieved from http://www.diggingintodata.org/ on 31 December 2009.

Digital Humanities Manifesto (2009). UCLA. Retrieved from http://dev.cdh.ucla.edu/digitalhumanities/2008/12/15/digital-humanities-manifesto/ on 3 August 2009.

Directory of Open Access Journals. (2009). Open Society Initiative, Scholarly Publishing and Academic Resources Coalition. Retrieved from http://www.doaj.org on 16 August 2009.

Directory of Open Access Repositories. (2008). University of Nottingham, UK and University of Lund, Sweden. Retrieved from www.opendoar.org on 16 August 2009.

Drucker, J. (2009). Blind Spots: Humanists must plan their digital future. Chronicle of Higher Education, 55(30): B6. Retrieved from http://chronicle.com/free/v55/i30/30b00601.htm on 25 June 2009.

Duguid, P. (2007). Inheritance and loss? A brief survey of Google Books. First Monday, 12(8). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1972/1847 on 3 September 2009.

eBird. (2009). Cornell Lab of Ornithology and Audobon Society. Retrieved from http://ebird.org/content/ebird/ on 9 September 2009.

Edwards, P. N., Jackson, S. J., Bowker, G. C. & Knobel, C. P. (2007). Understanding Infrastructure: Dynamics, Tensions, and Design. National Science Foundation: University of Michigan. NSF Grant 0630263. Retrieved from http://hdl.handle.net/2027.42/49353 on 26 July 2007.

Electronic Cultural Atlas Initiative. (2009). Retrieved from http://www.ecai.org on 15 September 2009.

EPrints. (2010). Retrieved from http://www.eprints.org/ on 2 January 2010.

Friedlander, A. (2008). Head in the Clouds and Boots on the Ground: Science, Cyberinfrastructure and CLIR. Kanazawa Institute of Technology Library Roundtable. Retrieved from http://www.clir.org/pubs/resources/articles.html on 15 August 2008.

Friedlander, A. (2009). Asking questions and building a research agenda for digital scholarship. In. Working Together or Apart: Promoting the Next Generation of Digital Scholarship. Washington, DC, Council on Library and Information Resources. CLIR Publication No. 145: 1-15. Retrieved from http://www.clir.org on 15 June 2009.

Frischer, B. (2004). Testimony to the Commission on Cyberinfrastructure for the Humanities and Social Sciences. American Council of Learned Societies. Retrieved from

http://www.acls.org/cyberinfrastructure/cyber_meeting_notes_october.htm#frischer_summary on 6 August 2006.

Frischer, B. (2009). Art and Science in the Age of Digital Reproduction: From Mimetic Representation to Interactive Virtual Reality. I Congreso Internacional de Arqueología e Informática Gráfica, Patrimonio e Innovación, Sevilla 17-20 de Junio de 2009.

Galison, P. (1997). Image and Logic: A Material Culture of Microphysics. Chicago: University of Chicago Press.

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. & Heber, G. (2005). Scientific data management in the coming decade. CT Watch Quarterly, 1(1). Retrieved from http://www.ctwatch.org/quarterly/articles/2005/02/scientific-data-management/ on 25 August 2006.

Gray, J. & Szalay, A. (2002). The world-wide telescope. Communications of the ACM, 45(11): 51-55.

Hamma, K. (2009). Professionally indisposed to change. EDUCAUSE Review, 44(2): 8-9.

Hey, T., Tansley, S. & Tolle, K. (Eds.). (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/ on 16 December 2009.

HyperCities. (2009). UCLA. Retrieved from http://www.hypercities.com on 17 August 2009.

Institute for Advanced Technology in the Humanities. (2009). University of Virginia. Retrieved from http://www.iath.virginia.edu on 6 August 2009.

Ivanhoe: a game of critical interpretation. (2009). Retrieved from http://www.speculativecomputing.org/ivanhoe/index.html on 3 September 2009.

Jaschik, S. (2008). Publishing and values. Inside Higher Ed. Retrieved from http://www.insidehighered.com/news/2007/08/22/anthro on 10 March 2009.

Jaschik, S. (2009). Farewell to the printed monograph. Inside Higher Ed. Retrieved from http://www.insidehighered.com/news/2009/03/23/michigan on 24 March 2009.

Journal of Post Modern Culture. (2000). Retrieved from http://pmc.iath.virginia.edu/ on 2 September 2009.

Journal of the Society of Architectural Historians. (2009). Retrieved from http://www.sah.org/index.php?src=gendocs&ref=JSAH&category=Publications on 2 September 2009.

Kanfer, A. G., Haythornthwaite, C., Bruce, B. C., Bowker, G. C., Burbules, N. C., Porac, J. F. & Wade, J. (2000). Modeling distributed knowledge processes in next generation multidisciplinary alliances. Information Systems Frontiers, 2(3-4): 317-331.

King, C. J., Harley, D., Earl-Novell, S., Arter, J., Larence, S. & Perciali, I. (2006). Scholarly Communication: Academic Values and Sustainable Models. Andrew W. Mellon Foundation, Center for Studies in Higher Education, University of California, Berkeley. Retrieved from http://cshe.berkeley.edu/publications/publications.php?id=230 on 28 July 2006.

Kurtz, M. J. & Bollen, J. (2010). Usage bibliometrics. In Cronin, B. (Ed.). Annual Review of Information Science and Technology. Medford, NJ, Information Today. 44.

Long-Lived Digital Data Collections. (2005). National Science Board. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/ on 18 April 2009.

Lynch, C. A. (2002). Digital collections, digital libraries and the digitization of cultural heritage information. First Monday, 7(5). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/949/870 on 8 September 2009.

Lynch, C. A. (2003). Life after graduation day: Beyond the academy's digital walls. EDUCAUSE Review, 38(5): 12-13.

Lynch, C. A. & Garcia-Molina, H. (1995). Interoperability, scaling, and the digital libraries research agenda. IITA Digital Libraries Workshop. Retrieved from http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html on 1 October 2006.

Mahoney, A. (2002). Finding texts in Perseus. New England Classical Journal, 29(1): 32-34.

Manovich, L. (2009). Cultural Analytics. Software Studies Initiative, University of California, San Diego. Retrieved from http://lab.softwarestudies.com/2008/09/cultural-analytics.html on 1 September 2009.

Marchionini, G. & Crane, G. R. (1994). Evaluating hypermedia and learning: Methods and results from the Perseus project. ACM Transactions on Information Systems, 12(1): 5-34.

Maryland Institute for Technology in the Humanities. (2009). University of Virginia. Retrieved from http://mith.umd.edu/ on 6 August 2009.

Monastersky, R. (2005). The number that's devouring science. Chronicle of Higher Education, 52(8): A12-A17.

National Centre for Text Mining. (2009). Retrieved from http://www.nactem.ac.uk/ on 2 September 2009.

Nunberg, G. (2009). Google's Book Search: A Disaster for Scholars. Chronicle of Higher Education. Retrieved from http://chronicle.com/article/Googles-Book-Search-A/48245/ on 3 September 2009.

OECD Principles and Guidelines for Access to Research Data from Public Funding (2007). Organisation for Economic Co-Operation and Development.

OER Commons. (2009). Retrieved from http://www.oercommons.org/ on 18 April 2009.

Olson, G. M. & Olson, J. S. (2000). Distance matters. Human-Computer Interaction, 15(2-3): 139-178.

Open Archives Initiative Protocol for Metadata Harvesting. (2009). Retrieved from http://www.openarchives.org/pmh/ on 12 August 2009.

Open Content Alliance. (2009). Retrieved from http://www.opencontentalliance.org/ on 16 August 2009.

Open Education. (2009). A Project of Creative Commons. Retrieved from http://opened.creativecommons.org/Main_Page on 6 August 2009.

PAN-STARRS. (2009). Panoramic Survey Telescope & Rapid Response System. Retrieved from http://pan-starrs.ifa.hawaii.edu/public/ on 14 September 2009.

Pea, R., Wulf, W. A., Elliott, S. W. & Darling, M. A. (2003). Planning for Two Transformations in Education and Learning Technology: Report of a Workshop. Washington, D.C.: National Academies Press. Retrieved from http://www.nap.edu/catalog/10789.html on 8 September 2009.

Perseus Digital Library. (2009). Tufts University. Retrieved from http://www.perseus.tufts.edu/hopper/ on 3 September 2009.

Poe, M. (2001). Note to self: Print monograph dead; invent new publishing model. The Journal of Electronic Publishing, 7(2). Retrieved from http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.204 on 16 August 2009.

Presner, T. S. (2010, forthcoming). HyperCities: Building a Web 2.0 learning platform. In Natsina, A. & Tagialis, T. (Eds.). Teaching Literature at a Distance. London and New York, Continuum books.

Presner, T. S. & Johanson, C. (2009). The Promise of Digital Humanities: A White Paper. 1-19. Retrieved from http://www.digitalhumanities.ucla.edu/ on 12 August 2009.

Project Bamboo. (2009). Retrieved from http://projectbamboo.org/ on 15 September 2009.

Reedijk, J. & Moed, H. (2008). Is the impact of journal impact factors decreasing? Journal of Documentation, 64(2): 183-192.

Reference Model for an Open Archival Information System. (2002). Recommendation for Space Data System Standards, Consultative Committee for Space Data Systems Secretariat, Program Integration Division (Code M-3), National Aeronautics and Space Administration. Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf on 4 October 2006.

Research Papers in Economics. (2009). University of Connecticut. Retrieved from http://www.repec.org/ on 12August 2009.

Rice University Press Mission Statement. (2008). Retrieved from http://rup.rice.edu/about/mission on 1 September 2009.

Rome Reborn. (2009). Retrieved from http://www.romereborn.virginia.edu/ on 17 August 2009.

Samuelson, P. (2009). The Dead Souls of the Google Book Search Settlement. Communications of the Association for Computing Machinery, 52(7): 28-30.

Scheiner, S. M. (2004). Experiments, observations, and other kinds of evidence. In Taper, M. L. & Lele, S. R. (Eds.). The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations. Chicago, University of Chicago Press: 51-66.

Science Commons. (2009). A Project of Creative Commons. Retrieved from http://sciencecommons.org/about/index.html on 6 August 2009.

Semantic Web Activity: W3C. (2009). Retrieved from http://www.w3.org/2001/sw/ on 14 April 2009.

Serving and Archiving Virtual Environments. (2009). Serving and Archiving Virtual Environments. Retrieved from http://www3.iath.virginia.edu/save/ on 13 August 2009.

SHERPA/RoMEO: Publisher copyright policies & self-archiving. (2009). Retrieved from http://www.sherpa.ac.uk/romeo/ on 12 August 2009.

Siegfried, S. L., Bates, M. J. & Wilde, D. N. (1993). A profile of end-user searching behavior by humanities scholars - The Getty online searching project report no. 2. Journal of The American Society for Information Science, 44(5): 273-291.

Sloan Digital Sky Survey. (2006). Retrieved from http://www.sdss.org/ on 15 August 2006.

Smith, D. A., Mahoney, A. & Crane, G. R. (2002). Integrating harvesting into digital library content. 2nd ACM IEEE-CS Joint Conference on Digital Libraries, Portland, OR, New York: ACM. 183-184. Retrieved from http://www.perseus.tufts.edu/Articles/oaishort.pdf on 4 October 2006.

Stone, S. (1982). Humanities scholars: Information needs and uses. Journal of Documentation, 38(4): 292-313.

Survey Research Center, Institute for Social Research. (2009). University of Michigan. Retrieved from http://www.isr.umich.edu/src/ on 16 August 2009.

Survey Research Center, UC-Berkeley. (2009). University of California, Berkeley. Retrieved from http://srcweb.berkeley.edu/ on 16 August 2009.

Szalay, A. (2008). Jim Gray, astronomer. Communications of the ACM, 51(11): 59-65.

The Case of the Textbook: Open or Closed? (2009). EDUCAUSE Review, 44(1): 13.

The effect of open access and downloads ('hits') on citation impact: a bibliography of studies. (2009). The Open Citation Project - Reference Linking and Citation Analysis for Open Archives. Retrieved from http://opcit.eprints.org/oacitation-biblio.html on 16 August 2009.

The Facts about Open Access: A Study of the Financial and Non-Financial Effects of Alternative Business Models on Scholarly Journals (2005). Kaufman-Wills Group LLC: Association of Learned and Professional Society Publishers. Retrieved from http://sippi.aaas.org/Pubs/ on 27 September 2007.

Thierstein, J. (2009). Education in the Digital Age. EDUCAUSE Review, 44(1): 33-34.

Tibbo, H. R. (2003). Primarily history in America: How U.S. historians search for primary materials at the dawn of the digital age. The American Archivist, 66(Spring-Summer): 9-50.

Tibetan and Himalayan Library. (2009). University of Virginia. Retrieved from http://www.thlib.org/index.php on 1 September 2009.

U.K. Research Council e-Science Programme. (2009). Retrieved from http://www.rcuk.ac.uk/escience/default.htm on 13 August 2009.

UC and the Google Book Settlement: Frequently Asked Questions. (2009). University of California, Office of Scholarly Communication. Retrieved from http://osc.universityofcalifornia.edu/google/faq.html on 3 September 2009.

UK Data Archive. (2009). Retrieved from http://www.data-archive.ac.uk/about/about.asp on 16 August 2009.

University of California Publishing Services. (2009). Retrieved from http://www.ucpress.edu/pubservices/ on 2 September 2009.

Unsworth, J., Courant, P., Fraser, S., Goodchild, M., Hedstrom, M., Henry, C., Kaufman, P. B., McGann, J., Rosenzweig, R. & Zuckerman, B. (2006). Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for Humanities and Social Sciences. American Council of Learned Societies. Retrieved from http://www.acls.org/cyberinfrastructure/cyber.htm on 17 July 2007.

Van House, N. A. (2004). Science and technology studies and information studies. In Cronin, B. (Ed.). Annual Review of Information Science and Technology. Medford, NJ, Information Today. 38: 3-86.

Vectors: Journal of Culture and Technology in a Dynamic Vernacular. (2009). Retrieved from http://www.vectorsjournal.org/ on 8 September 2009.

Whalen, M. (2009). What's wrong with this picture? An examination of art historians' attitudes about electronic publishing opportunities and the consequences of their continuing love affair with print. Art Documentation, 28(2): 13-22.

Wiberley, S. E. (2003). A methodological approach to developing bibliometric models of types of humanities scholarship. Library Quarterly, 73(2): 121-159.

Wiberley, S. E. & Jones, W. G. (1994). Humanists Revisited - A Longitudinal Look At The Adoption Of Information Technology. College & Research Libraries, 55(6): 499-509.

Willinsky, J. (2006). The Access Principle: The Case for Open Access to Research and Scholarship. Cambridge, MA: MIT Press.

Willinsky, J. (2009). Toward the Design of an Open Monograph Press. Journal of Electronic Publishing, 12(1). Retrieved from http://dx.doi.org/10.3998/3336451.0012.103 on 1 April 2009.

R. Rogers (2009). *The end of the virtual - Digital methods*. Amsterdam: Amsterdam University Press.

Vossiuspers UvA

Digital methods may be contrasted with what has come to be known as virtual methods, a currently dominant approach to the study of the Internet. Virtual methods, rooted in the U.K. Virtual Society? program (1997-2002), sought to ground cyberspace by demonstrating how it was hardly a realm apart. Whereas virtual methods have made great strides, they rely on methods imported from the humanities and the social sciences. Do the methods have to change, owing to the specificity of the medium and its objects? With the end of the virtual, Richard Rogers proposes that Internet research may be put to new uses, given an emphasis on natively digital as opposed to digitized methods. How to capture and analyze hyperlinks, tags, search engine results, archived websites, and other digital objects? What may one learn from how online devices make use of the objects, and how may such uses be repurposed for social and cultural research? Ultimately, Rogers proposes a research practice that grounds claims about cultural change and societal conditions in online dynamics.

Richard Rogers is Professor of New Media & Digital Culture at the University of Amsterdam.

The End of the Virtual

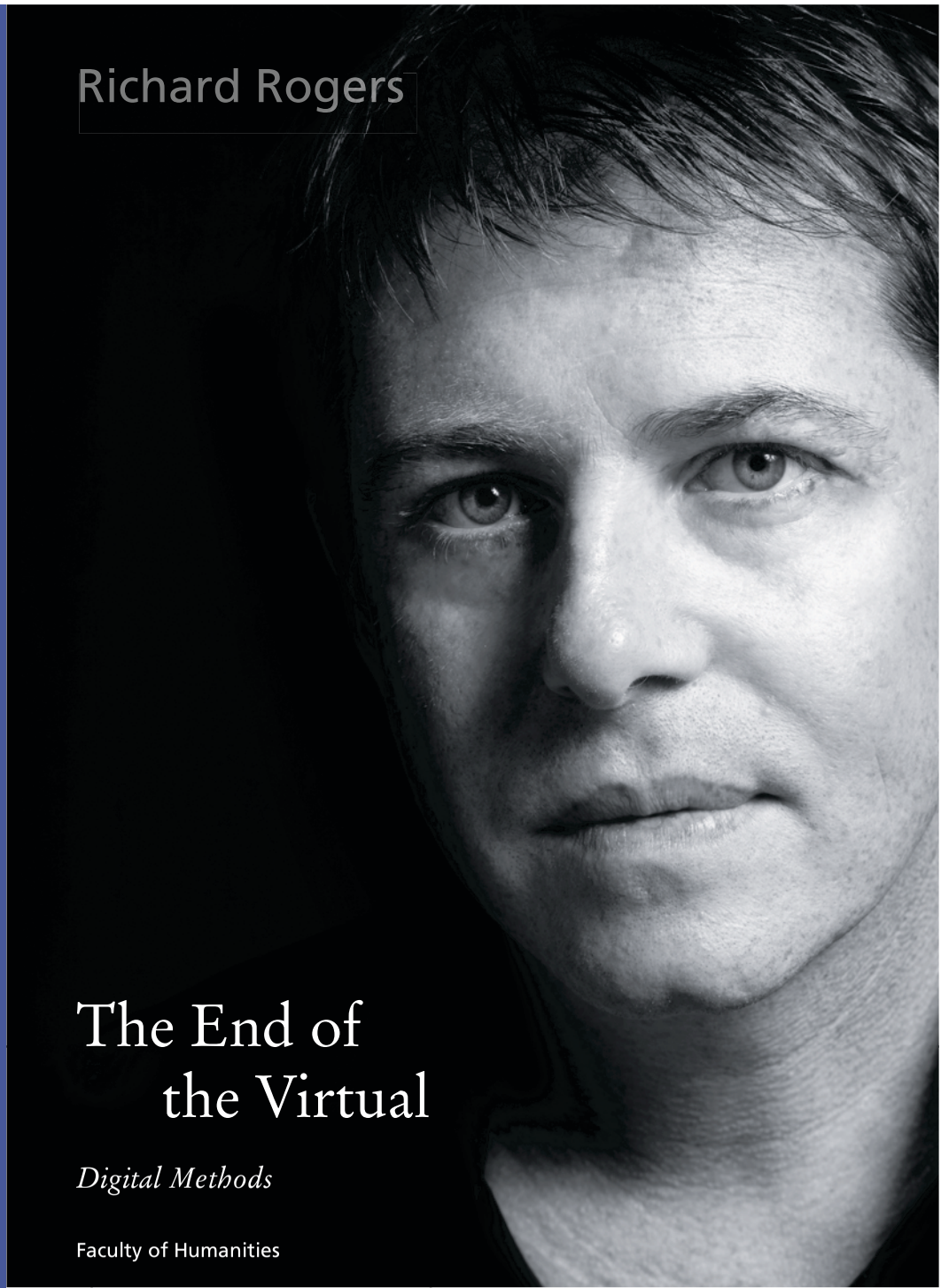
Page 83

9 789 056 295936

Oratiereeks 339

University of Amsterdam

Richard Rogers



The End of the Virtual

Digital Methods

Faculty of Humanities

The End of the Virtual

Vossiuspers UvA is an imprint of Amsterdam University Press.
This edition is established under the auspices of the Universiteit van Amsterdam.
This publication was made possible in part by a grant received from the Mondriaan Interregeling for the Digital Methods Initiative.

Cover design: Nauta & Haagen, Oss
Lay-out: JAPES, Amsterdam
Cover illustration: Carmen Freudenthal, Amsterdam

ISBN 978 90 5629 593 6
e-ISBN 978 90 4851 128 0

The End of the Virtual

Digital Methods

Inaugural Lecture

delivered on the appointment to
the Chair of New Media & Digital Culture
at the University of Amsterdam
on 8 May 2009

by

Richard Rogers

Vossiuspers UvA

Geachte aanwezigen,

Situating Digital Methods in Internet Research

Arguably, there is an ontological distinction between the natively digital and the digitized; that is, between the objects, content, devices and environments 'born' in the new medium, as opposed to those which have 'migrated' to it. Should the current methods of study change, however slightly or wholesale, given the focus on objects and content of the *medium*? The research program proposed here thereby engages with 'virtual methods' importing standard practices from the social sciences and the humanities. The distinction between the natively digital and the digitized also could apply to current research methods. What type of Internet research may be performed with digitized methods (such as online surveys and directories) compared to those that are natively digital (such as recommendation systems and folksonomy)?

Second, I propose that Internet research may be put to new uses, given an emphasis on natively digital methods as opposed to the digitized. I will strive to shift the attention from the opportunities afforded by transforming ink into bits, and instead inquire into how research *with* the Internet may move beyond the study of online culture alone. How to capture and analyze hyperlinks, tags, search engine results, archived websites, and other digital objects? What may one learn from how online devices (e.g. engines and recommendation systems) make use of the objects, and how may such uses be repurposed for social and cultural research? Ultimately, I propose a research practice which grounds claims about cultural change and societal conditions in online dynamics, introducing the term 'online groundedness.' The overall aim is to rework method for Internet research, developing a novel path of study: digital methods.

To date, the methods employed in Internet research have served to critique the persistent idea of the Internet as a virtual realm apart. Such thinking arose from

the discourse surrounding virtual reality in the late 1980s and early 1990s, and the Internet came to stand for a virtual realm, with opportunities for redefining consciousness, identity, corporality, community, citizenry and (social movement) politics.[1] Indeed, in 1999, in one of the first efforts to synthesize Internet research, the communications scholar Steve Jones invited researchers to move beyond the perspective of the Internet as a realm apart, and opened the discussion of method.[2] How would social scientists study the Internet, if they were not to rely on the approaches associated with it to date: human-computer interaction, social psychology and cybercultural studies?[3] In their ground-breaking work on Internet usage in Trinidad and Tobago, the ethnographers Daniel Miller and Don Slater challenged the idea of cyberspace as a realm apart where all 'inhabiting' it experienced its identity-transforming affordances, regardless of physical location.[4] Slater and Miller grounded the Internet, arguing that Trinis appropriated the medium it to fit their own cultural practices. Although it was a case study, the overall thrust of the research was its potential for generalization. If Trinis were using the Internet to stage Trini culture, the expectation is that other cultures are doing the same.

The important *Virtual Society?* program (1997-2002) marked another turning point in Internet research, debunking the myth of cyberspace's transformative capacities through multiple empirical studies about Internet users. The program ultimately formulated five 'rules of virtuality'.[5] In what is now the classic digital divide critique, researchers argued that the use of new media is based on one's situation (access issues), and the fears and risks are unequally divided (skills issues). With respect to the relationship between the real and the virtual, virtual interactions supplement rather than substitute for the 'real,' and stimulate more real interaction, as opposed to isolation and desolation. Finally, the research found that identities are grounded in both the online as well as the offline. Significantly, the program settled on approaches subsequently characterized as virtual methods, with an instrumentarium for studying users. Surveys, interviews, observation and participant-observation became the preferred methods of inquiry. In the humanities, subsequent user studies – concentrating on the amateur, the fan, and the 'produser' – also are grappling with the real and virtual divide, seeking to demonstrate and critique the reputational status of online culture.[6] The argument advanced here is that virtual methods and user studies in the social sciences and the humanities have shifted the attention away from the *data* of the medium, and the opportunities for study of far more than online culture.

How may one rethink user studies with data (routinely) collected by software? User studies to date have relied on accounts favoring observation, interviews and surveys, owing, in one reading, to the difference in armatures between social scientific and humanities computing, on the one hand, and the large commercial companies, with their remarkable data collection achievements, on the other. In a sense, Google, Amazon and many other dominant Web devices are already conducting user studies, however infrequently the term is used. User inputs (preferences, search history, purchase history, location) are captured and analyzed so as to tailor results. Taking a lead from such work, new media theorist Lev Manovich has called for a methodological turn in Internet research, at least in the sense of data collection. With 'cultural analytics,' named after Google Analytics, the proposal is to build massive collection, storage and analytical facilities for humanities computing.[7] One distinguishing feature of the methodological turn is its marked departure from the reliance on (negotiated) access to commercial data sets, e.g. AOL's set of users' search engine queries, Linden Lab's set of the activities of millions of users in Second Life, or Sony's for Everquest, however valuable the findings have been.[8]

In a sense, the research program is one answer to the question, what would Google do? The programs could be situated in the larger context of the extent and effects of 'googlization'. Until now, the googlization critique has examined the growing 'creep' of Google; its business model and its aesthetics, across information and knowledge industries.[9] Library science scholars in particular concern themselves with the changing locus of access to information and knowledge (from public shelves and stacks to commercial servers). The 'Google effect' also may be couched in terms of supplanting surfing and browsing with search. It also may be studied in terms of the demise of the expert editor, and the rise of the back-end algorithm, themes to which I will return later. Here, however, the point is that they also may be studied in terms of models for research – ones that seek to replicate the scale of data collection as well as analysis.

The proposal I am putting forward is more modest, yet still in keeping with what are termed registrational approaches to user studies. Online devices and software installed on the computer (e.g. browsers) register users' everyday usage. Browser histories would become a means to study use. The larger contention is that data collection, in the methodological turn described above, could benefit from thinking about how computing may have techniques which can be appro-

priated for research. Thus the proposal is to consider first and foremost the availability of computing *techniques*.

I would like to suggest inaugurating a new era in Internet research, which no longer concerns itself with the divide between the real and the virtual. It concerns a shift in the kinds of questions put to the study of the Internet. The Internet is employed as a site of research for far more than *just* online culture. The issue no longer is how much of society and culture is online, but rather how to diagnose cultural change and societal conditions using the Internet. The conceptual point of departure for the research program is the recognition that the Internet is not only an object of study, but also a source. Knowledge claims may be made on the basis of data collected and analyzed by devices such as search engines. One of the more remarkable examples is Google Flu Trends, a non-commercial (Google.org) project launched in 2008, which anticipates local outbreaks of influenza by counting search engine queries for flu, flu symptoms and related terms, and 'geo-locating' the places where the queries have been made. It thereby challenges existing methods of data collection (emergency room reports), and reopens the discussion of the Web as anticipatory medium, far closer to the ground than one might expect.[10]

Where did the 'grounded Web,' and its associated geo-locative research practice, originate? The 'end of cyberspace' as a placeless space (as Manuel Castells put it) may be located in the technical outcomes of the famous Yahoo lawsuit, brought by two non-governmental organizations in France in 2000.[13] At the time, French Web users were able to access the Nazi memorabilia pages on Yahoo.com in the United States, and the French organizations wanted the pages blocked – in France. IP-to-geo (address location) technology was developed specifically to channel content nationally; when one types google.com into a browser in France, now google.fr is returned by default. This 'grounding' of the Web has been implemented by major content-organizing projects such as YouTube; online television is served geographically, too.

Diagnostic work such as Google Flu Trends, whereby claims about societal conditions are made on the basis of captured Internet practices, leads to new theoretical notions. For the third era of Internet research, the digital methods program introduces the term *online groundedness*, in an effort to conceptualize research which follows the medium, captures its dynamics, and makes grounded claims about cultural and societal change. Indeed, the broader theoretical goal of digital methods is to rethink the relationship between the Web and the ground. Like the

ethnographers before them, the researchers in the UK *Virtual Society?* program needed to visit the ground in order to study the Web. Here the digital methods research program actually complicates the sequence in which one's findings are grounded.[12] For example, journalism has methodological needs, now that the Internet has become a significant meta-source, where the traditional question normally concerns the trustworthiness of a source. Snowballing from source to source was once a social networking approach to information-checking, methodologically speaking. Who else should I speak to? That question comes at the conclusion of the interview, if trust has been built. The relationship between 'who I should speak to' and 'who else do you link to' is asymmetrical for journalism, but the latter is what search engines ask when recommending information. How to think through the difference between source recommendations from verbal and online links? Is search the beginning of the quest for information, ending with some grounded interview reality beyond the net, whereby we maintain the divide between the real and the virtual? Or is that too simplistic? Our ideal source set divide (real and virtual, grounded or googled) raises the question of what comes next. What do we 'look up' upon conclusion of the interview to check the reality? The Internet may not be changing the hierarchy of sources for some (i.e. the restrictions on citing Wikipedia in certain educational settings), but it may well be changing the order of checking, and the relationship of the Web to the ground.

I developed the notion of online groundedness after reading a study conducted by the Dutch newspaper *NRC Handelsblad*. The investigation into right-wing and extremist groups in the Netherlands explored whether the language used was becoming harsher over time, perhaps indicating a 'hardening' of right-wing and hate culture more generally. Significantly, the investigators elected to use the Internet Archive, over an embedded researcher (going native), or the pamphlets, flyers and other ephemera at the Social History Institute.[13] They located and analyzed the changes in tone over time on right-wing as well as extremist sites, finding that right-wing sites were increasingly employing more extremist language. Thus the findings made about culture were grounded through an analysis of websites. Most significantly, the online became the baseline against which one might judge a societal condition.

RICHARD ROGERS

Follow the Medium: The Digital Methods Research Program

Why follow the medium? A starting point is the recognition that Internet research is often faced with unstable objects of study. The instability is often discussed in terms of the ephemerality of websites and other digital media, and the complexities associated with *fixing* them, to borrow a term from photography. How to make them permanent, so that they can be carefully studied? In one approach, vintage hardware and software are maintained so as to keep the media 'undead.' Another technique, as practiced in game environments, addresses ephemerality through simulation/emulation, which keeps the nostalgic software, like Atari games, running on current hardware. The ephemerality issue, however, is much larger than issues of preservation. The Internet researcher is often overtaken by events of the medium, such as software updates that 'scoop' one's research.

As a research practice, following the medium, as opposed to striving to fix it, may also be discussed in a term borrowed from journalism and the sociology of science – 'scooping.' Being the first to publish is to 'get the scoop.' 'Being scooped' refers to someone else having published the findings first. Sociologist of science Michael Lynch has applied this term to the situation in which one's research subjects come to the same or similar conclusions as the researchers, and go on record with their findings first. The result is that the '[research subjects] reconfigure the field in which we previously thought our study would have been situated'.[14] In Internet research, being scooped is common. Industry analysts, watchdogs and bloggers routinely coin terms (googlization) and come to conclusions which shape ongoing academic work. I would like to argue, however, that scooping is also done by the objects themselves, which are continually reconfigured. For example, Facebook, the social networking site, has been considered a walled garden' or relatively closed community system, where by default only 'friends' can view information and activities concerning other friends. The walled garden is a series of concentric circles: a user must have an account to gain access, must 'friend' people to view their profiles, and must change privacy default settings to let friends of friends view one's own profile. Maximum exposure is opening profiles to friends of friends. In March 2009, Facebook changed a setting; users may now make their profile open to all other users with accounts, as opposed to just friends, or friends of friends, as in its previous configuration. Which types of research would be 'scooped' by Facebook's flipping a switch? Facebook serves as one notable example

of the sudden reconfiguration of a research object, which is common to the medium.

More theoretically, following the medium is a particular form of medium-specific research. Medium specificity is not only how one sub-divides disciplinary commitments in media studies according to the primary objects of study: film, radio, television, etc. It is also a particular plea to take seriously ontological distinctiveness, though the means by which the ontologies are constructed differ. To the literary scholar and media theorist Marshall McLuhan, media are specific in how they engage the senses.[15] Depth, resolution and other aesthetic properties have effects on how actively or passively one processes media. One is filled by media, or one fills it in. To the cultural theorist Raymond Williams, medium specificity lies elsewhere. Media are specific in the forms they assume – forms shaped by the dominant actors to serve interests.[16] For example, creating 'flow,' the term for how television sequences programming so as to keep viewers watching, boosts viewer ratings and advertising. Thus, to Williams, media are not a priori distinct from one another, but can be made so. To Katherine Hayles, the specificity of media resides in their materiality; a book specifies, whilst text does not.[17] Her proposal for media-specific analysis is a comparative media studies program, which takes materially instantiated characteristics of media (such as hypertext in digital media), and enquires into their (simulated) presence in other media (such as print). One could take other media traits and study them across media. For example, as Alexander Galloway has argued, flow is present not only in radio and television, but also on the Web, where dead links disrupt surfing.[18]

Hayle's point of departure may be seen in Mathew Fuller's work on Microsoft Word and Adobe Photoshop, which studies how particular software constrains or enables text.[19] To Fuller, a Microsoft document or a Photoshop image are specific outputs of software, distinctive from some document or some image. An accompanying research program would study the effects of (software) features, as Lev Manovich also points to in his work on the specificity of computer media. With these media Manovich's ontology moves beyond the outputs of media (Hayle's hypertextual print, Fuller's Word document and Photoshop image).[20] Computer media are metamedia in that they incorporate prior media forms, which is in keeping with the remediation thesis put forward by Jay David Bolter and Richard Grusin.[21] Yet, to Manovich, computer media not only refashion the outputs of other media; they also embed their forms of *production*.

The medium specificity put forward here lies not so much in McLuhan's sense engagement, Williams's socially shaped forms, Hayles's materiality, or other theorists' properties and features. Rather, it is situated in method. Previously I described such work 'Web epistemology'.[22] On the Web, information, knowledge and sociality are organized by recommender systems – algorithms and scripts that prepare and serve up orders of URLs, media files, friends, etc. In a sense, Manovich has shifted the discussion in this direction, both with the focus on forms of production (method as craft) as well as with the methodological turn associated with the cultural analytics initiative. I would like to take this turn further, and propose that the under-interrogated methods of the Web also are worthy of study, both in and of themselves as well as in the effects of their spread to other media, e.g. TV shows recommended to Tivo users on the basis of their profiles.

Initial work in the area of Web epistemology arose within the context of the politics of search engines. [23] It sought to consider the means by which sources are adjudicated by search engines. Why, in March of 2003, were the US White House, the Central Intelligence Agency, the Federal Bureau of Investigation, the Heritage Foundation and news organizations such as CNN the top returns for the query 'terrorism'? The answer lies somewhat in how hyperlinks are handled. Hyperlinks, however, are but one digital object, to which may be added: the thread, tag, PageRank, Wikipedia edit, robots.txt, post, comment, trackback, pingback, IP address, URL, whois, timestamp, permalink, social bookmark and profile. In no particular order, the list goes on. The proposal is to study how these objects are handled, specifically, in the medium, and learn from medium method.

In the following, I would like to introduce a series of medium objects, devices, spaces, as well as platforms, first touching briefly on how they are often studied with digitized methods and conceptual points of departure from outside the medium. Subsequently, I would like to discuss the difference it makes to research if one were to follow the medium – by learning from and reapplying how digital objects are treated by devices, how websites are archived, how search engines order information and how geo-IP location technology serves content nationally or linguistically. What kinds of research can be performed through hyperlink analysis, repurposing insights from dominant algorithms? How to work with the Internet archive for social research? Why capture website histories? How may search engine results be studied so as to display changing hierarchies of credibility, and the differences in source reliance between the Web, the blogosphere and news

sphere? Can geo-IP address location technology be reworked so as to profile countries and cultures? How may the study of social networking sites reveal cultural tastes and preferences? How are software robots changing how quality content is maintained on Wikipedia? What would a research bot do? Thus, from the micro to the macro, I treat the hyperlink, website, search engine and spheres (including national webs). I finally turn to social networking sites, as well as Wikipedia, and seek to learn from these profiling and bot cultures (respectively), and rethink how to deploy them analytically. The overall purpose of following the medium is to reorient Internet research to consider the Internet as a source of data, method and technique.

The Link

How is the hyperlink most often studied? There are at least two dominant approaches to studying hyperlinks: hypertext literary theory and social network theory, including small world and path theory.[24] To literary theorists of hypertext, sets of hyperlinks form a multitude of distinct pathways through text. The surfer, or clicking text navigator, may be said to author a story by choosing routes (multiple clicks) through the text.[25] Thus the new means of authorship, as well as the story told through link navigation, are both of interest. For small world theorists, the links that form paths show distance between actors. Social network analysts use pathway thought, and zoom in on how the ties, uni-directional or bi-directional, position actors.[26] There is a special vocabulary that has been developed to characterize an actor's position, especially an actor's centrality, within a network. For example, an actor is 'highly between' if there is a high probability that other actors must pass through him to reach each other.

How do search engines treat links? Arguably, theirs is a scientometric (and associational sociology) approach. As with social network analysis, the interest is in actor positioning, but not necessarily in terms of distance from one another, or the means by which an actor may be reached through networking. Rather, ties are reputational indicators, and may be said to define actor standing. Additionally, the approach does not assume that the ties between actors are friendly, or otherwise have utility, in the sense of providing empowering pathways, or clues for successful networking.

Here I would like to explore how engines treat links as markers of impact and reputation. How may an actor's reputation be characterized by the types of hyperlinks given and received? Actors can be profiled not only through the quantity of links received, as well as the quantity received from others who themselves have received many links, in the basic search engine algorithm. Actors may also be profiled by examining which particular links they give and receive.[27] In previous research, my colleagues and I found linking tendencies among domain types, i.e. governments tend to link to other governmental sites only; non-governmental sites tend to link to a variety of sites, occasionally including critics. Corporate websites tend not to link, with the exception of collectives of them – industry trade sites and industry 'front groups' do link, though. Academic and educational sites typically link to partners and initiatives they have created. Taken together, these linking proclivities of organizational types display an everyday 'politics of association'.[28] For example, in work my colleagues and I conducted initially in 1999, we found that while Greenpeace linked to governmental sites, government did not link back. Novartis, the multinational corporation, linked to Greenpeace, and Greenpeace did not link back. When characterizing an actor according to inlinks and outlinks, one notices whether there is some divergence from the norms, and more generally whether particular links received may reveal something about an actor's reputation. A non-governmental organization receiving a link from a governmental site could be construed as a reputation booster, for example.[29]

Apart from capturing the micro-politics of hyperlinks, analysis of links also may be put to use in more sophisticated sampling work. Here the distinction between digitized and natively digital method stands out in greater relief. The Open Net Initiative at the University of Toronto conducts Internet censorship research by building lists of websites (from online directories such as the Open Directory Project and Yahoo). The researchers subsequently check whether the sites are blocked in a variety of countries. It is important work that sheds light on the scope as well as technical infrastructure of state Internet censorship practices worldwide.[30] In the analytical practice, sites are grouped by category: famous bloggers, government sites, human rights sites, humor, women's rights, etc.; there are approximately forty categories. Thus censorship patterns may be researched by site type across countries.

The entire list of websites checked per country (some 3000) is a sample, covering of course only the smallest fraction of all websites as well as those of a particular subject category. How would one sample websites in a method following the medium, learning from how search engines work (link analysis) and repurposing it for social research? My colleagues and I contributed to the Open Net Initiative work by employing a method which crawls all the websites in a particular category, captures the hyperlinks from the sites, and locates additional key sites (by co-link analysis) that are not on the lists. I dubbed the method 'dynamic URL sampling', in an effort to highlight the difference between manual URL-list compilation, and more automated techniques of finding significant URLs. Once the new sites are found, they are checked for connection stats (through proxies initially, and later perhaps from machines located in the countries in question), in order to determine whether they are blocked. In the research project on 'social, political and religious' websites in Iran, researchers and I crawled all the sites in that ONI category, and through hyperlink analysis, found some thirty previously unknown blocked sites. Significantly, the research was also a page-level analysis (as opposed to host only), with one notable finding being that Iran was not blocking the BBC news front page (as ONI had found), but only its Persian-language page. The difference between the two methods of gathering lists of websites for analysis – manual directory-style work and dynamic URL sampling – shows the contribution of medium-specific method.

The Website

Up until now, investigations into websites have been dominated by user and 'eyeball studies,' where attempts at a navigation poetics are met with such sobering ideas as 'don't make me think'.[31] Many methods for studying websites are located over the shoulder, where one observes navigation or the use of a search engine, and later conducts interviews with the subjects. In what one may term classic registrational approaches, a popular technique is eye-tracking. Sites load and eyes move to the upper left of the screen, otherwise known as the golden triangle. The resulting heat maps provide site redesign cues. For example, Google.com has moved its services from above the search box (tabs) to the top left corner of the page (menu). Another dominant strand of website studies lies in feature analysis,

where sites are compared and contrasted on the basis of levels of interactivity, capacities for user feedback, etc.[32] The questions concern whether a particular package of features result in more users, and more attention. In this tradition, most notably in the 9/11 special collection, websites are often archived for further study. Thus much of the work lies in the archiving of sites prior to the analysis. One of the crucial tasks ahead is further reflection upon the means by which websites are captured and stored, so as to make available the data upon which findings are based. Thus the digital methods research program engages specifically with the website as archived object, made accessible, most readily, through the Internet Archive's Wayback Machine. The research program specifically asks which types of website study are enabled and constrained by the Wayback Machine.

In order to answer that question, the work first deconstructs, or unpacks, the Internet Archive and its Wayback Machine. In which sense does the Internet Archive, as an object formed by the archiving process, embed particular preferences for how it is used, and for the type of research performed using it? Indeed, Web archiving scholar Niels Brügger has written: '[U]nlike other well-known media, the Internet does not simply exist in a form suited to being archived, but rather is first formed as an object of study in the archiving, and it is formed differently depending on who does the archiving, when, and for what purpose.'[33] The idea that the object of study is constructed by the very means by which it is tamed, and captured by method and technique, is a classic point from the sociology and philosophy of science and elsewhere.[34] Thus the initial research questions are, which methods of research are privileged by the specific form assumed by the Web archive, and which are precluded? For example, when one uses the Internet Archive (archive.org), what stands out for everyday Web users accustomed to search engines is not so much the achievement of the existence of an archived Internet. Rather, the user is struck by how the Internet is archived, and, particularly, how it is queried. One queries a URL, as opposed to keywords, and receives a list of stored pages associated with the URL from the past. In effect, the Internet Archive, through the interface of the Wayback Machine, has organized the story of the Web into the histories of individual websites.

Which research approaches are favored by the current organization of websites by the Internet Archive? With the Wayback Machine, one can study the evolution of a single page (or multiple pages) over time; for example, by reading or collect-

ing snapshots from the dates when a page was indexed. How can such an arrangement of historical sites be put to use? Previously I mentioned the investigative reporting work done by *NRC Handelsblad* in their analysis of the rise of extremist language in the Netherlands. The journalists read some hundred websites from the Internet archive, some dating back a decade. It is work that should be built upon, methodologically as well as technically. One could scrape the pages of the right-wing and extremist sites from the Internet Archive, place the text (and images) in a database, and systematically query it for the presence of particular keywords over time. As *NRC Handelsblad* did, one could determine changes in societal conditions through the analysis of particular sets of archived sites.

How else to perform research with the Internet Archive? The digital methods program has developed means to capture the history of sites by taking snapshots and assembling them into a movie, in the style of time-lapse photography.[35] To demonstrate how to use the Internet archive for capturing such evolutionary histories, my colleagues and I took snapshots of Google's front pages from 1998 up to the end of 2007. The analysis concerned the subtle changes made to the interface, in particular the tabs. We found that the Google directory project, organizing the Web by topic, undertaken by human editors, has been in decline. After its placement on the Google front page in 2001, it was moved in 2004 under the 'more' button, and in 2006 under 'even more.' By late 2007, with the removal of the 'even more' option, one had to search Google in order to find its directory.[36] The larger issue of the demise of the human editor, read in this case from the evolution of Google's interface, has far-reaching implications for how knowledge is collected and ordered. Indeed, after examining Google, researchers and I turned to Yahoo, the original Web directory, and found that there, too, the directory had been replaced by the back-end algorithm. In examining the outputs of a query in the directory, we also learned that at Yahoo the results are no longer ordered alphabetically, in the egalitarian style of information and source ordering inherited from encyclopedias. Yahoo is listing its directory sources according to *popularity*, in the well-known style of recommendation systems more generally.

Are the histories of search engines, captured from their interface evolutions, indicating changes in how information and knowledge are ordered more generally? A comparative media studies approach would be useful, with one of the more poignant cases being the online newspaper. With the *New York Times*, for example, articles are still placed on the front page and in sections, but are also listed by

'most emailed' and 'most blogged', providing a medium-specific recommender system for navigating the news. The impact of recommender systems – the dominant means on the Web by which information and knowledge are ordered – may also be studied through user expectations. Are users increasingly expecting Web-like orderings at archives, libraries, tourist information centers and other sites of knowledge and information queries?

The Search Engines & the Spheres

The study of search engines was jolted by the now infamous AOL search engine data release in 2006, where 500,000 users' searches over three months were put online, with frightening and often salacious press accounts about the level of intimate detail revealed about searchers, even if their histories are made anonymous and decoupled from geography (no IP address). One may interpret the findings from the AOL case as a shift in how one considers online presence, if that remains the proper term. A person may be 'googled', and his or her self-authored presence often appears at or towards the top of the returns. Generally speaking, what others have written about a person would appear lower down in the rankings. However, with search engine queries stored, a third set of traces could come to define an individual. This opens up intriguing policy questions. How long may an engine company keep search histories? Thus search engines are being studied in the legal arena, especially in terms of how data retention laws may be applied to search criteria.

Previously, I mentioned another strand in search engine studies, summed up in the term googlization. It is a political-economy style critique, considering how Google's free-service-for-profile model may be spreading across industries and (software) cultures. I have covered the critique elsewhere, striving to propose a research agenda for googlization scholars which includes front-end and back-end googlization. Front-end googlization would include the study of the information politics of the interface (including the demise of the human-edited directory). Back-end googlization concerns the rise of the algorithm that recommends sources hierarchically, instead of alphabetically, as mentioned above. The significance of studying the new information hierarchies of search engines also should be viewed in light of user studies. A small percentage of users set preferences to more than

ten results per page; typically they do not look past the first page of results; and they increasingly click the results appearing towards the top.[37] Thus the power of search engines lies in the combination of its ranking practices (source inclusion in the top results) together with the users' apparent 'respect' for the orderings (not looking further). Google's model also relies on registrational interactivity, where a user's preferences as well as history are registered, stored and employed, increasingly, to serve customized results. Prior to the Web and search engine algorithms and recommendation systems, interactivity was 'consultational,' with pre-loaded information 'called up'.[38] A query would return the same information for all users at any given time. Now the results are dynamically generated, based on one's registered preferences, history and location.

The different orders of sources and things served by engines are under-studied, largely because they are not stored, and made available for research, apart from the AOL data release, or other negotiated agreements with search engine companies. Google once made available an API (application programming interface) allowing data collection. A limited number of queries could be made per day, and the results repurposed. Researchers relying on the API were scooped by Google when it discontinued the service in late 2006. With its reintroduction in a different form in 2009, Google emphasized, however, that automated queries and the permanent storage of results violated the terms of service. How to study search engine results under such conditions? Now we scrape Google, and post a notice appreciating Google's forbearance.[39]

What may be found in Google's search engine results? As I have remarked, search engines, a crucial point of entry to the Web, are epistemological machines in the sense that they crawl, index, cache and ultimately order content. Earlier I described the Web, and particularly a search engine-based Web, as a potential collision space for alternative accounts of reality.[40] The phrasing built upon the work of the sociologist C. Wright Mills, who characterized the purpose of social research as 'no less than to present conflicting definitions of reality itself'.[41] Are engines placing alternative accounts of reality side by side, or do the results align with the official and the mainstream? Storing and analyzing search engine results could answer such questions. Such has been the purpose of the software project called the Issue Dramaturg, so called for the potential drama within the top results, whereby sites may climb to or suddenly fall from the top. It is important to point out that top engine placements are highly sought after; organizations make

use of search engine optimization techniques so as to boost site visibility. There are white hat and black hat techniques; that is, those accepted by engines and those that prompt engines to delist websites from results until there is compliance again with engine etiquette.

In the Issue Dramaturg project, my team stored Google search engine results for the query ', 9/11', as well as other keywords for two purposes. The one is to enquire into source hierarchies, as described above. Which sources are privileged? Which are 'winning' the competition to be the top sources returned for particular queries? The other purpose has been to chart particular sources, in the approach to engine studies I have termed 'source distance'. For the query 9/11, how far from the top of the engine returns are such significant actors as the New York City government and the *New York Times*? Are such sources prominent, or do they appear side by side with sources that challenge more official and familiar views? Apart from the New York City government and the *New York Times,* another actor we have monitored is the 9/11 truth movement (911truth.org). For months between March and September 2007, the 9/11 truth movement's site appeared in the top five results for the query 9/11, and the other two were well below result fifty. In mid-September 2007, around the anniversary of the event, there was drama. 911truth.org fell precipitously to result two hundred, and subsequently out of the top one thousand, the maximum number of results served by Google. We believe that is one of the first fully documented cases of the apparent removal of a website in Google – from a top five placement for six months to a sub-one thousand ranking.[42] The case leads to questions of search engine result stability and volatility, and opens up an area of study.

However dominant it may be, there are more search engines than Google's Web search. What is less appreciated perhaps is that there are other dominant engines per section or sphere of the Web. For the blogosphere, there is Technorati; for the newssphere, Google News; and for the tagosphere or social bookmarking space, Delicious. Indeed, thinking of the Web in terms of spheres refers initially to the name of one of the most well-known, the blogosphere, as well as to scholarship that seeks to define another realm, the Web sphere.[43] The sphere in blogosphere refers in spirit to the public sphere; it also may be thought of as the geometrical form, where all points on the surface are the same distance from the center or core. One could think about such an equidistant measure as an egalitarian ideal, where every blog, or even every source of information, is knowable by

the core, and vice versa. On the Web, however, it has been determined that certain sources are central. They receive the vast majority of links as well as hits. Following such principles as the rich get richer (aka Pareto power law distributions), the sites receiving attention tend to garner only more. The distance between the center and other nodes may only grow, with the ideal of a sphere being a fiction, though a useful one. I would like to suggest an approach examining the question of distance from core to periphery, and operationalize it as the measure of differences in rankings between sources per sphere. Cross-spherical analysis is a digital method for measuring and learning from the distance between sources in different spheres on the Web.

Conceptually, a sphere is considered to be a device demarcated source set, i.e. the pure PageRank of all sources on the Web (most influential sites by inlink count), or indeed analogous pageranks of all sources calculated by the dominant engines per sphere, such as Technorati, Google News and Delicious. Thus, to study a sphere, we propose first to allow the engines to demarcate it. In sphere analysis, one considers which sources are most influential, not only overall but per query. Cross-spherical analysis compares the sources returned by each sphere for the same query. It can therefore be seen as comparative ranking research. Most importantly, with cross-spherical analysis, one may think through the consequences of each engine's treatment of links, freshness, tags, etc. Do particular sources tend to be in the core of one sphere, and not in others? What do comparisons between sources, and source distances, across the spheres tell us about the quality of the new media? What do they tell us about current informational commitments in particular cultures?

In a preliminary analysis, my colleagues and I studied which animals are most associated with climate change on the (English-language) Web, in the news and in the blogosphere. We found that the Web has the most diverse set of animals associated with climate change. The news favored the polar bear, and the blogosphere amplified, or made more prominent, the selection in the news sphere. Here we cautiously concluded that the Web may be less prone to the creation of media icons than the news, which has implications for studies of media predicated upon a publicity culture. The blogosphere, moreover, appeared parasitically connected to the news as opposed to providing an alternative to it.

The Webs

As mentioned above, Internet research has been haunted by the virtual/real divide. One of the reasons for such a divide pertains to the technical arrangements of the Internet, and how they became associated with a virtual realm, cyberspace. Indeed, there was meant to be something distinctive about cyberspace, technologically.[44] The protocols and principles, particularly packet switching and the end-to-end principle, initially informed the notion of cyberspace as a realm free from physical constraints. The Internet's technical indifference to the geographical location of its users spawned ideas not limited to placeless-ness. In its very architecture, the Internet also supposedly made for a space untethered to the nation-states, and their divergent ways of treating flows of information. One recalls the famous quotation attributed to John Gilmore, co-founder with John Perry Barlow of the Electronic Frontier Foundation. 'The Internet treats censorship as a malfunction, and routes around it'.[45] Geography, however, was built into cyberspace from the beginning, if one considers the locations of the original thirteen root servers, the unequal distributions of traffic flows per country, as well as the allotment of IP addresses in ranges, which later enabled the application of geo-IP address location technology to serve advertising and copyright needs. Geo-IP technology, as well as other technical means (aka locative technology), also may be put to use for research that takes the Internet as a site of study, and inquires into what may be learned about societal conditions across countries. In the digital methods research program, my colleagues and I have dubbed such work national Web studies.

Above I discussed the research by British ethnographers, who grounded cyberspace through empirical work on how Caribbean Internet users appropriated the medium to fit their own cultural practices. This is of course national Web studies, although with observational methods (from outside of the medium). To study the Web, nationally, one also may inquire into routinely collected data, for example by large enterprises such as Alexa's top sites by country (according to traffic). Which sites are visited most frequently per country, and what does site visitation say about a country's informational culture? Alexa pioneered registrational data collection with its toolbar, which users installed in their browsers. The toolbar provided statistics about the Website loaded in the browser, such as its freshness. All websites the user loaded, or surfed, also would be logged, and the logged URLs

would be compared with the URLs already in the Alexa database. Those URLs not in the database would be crawled, and fetched. Thus was born the Internet Archive.

The Internet Archive (1996-) was developed during the period of Internet history that one could term cyberspace. (I have developed periodizations of Internet history elsewhere, and will not further elaborate here.)[46] To illustrate the design and thought behind the Internet Archive, and the national Web archives sprouting up in many countries, it may be useful to point out that the Internet Archive was built for surfing – an Internet usage type that arguably has given way to search.[47] At the Wayback Machine of the Internet Archive, type in a single URL, view available pages, and browse them. If one reaches an external link, the Internet Archive looks up the page closest in date to the site one is exiting, and loads it. If no site exists in the Internet Archive, it connects to the live website. It is the continuity of flow, from Website to Website, that is preserved.[48] National Web archives, on the other hand, have ceased to think of the Web in terms of cyberspace. Instead, their respective purposes are to preserve national Webs. For the purposes of contributing method to Internet research, the initial question is, how would one demarcate a national Web?

At the National Library in the Netherlands, for example, the approach is similar to that of the Internet censorship researchers, discussed above. It is a digitized method, that is, a directory model, where an expert chooses significant sites based on editorial criteria. These sites are continually archived with technology originally developed in the Internet Archive project. At the time of writing, approximately one thousand national websites are archived in the Netherlands – a far cry from what is saved in the Internet Archive.[49] In accounting for the difference in approaches and outcomes of the two projects, I would like to observe that the end of the virtual, and the end of cyberspace, have not been kind to Web archiving; the return of the nation-state and the application of certain policy regimes (especially copyright) have slowed efforts dramatically. Would digital methods aid in redressing the situation? I would like to invite national Web archivists to consider a registrational approach, e.g. the Alexa model adapted for a national context.

Social Networking Sites & Post-demographics

'We define social networking websites here as sites where users can create a profile and connect that profile to other profiles for the purposes of making an explicit personal network.'[50] Thus begins the study of American teenage use of such sites as MySpace and Facebook, conducted for the Pew Internet & American Life Project. Surveys were taken. 91% of the respondents use the sites to 'manage friendships'; less than a quarter use the sites to 'flirt'. Other leading research into social networking sites considers such issues as presenting oneself and managing one's status online, the different 'social classes' of users of MySpace and Facebook, and the relationship between real-life friends and 'friended' friends.[51] Another set of work, often from software-making arenas, concerns how to make use of the copious amounts of data contained in online profiles, especially interests and tastes. I would like to dub this latter work 'post-demographics.' Post-demographics could be thought of as the study of the data in social networking platforms, and, in particular, how profiling is, or may be, performed. Of particular interest here are the potential outcomes of building tools on top of profiling platforms. What kinds of findings may be made from mashing up the data, or what may be termed meta-profiling?

Conceptually, with the 'post' prefixed to demographics, the idea is to stand in contrast to how the study of demographics organizes groups, markets and voters in a sociological sense. It also marks a theoretical shift from how demographics have been used 'bio-politically' (to govern bodies) to how post-demographics are employed 'info-politically,' to steer or recommend certain information to certain people.[52] The term post-demographics also invites new methods for the study of social networks, where the traditional demographics of race, ethnicity, age, income, and educational level – or derivations thereof such as class – give way to tastes, interests, favorites, groups, accepted invitations, installed apps and other information comprising an online profile and its accompanying baggage. That is, demographers normally would analyze official records (births, deaths, marriages) and survey populations, with census-taking being the most well known of those undertakings. Profilers, however, have users input data themselves in platforms that create and maintain social relations. They capture and make use of information from users of online platforms.

Perhaps another means of distinguishing between the two types of thought and practice is with reference to the idea of digital natives, those growing up with online environments, and unaware of life prior to the Internet, especially with the use of manual systems that came before it, like a library card catalog.[53] The category of digital natives, however, takes a generational stance, and in that sense is a traditional demographic way of thinking. The post-demographic project would be less interested in new digital divides (digital natives versus non-natives) and the emergent narratives surrounding them (e.g. moral panics), but rather in how profilers recommend information, cultural products, events or other people (friends) to users, owing to common tastes, locations, travel destinations and more. There is no end to what *could* be recommended, if the data are rich and stored. How to study the data?

With post-demographics, the proposal is to make a contribution to Internet research by learning from those profilers and researchers who both collect as well as harvest (or scrape) social networking sites' data for further analysis or software-making, such as mash-ups. How do social networking sites make their data available to profilers? Under the developers' menu item at Facebook, for example, one logs in and views the fields available in the API (or application programming interface). Sample scripts are provided, as in 'get friends of user number x,' where x is yourself. Thus the available scripts generally follow the privacy culture, in the sense that the user decides what the profiler can see. It becomes more interesting to the profiler when many users allow access, by clicking 'I agree' on a third-party application.

Another set of profiling practices are not interested in personal data per se, but rather in tastes and especially taste relationships. One may place many profiling activities in the category of depersonalized data analysis, including Amazon's seminal recommendation system, where it is not highly relevant which person also bought a particular book, but rather that people have done so. Supermarket loyalty cards and the databases storing purchase histories similarly employ depersonalized information analysis, where like Amazon, of interest is the quantity of particular items purchased as well as the purchasing relationships (which chips with which soft drink). Popular products are subsequently boosted. Certain combinations may be shelved together.

While they do not describe themselves as such, of course the most significant post-demographic machines are the social networking platforms themselves, col-

lecting user tastes, and showing them to others, be they other friends, everyday peoplewatchers or profilers. Here I would like to describe briefly one piece of software my research team built on top of the large collection device, MySpace, and the kinds of post-demographic analytical practices which resulted.

Elfriendo.com is the outcome of reflecting on how to make use of the profiles on the social networking platform, MySpace. At Elfriendo.com, enter a single interest, and the tool creates a new profile on the basis of the profiles of people expressing that single interest. One may also compare the compatibility of interests, i.e. whether one or more interests, tunes, movies, TV shows, books and heroes are compatible with other ones. Is Christianity compatible with Islam, in the sense that those people with one of the respective interests listen to the same music and watch the same television programs? Elfriendo answers those sorts of questions by analyzing sets of friends' profiles, and comparing interests across them. Thus a movie, TV show, etc. has an aggregate profile, made up of other interests. (To wit, Eminem, the rapper, appears in both the Christianity and Islam aggregate profiles, in early February 2009.) One also may perform a semblance of post-demographic research with the tool, gaining an appreciation of relational taste analysis with a social networking site, more generally.[54]

It is instructive to state that MySpace is more permissive and less of a walled garden than Facebook, in that it allows the profiler to view a user's friends (and his/her friends' profiles), without you having friended anybody. Thus, one can view all of Barack Obama's friends, and their profiles. Here, in an example, one queries Elfriendo for Barack Obama as well as John McCain, and the profiles of their respective sets of friends are analyzed. The software counts the items listed by the friends under interests, music, movies, TV shows, books and heroes. What does this relational taste counting practice yield? The results provide distinctive pictures of the friends of the two presidential candidates campaigning in 2008. The compatibility level between the interests of the friends of the two candidates is generally low. The two groups share few interests. The tastes of the candidates' friends are not compatible for movies, music, books and heroes, though for TV shows the compatibility is 16%. There seem to be particular media profiles for each set of candidate's friends, where those of Obama watch the Daily Show, and those of McCain watch Family Guy, Top Chef and America's Next Top Model. Both sets of friends watch Lost. The findings may be discussed in terms of voter post-demographics, in that the descriptions of voter profiles are based on media

tastes and preferences as opposed to educational levels, income and other standard indicators.

Wikipedia & Networked Content

At present, approaches to the study of Wikipedia have followed from certain qualities of the online encyclopedia, all of which appear counter-intuitive at first glance. One example is that Wikipedia is authored by so-called amateurs, yet is surprisingly encyclopedia-like, not only in form but in accuracy.[55] The major debate concerning the quality of Wikipedia vis-à-vis *Encyclopedia Britannica* has raised questions relevant to digital methods, in that the Web-enabled collective editing model has challenged the digitized work of a set of experts. However, research has found that there is only a tiny ratio of editors to users in Web 2.0 platforms, including Wikipedia. This is otherwise known as the myth of user-generated content.[56] Wikipedia co-founder Jimbo Wales, has often remarked that the dedicated community is indeed relatively small, at just over 500 members. Thus the small cadre of Wikipedia editors could be considered a new elite, leading to exercises in relativizing the alleged differences between amateurs and experts, such as through a study of the demographics of Wikipedians.[57] Another example of a counter-intuitive aspect of Wikipedia is that the editors are unpaid, yet committed and highly vigilant. The vigilance of the crowd, as it is termed, is something of a mythical feature of a quality-producing Web, until one considers how vigilance is performed. Who is making the edits? One approach to the question lies in the Wikiscanner project (2007-), developed by Virgil Griffith studying at the California Institute of Technology. The Wikiscanner outs anonymous editors by looking up the IP address of the editor and checking it against a database with the IP address locations (geoIP technology). Wikipedia quality is ensured, to Griffith, by scandalizing editors making self-serving changes, such as a member of the Dutch royal family, who embellished an entry and made the front-page of the newspaper after a journalist used the tool.

How else are vandals kept at bay on Wikipedia, including those experimenters and researchers making erroneous changes to an entry, or creating a new fictional one, in order to keep open the debate about quality?[58] Colleagues and I have contributed to work about the quality of Wikipedia by introducing the term net-

worked content.[59] It refers to content held together by human authors and non-human tenders, including bots and alert software which revert edits or notify Wikipedians of changes made. Indeed, when looking at the statistics available on Wikipedia on the number of edits per Wikipedian user, it is remarkable to note that the bots are by far the top editors. The contention, which is being researched in the digital methods program, is that the bots and the alert software are significant agents of vigilance, maintaining the quality of Wikipedia.

From the Wikiscanner project and the bots statistics related above, it is worth emphasizing that Wikipedia is a compendium of network activities and events, each logged and made available as large data sets. Wikipedia also has in-built reflection or reflexivity, as it shows the process by which an entry has come into being, something missing from encyclopedias and most other *finished* work more generally. One could study the process by which an entry matures; the materials are largely the revision history of an entry, but also its discussion page, perhaps its dispute history, its lock-downs and re-openings. Another approach to utilizing Wikipedia data would rely on the edit logs of one or more entries, and repurpose the Wikiscanner's technical insights by looking up where they have been made. 'The places of edits' show subject matter concerns and expertise by organization and by country.

Conclusion. The End of the Virtual – Grounding Claims Online

My aim is to set into motion a transformation in how and why one performs research using the Internet. The first step is to move the discussion away from the limitations of the virtual (how much culture and society are online) to the limitations of current method (how to study culture and society, and ground findings with the Internet).

I would like to conclude with a brief discussion of these limitations in Internet research as well as a proposal for renewal. First, the end of cyberspace and its placeless-ness, and the end of the virtual as a realm apart, are lamentable for particular research approaches and other projects. In a sense, the real/virtual divide served specific research practices.[60] Previously I mentioned that Internet archiving thrived in cyberspace, and more recently, it suffers without it. Where

cyberspace once enabled the idea of massive website archiving, the grounded Web and the national Webs are shrinking the collections.

Indeed, I have argued that one may learn from the methods employed in the medium, moving the discussion of medium specific theory from ontology (properties and features) to epistemology (method). The Internet, and the Web more specifically, have their ontological objects, such as the link and the tag. Web epistemology, among other things, is the study of how these natively digital objects are handled by devices. The insights from such a study lead to important methodological distinctions, as well as insights about the purpose of Internet research. Where the methodological distinction is concerned, one may view current Internet methods as those that follow the medium (and the dominant techniques employed in authoring and ordering information, knowledge and sociality) and ones that remediate or digitize existing method. The difference in method may have significant outcomes. One reason for the fallowing of the Web archiving efforts may lie in the choice of a digitized method (editorial selection) over a digital one (registrational data collection), such as that employed in the original Internet Archive project, where sites surfed by users were recorded. Indeed, I have employed the term digital methods so that researchers may consider the value and the outcomes of one approach over another. As a case in point, the choice of dynamic URL sampling over the editorial model could be beneficial to Internet censorship research, as I discussed.

Third, and finally, I have argued that the Internet is a site of research for far more than online culture and its users. With the end of the virtual/real divide, however useful, the Internet may be rethought as a source of data about society and culture. Collecting it and analyzing it for social and cultural research requires not only a new outlook about the Internet, but method, too, to ground the findings. Grounding claims in the online is a major shift in the purpose of Internet research, in the sense that one is not so much researching the Internet, and its users, as studying culture and society *with the Internet*. I hope you will join me in this urgent project.

Ik heb gezegd.

Notes

1. Barlow, 1996; Benedict, 1991; Dibbell, 1998; Rheingold, 1991; Rheingold; 1993; Shaviro, 2008; Stone, 1995; Turkle, 1995.
2. Jones, 1999.
3. Hine, 2000.
4. Slater & Miller, 2000.
5. Woolgar, 2002.
6. Jenkins, 2006; Keen, 2007; Bruns, 2008.
7. Manovich, 2007. See also Manovich, 2008; Lazer et al., 2009.
8. Manovich, 2008.
9. Jeanneney, 2007; Vaidhyanathan, 2007; Rogers, 2009.
10. Rogers, 2003.
11. Castells, 1996; Goldsmith & Wu, 2006; Rogers, 2008.
12. Marres & Rogers, 2008.
13. *NRC Handelsblad*, 2007.
14. Lynch, 1997.
15. McLuhan, 1964.
16. Williams, 1974.
17. Hayles, 2004.
18. Galloway, 2004.
19. Fuller, 2003.
20. Manovich, 2008.
21. Bolter & Grusin, 1999.
22. Rogers, 2004.
23. Introna & Nissenbaum, 2000.
24. Landow, 1994; Watts, 1999; Park & Thewall, 2003
25. Elmer, 2001.
26. Krebs, 2002.
27. cf. Beaulieu, 2005.
28. Marres & Rogers, 2000; Rogers, 2002.
29. The Issue Crawler software, with particular allied tools, has been developed specifically to perform such hyperlink analysis. The software crawls websites, and links are gathered and stored. The crawler-analytical modules are adaptations from scientometrics (co-link analysis) and social networking analysis (snowball). Once a network is located with the Issue Crawler, individual actors may be profiled, using the actor profiler tool. The actor profiler shows, in a graphic representation, the inlinks and outlinks of the top ten network actors. The other technique for actor profiling relies on a

scraper that would capture all outlinks from a site, and a scraper of a search engine, the Yahoo inlink ripper, which provides a list of the links made to a website.

30. Diebert et al, 2006.
31. Krug, 2000; Dunne, 2005.
32. Foot & Schneider, 2006.
33. Brügger, 2005, 1.
34. Latour & Woolgar, 1986; Knorr-Cetina, 1999; Walker, 2005.
35. Screen-capturing software has been employed previously for the analysis of Wikipedia pages, showing the evolution of entries and thus how Wikipedians build knowledge.
36. The 'even more' button returned to the interface of Google.com in 2008.
37. Spink & Jansen, 2004.
38. Jensen, 1999.
39. The notice appears on the credits page of the Issue Dramaturg, http://issuedramaturg.issuecrawler.net/.
40. Rogers, 2004.
41. C. Wright Mills, 1971, 212; Rogers & Marres, 2002.
42. Rogers, 2009.
43. Foot & Schneider, 2002; Schneider & Foot, 2002.
44. Chun, 2006.
45. Boyle, 1997.
46. Rogers, 2008.
47. Shirky, 2005.
48. Galloway, 2004.
49. See Weltevrede, 2009.
50. Lenhart & Madden, 2007.
51. Boyd & Ellison, 2007.
52. Foucault, 1998; Rogers, 2004.
53. Prensky, 2001.
54. One gains a sense of how analysis may be performed, and the kinds of findings that may be made, because Elfriendo captures the top 100 profiles, thus providing an indication, as opposed to a grounded finding from a proper sampling procedure.
55. Giles, 2005.
56. Swartz, 2006.
57. Van Dijck, 2009.
58. Chesney, 2006; Read, 2006; Magnus, 2008.
59. Niederer, 2009.
60. For the edits may be traced.

References

Barlow, J. P., 'A Declaration of the Independence of Cyberspace,' Davos, Switzerland, 1996, http://homes.eff.org/~barlow/Declaration-Final.html (accessed 28 January 2009)

Beaulieu, A., 'Sociable Hyperlinks: An Ethnographic Approach to Connectivity', in: C. Hine (ed.), *Virtual Methods: Issues in Social Research on the Internet*. Berg, Oxford, 2005, pp. 183-197

Benedict, M., 'Cyberspace: Some Proposals', in: M. Benedict (ed.), *Cyberspace – First Steps*. Cambridge: MIT Press, Cambridge, MA, 1991, pp. 119-224

Bolter, J. D. and R. Grusin, *Remediation: Understanding New Media*. MIT Press, Cambridge, MA, 1999

Boyd, D. and N. Ellison, 'Social network sites: Definition, history, and scholarship,' in: *Journal of Computer-Mediated Communication*, 13(1), 2007

Boyle, J., 'Foucault in Cyberspace', in: *Univ. Cincinnati Law Review*, 66, 1997, pp. 177-205

Brügger, N., *Archiving Websites: General Considerations and Strategies*. Centre for Internet Research, Aarhus, 2005

Bruns, A., *Blogs, Wikipedia, Second Life, and Beyond: From Production to Producage*. Peter Lang, New York, 2008

Castells, M., *The Information Age: Economy, Society and Culture – The Rise of the Network Society*. Blackwell, Malden, MA, 1996

Chesney, T., 'An empirical examination of Wikipedia's credibility', in: *First Monday*, 11(11), 2006

Chun, W., *Control and Freedom: Power and Paranoia in the Age of Fiber*. MIT Press, Cambridge, MA, 2006

Contractor, N., 'Digital Traces: An Exploratorium for Understanding and Enabling Social Networks', presentation at the annual meeting of the American Association for the Advancement of Science (AAAS), 2009

Dibbell, J., *My Tiny Life: Crime and Passion in a Virtual World*. Henry Holt, New York, 1998

Diebert, R., J. Palfrey, R. Rohozinski, and J. Zittrain (eds.), *Access Denied: The practice and policy of global Internet filtering*. MIT Press, Cambridge, MA, 2008

van Dijck, J., 'Users Like You: Theorizing Agency in User-Generated Content', in: *Media, Culture and Society*, 31(1), 2009, pp. 41-58

Dunne, A., *Hertzian Tales: Electronic Products, Aesthetic Experience, and Critical Design*. MIT Press, Cambridge, MA, 2005

Elmer, G., 'Hypertext on the Web: The Beginnings and Ends of Web Path-ology', in: *Space and Culture*, 10, 2001, pp. 1-14

Elmer, G., *Profiling Machines*. MIT Press, Cambridge, MA, 2004

Foot, K. and S. Schneider, 'Online Action in Campaign 2000: An Exploratory Analysis of the U.S. Political Web Sphere, in: *Journal of Broadcast and Electronic Media*, 46(2), 2002, pp. 222-244

Foot, K. and S. Schneider, *Web Campaigning*. Cambridge, MA: MIT Press

Foucault, M., *The History of Sexuality Vol.1: The Will to Knowledge*. Penguin, London, 1998

Fuller, M., *Behind the Blip: Essays on the Culture of Software*. Autonomedia, Brooklyn, 2003

Galloway, A., *Protocol: How Control Exists After Decentralization*. MIT Press, Cambridge, MA, 2004

Giles, J., 'Internet encyclopedias go head to head', in: *Nature*, 438, 2005, pp. 900-901

Goldsmith, J. and T. Wu, *Who Controls the Internet? Illusions of a Borderless World*. Oxford, New York, 2006

Hayles, K., 'Print Is Flat, Code Is Deep: The Importance of Media-Specific Analysis', *Poetics Today*, 25(1), 2004, pp. 67-90

Hine, C., *Virtual Ethnography*. Sage, London, 2000

Hine, C. (ed.), *Virtual Methods: Issues in Social Research on the Internet*. Berg, Oxford, 2005

Introna, L. and H. Nissenbaum, 'Shaping the Web: Why the Politics of Search Engines Matters', *The Information Society*, 16(3), 2000, pp. 1-17

Jeanneney, J.-N., *Google and the Myth of Universal Knowledge*. University of Chicago Press, Chicago, 2007

Jenkins, H., *Convergence Culture: Where Old and New Media Collide*. NYU Press, New York, 2006

Jensen, J., 'Interactivity: Tracking a New Concept in Media and Communication Studies', in: P. Mayer (ed.), *Computer Media and Communication*. Oxford University Press, Oxford, 1999, pp. 160-188

Jones, S., 'Studying the Net: Intricacies and Issues.' in: S. Jones (ed.), *Doing Internet Research: Critical Issues and Methods for Examining the Net*. Sage, London, 1999, pp. 1-28

Keen, A., *The Cult of the Amateur: How Today's Internet is Killing Our Culture*. Nicholas Brealey, London, 2007

Knorr-Cetina, K., *Epistemic Cultures*. Harvard University Press, Cambridge, MA, 1999

Krebs, V., 'Mapping Networks of Terrorist Cells', in: *Connections*, 24(3), 2002, 43-52

Krug, S., *Don't Make Me Think! A Common Sense Approach to Web Usability*. New Riders, Indianapolis, IN, 2000

Landow, G., *Hyper/Text/Theory*. Johns Hopkins University Press, Baltimore, MD, 1994

Latour, B. and S. Woolgar, *Laboratory Life*. Princeton University Press, Princeton, NJ, 1986

Lazer, D. et al., 'Computational Social Science', in: *Science*, 323, 2009, pp. 721-723

Lenhart, A. and M. Madden, 'Social Networking Websites and Teens', Pew Internet Project Data Memo, Pew Internet & American Life Project, Washington, DC, 2007

Lynch, M., 'A sociology of knowledge machine'. in: *Ethnographic Studies*, 2, 1997, pp. 16-38

Magnus, P.D., 'Early response to false claims in Wikipedia', in: *First Monday*, 13(9), 2008

Manovich, L., 'Cultural Analytics.' unpublished ms., www.manovich.net/cultural_analytics.pdf (accessed 28 January 2009)
Manovich, L., *Software Takes Command*. unpublished ms., www.manovich.net/ (accessed 10 April 2009)
Marres, N. and R. Rogers, 'Depluralising the Web, Repluralising Public Debate. The GM Food Debate on the Web,' in: R. Rogers (ed.), *Preferred Placement*. Jan van Eyck Editions, Maastricht, 2000, pp. 113-135
Marres, N. and R. Rogers, 'Subsuming the Ground: How Local Realities of the Ferghana Valley, Narmada Dams and BTC Pipeline are put to use on the Web', *Economy & Society*, 37(2), 2008, pp. 251-281
McLuhan, M., *Understanding Media: The Extensions of Man*. McGraw Hill, New York, 1964
Miller, D. and D. Slater, *The Internet: An Ethnographic Approach*. Berg, Oxford, 2000
Mills, C. Wright, *The Sociological Imagination*. Penguin, Harmondsworth, 1971
Niederer, S., 'Wikipedia and the Composition of the Crowd,' unpublished ms., 2009
NRC Handelsblad. 28 August 2007
Park, H. and M. Thewall, 'Hyperlink Analyses of the World Wide Web: A Review', in: *Journal of Computer-Mediated Communication*, 8(4), 2003
Prensky, M., 'Digital Natives, Digital Immigrants', *On the Horizon*. 9(5), 2001
Read, B., 'Can Wikipedia Ever Make the Grade?' *Chronicle of Higher Education*, 53(10), 2006, p. A31
Reingold, H., *Virtual Reality: Exploring the Brave New Technologies*. Summit, New York, 1991
Rheingold, H., *The Virtual Community: Homesteading on the Electronic Frontier*. Addison-Wesley, Reading, MA, 1993
Rogers, R., 'Operating Issue Networks on the Web,' in: *Science as Culture*, 11(2), 2002, pp. 191-214
Rogers and Marres, N., 'French scandals on the Web, and on the streets: A small experiment in stretching the limits of reported reality', in: *Asian Journal of Social Science*, 30 (2), 2002, pp. 339-353
Rogers, R., 'The Viagra Files: The Web as Anticipatory Medium', in: *Prometheus*, 21(2), 2003, pp. 195-212
Rogers, R., *Information Politics on the Web*. MIT Press, Cambridge, MA, 2004
Rogers, R., 'The Politics of Web Space,' unpublished ms., 2008
Rogers, R., 'The Googlization Question, and the Inculpable Engine', in: Stalder, F. and K. Becker (eds.), *Deep Search: The Politics of Search Engines*. Edison, NJ: Transaction Publishers, 2009
Schneider, S. and K. Foot, 'Online structure for political action: Exploring presidential Web sites from the 2000 American election', *Javnost*, 9(2), 2002, pp. 43-60
Shaviro, S., 'Money for Nothing: Virtual Worlds and Virtual Economies', in: M. Ipe (ed.), *Virtual Worlds*. The Icfai University Press, Hyderabad, 2008, pp. 53-67.

Shirky, C., 'Ontology is Overrated: Categories, Links, and Tags', The Writings of Clay Shirky, 2005, www.shirky.com/writings/ontology_overrated.html (accessed 28 January 2009)

Spink, A. and B.J. Jansen, *Web Search: Public Searching on the Web*. Kluwer, Dordrecht, 2004

Stone, A.R., *The War of Desire and Technology at the Close of the Mechanical Age*. MIT Press, Cambridge, MA, 1995

Sunstein, C., *Infotopia: How Many Minds Produce Knowledge*. Oxford University Press, New York, 2006

Swartz, A., 'Who writes Wikipedia?' Raw Thoughts blog entry, 4 September 2006, www.aaronsw.com/weblog/whowriteswikipedia/ (accessed 22 August 2008)

Turkle, S., *Life on the Screen: Identity in the Age of the Internet*. Simon & Schuster, New York, 1995

Vaidhyanathan, S., 'Where is this book going?' The Googlization of Everything Blog, 25 September 2007, www.googlizationofeverything.com/2007/09/where_is_this_book_going.php (accessed 22 December 2008)

Walker, J., 'Feral Hypertext: When Hypertext Literature Escapes Control', *Proceedings of the Sixteenth ACM conference on Hypertext and Hypermedia*, 6-9 September 2005, Salzburg, Austria, pp. 46-53

Watts, D., *Small Worlds*. Princeton University Press, Princeton, 1999

Weltevrede, E., *Thinking Nationally with the Web: A Medium-Specific Approach to the National Turn in Web Archiving*. M.A thesis, University of Amsterdam, 2009

Williams, R., *Television: Technology and Cultural Form*. Fontana, London, 1974

Woolgar, S., 'Five Rules of Virtuality, in: S. Woolgar (ed.), *Virtual Society? Technology, Cyberbole, Reality*. Oxford University Press, Oxford, 2002, pp. 1-22

Manovich, L. (2009), “How to Follow Global Digital Cultures, or Cultural Analytics for Beginners,” in K. Becker and F. Stalder (eds.) *Deep Search: The Politics of Search beyond Google*. Innsbruck: Studienverlag, 198-212.

Lev Manovich

How to Follow Global Digital Cultures, or Cultural Analytics for Beginners

From "New Media" to "More Media"

Only fifteen years ago we typically interacted with relatively small bodies of information that were tightly organized in directories, lists and a priori assigned categories. Today we interact with a gigantic, global, not well organized, constantly expanding and changing information cloud in a very different way: we Google it.

The raise of search as the new dominant way for encountering information is one manifestation of the fundamental change in human's information environment.[1] We are living through an exponential explosion in the amounts of data we are generating, capturing, analyzing, visualizing, and storing – including cultural content. On August 25, 2008, Google's software engineers announced on googleblog.blogspot.com that the index of web pages, which Google is computing several times daily, has reached 1 trillion unique URLs.[2] During the same month, YouTube.com reported that users were uploaded 13 hours of new video to the site every minute.[3] And in November 2008, the number of images housed on Flickr reached 3 billions.[4]

The "information bomb" already described by Paul Virilio in 1998 has not only exploded.[5] It also led to a chain of new explosions that together produced cumulative effects larger than anybody could have anticipated. In 2008 International Data Corporation (IDC) forecasted that by 2011, the digital universe would be 10 times the size it was in 2006. This corresponds to a compound annual growth rate of %60.[6] (Of course, it is possible that the global economic crisis which begun in 2008 may slow this growth – but probably not too much.)

User-generated content is one of the fastest growing parts of this expanding information universe. According to IDC 2008 study, "Approximately 70% of the digital universe is created by individuals."[7] In other words, the size of media created by users competes well with the amounts of data collected and created by computer systems (surveillance systems, sensor-based

---

[1] This article draws on white paper Cultural Analytics that I wrote in May 2007. I am periodically updating this paper. For the latest version, visit http://lab.softwarestudies.com/2008/09/cultural-analytics.html.
[2] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.
[3] http://en.wikipedia.org/wiki/YouTube.
[4] http://blog.flickr.net/en/2008/11/03/3-billion/
[5] Paul Virilio. *Information Bomb*. (Original French edition: 1988.) Verso, 2006.
[6] IDC (International Data Corporation). *The Diverse and Exploding Information Universe*. 2008. (2008 research data is available at http://www.emc.com/digital_universe.)
[7] Ibid.

applications, datacenters supporting “cloud computing,” etc.) So if Friedrich Kittler - writing well before the phenomena is “social media” – noted that in a computer universe “literature” (i.e. texts of any kind) consists mostly of computer-generated files, the humans are now catching up.

The exponential growth of a number of both non-professional media producers in 2000s has led to a fundamentally new cultural situation and a challenge to our normal ways of tracking and studying culture. Hundreds of millions of people are routinely creating and sharing cultural content - blogs, photos, videos, map layers, software code, etc. The same hundreds of millions of people engage in online discussions, leave comments and participate in other forms on online social communication. As the number of mobile phones with rich media capabilities is projected to keep growing, this number is only going to increase. In early 2008, there were 2.2 mobile phones in the world; it was projected that this number will become 4 billion by 2010, with main growth coming from China, India, and Africa.

Think about this: the number of images uploaded to Flickr every week today is probably larger than all objects contained in all art museums in the world.

The exponential increase in the numbers of non-professional producers of cultural content has been paralleled by another development that has not been widely discussed. And yet this development is equally important in understanding what culture is today. The rapid growth of professional educational and cultural institutions in many newly globalize countries since the end of the 1990s - along with the instant availability of cultural news over the web and ubiquity of media and design software - has also dramatically increased the number of culture professionals who participate in global cultural production and discussions. Hundreds of thousands of students, artists, designers, musicians have now access to the same ideas, information and tools. As a result, often it is no longer possible to talk about centers and provinces. (In fact, based on my own experiences, I believe the students, culture professionals, and governments in newly globalized countries are often more ready to embrace latest ideas than their equivalents in "old centers" of world culture.)

If you want to see the effects of these dimensions of cultural and digital globalization in action, visit the popular web sites where the professionals and the students working in different areas of media and design upload their portfolios and samples of their work – and note the range of countries from which the authors come from. Here are examples of these sites: xplsv.tv (motion graphics, animation), coroflot.com (design portfolios from around the world), archinect.com (architecture students projects), infosthetics.com (information visualization projects). For example, when I checked on December 24, 2008, the first three projects in the “artists” list on xplsv.tv came from Cuba, Hungary, and Norway.[8] Similarly, on the same day, the set of entries on the first page of coroflot.com (the site where designers from around the world upload their portfolios; it contained 120,000+ portfolios by the beginning of 2009) revealed a similar global cultural geography. Next to the predictable 20th century Western cultural capitals - New York and Milan – I

[8] http://xplsv.tv/artists/1/, accessed December 24, 2008.

also found portfolios from Shanghai, Waterloo (Belgium), Bratislava (Slovakia), and Seoul (South Korea).[9]

The companies which manage these sites for professional content usually do not publish detailed statistics about their visitors – but here is another example based on the quantitative data which I do have access to. In the spring of 2008 we have created a web site for our research lab at University of California, San Diego: softwarestudies.com. The web site content follows the genre of "research lab site" so we did not expect many visitors; we also have not done any mass email promotions or other marketing. However, when I examined Google Analytics stats for softwarestudies.com at the end of 2008, I discovered that we had visitors from 100 countries. Every month people from 1000+ cities worldwide check out site.[10] Even more interestingly are the statistics for these cities. During a typical month, no American cities made it into "top ten list" (I am not counting La Jolla which is the location of UCSD where our lab is located). For example, in November 2008, New York occupied 13th place, San Francisco was at 27th place, and Los Angeles was at 42nd place. The "top ten" cities were from Western Europe (Amsterdam, Berlin, Porto), Eastern Europe (Budapest), and South America (Sao Paulo). What is equally interesting is the list of visitors per city followed a classical "long tail" curve. There was no sharp break anymore between "old world" and "new world," or between "centers" and "provinces." (See softwarestudies.com/softbook for more complete statistics.)

All these explosions which took place since the late 1990s – non-professionals creating and sharing online cultural content, culture professionals in newly globalized countries, students in Eastern Europe, Asia and South America who can follow and participate in global cultural processes via the web and free communication tools (email, Skype, etc) – redefined what culture is.

Before, cultural theorists and historians could generate theories and histories based on small data sets (for instance, "classical Hollywood cinema," "Italian Renaissance," etc.) But how can we track "global digital cultures" with their billions of cultural objects, and hundreds of millions of contributors? Before you could write about culture by following what was going on in a small number of world capitals and schools. But how can we follow the developments in tens of thousands of cities and educational institutions?

Introducing Cultural Analytics

The ubiquity of computers, digital media software, consumer electronics, and computer networks led to the exponential rise in the numbers of cultural producers worldwide and the media they create – making it very difficult, if not impossible, to understand global cultural developments and dynamics in any substantial details using 20th century theoretical tools and methods. But what if

[9] coroflot.com, visited December 24, 2008. The number of design portfolios submitted by users to coroflot.com grew from 90, 657 on May 7, 2008 to 120,659 on December 24, 2008.

[10] See http://lab.softwarestudies.com/2008/11/softbook.html.

we can we use the same developments – computers, software, and availability of massive amounts of "born digital" cultural content – to track global cultural processes in ways impossible with traditional tools?

To investigate these questions – as well as to understand how the ubiquity of software tools for culture creation and sharing changes what "culture" is theoretically and practically – in 2007 we established Software Studies Initiative (softwarestudies.com). Our lab is located at the campus of University of California, San Diego (UCSD) and it housed inside one of the largest IT research centers in the U.S. - California Institute for Telecommunications and Information (www.calit2.net). Together with the researchers and students working in our lab, we have been developing a new paradigm for the study, teaching and public presentation of cultural artifacts, dynamics, and flows. We call this paradigm **Cultural Analytics**.

Today sciences, business, governments and other agencies rely on computer-based quantitative analysis and interactive visualization of large data sets and data flows. They employ statistical data analysis, data mining, information visualization, scientific visualization, visual analytics, simulation and other computer-based techniques. Our goal is start systematically applying these techniques to the analysis of contemporary cultural data. The large data sets are already here – the result of the digitization efforts by museums, libraries, and companies over the last ten years (think of book scanning by Google and Amazon) and the explosive growth of newly available cultural content on the web.

We believe that a systematic use of large-scale computational analysis and interactive visualization of cultural patterns will become a major trend in cultural criticism and culture industries in the coming decades. What will happen when humanists start using interactive visualizations as a standard tool in their work, the way many scientists do already? If slides made possible art history, and if a movie projector and video recorder enabled film studies, what new cultural disciplines may emerge out of the use of interactive visualization and data analysis of large cultural data sets?

From Culture (few) to Cultural Data (many)

In April 2008, exactly one year later we founded Software Studies Initiative, NEH (National Endowment for Humanities, the main federal agency in the U.S. which provides grants for humanities research) announced a new "Humanities High-Performance Computing" (HHPC) initiative that is based on the similar insight:

> Just as the sciences have, over time, begun to tap the enormous potential of High-Performance Computing, the humanities are beginning to as well. Humanities scholars often deal with large sets of unstructured data. This might take the form of historical newspapers, books, election data, archaeological fragments, audio or video contents, or a

> host of others. HHPC offers the humanist opportunities to sort through, mine, and better understand and visualize this data."[11]

In describing the rationale for Humanities High-Performance Computing program, the officers at NEH start with the **availability of high-performance computers** that are already common in the sciences and industry. In January 2009, NEH together with NSF (National Science Foundation) has annonced another program Digging Into Data which has articulated their vision in more detail. This time the program statement put more emphasis on the **wide availability of cultural content (**both contemporary and historical) **in digital form** as the reason for begin applying data analysis and visualization to "cultural data.":

> With books, newspapers, journals, films, artworks, and sound recordings being digitized on a massive scale, it is possible to apply data analysis techniques to large collections of diverse cultural heritage resources as well as scientific data. How might these techniques help scholars use these materials to ask new questions about and gain new insights into our world?

We fully share the vision put forward by NEH Digtal Humanities. Massive amounts of cultural content and high-speed computers go well together – without the latter, it would be very time consuming to analyze petabytes of data. However, as we discovered in our lab, even with small cultural data sets consisting from hundreds, dozens or even only a few objects it is already viable to do Cultural Analytics: that is, to quantitatively analyze the structure of these objects and visualize the results revealing the patterns which lie below the unaided capacities of human perception and cognition.

Since Cultural Analytics aims to take advantage of the exponential increase in the amounts of digital content since the middle of the 1990s, it will be useful to establish taxonomy for the different types of this content. Such taxonomy may guide design of research studies as well as be used to group these studies once they start multiply.

To begin with, we have vast amounts of **media content** in digital form – games, visual design, music, video, photos, visual art, blogs, web pages. This content can be further broken down into a few categories. Currently, the proportion of "**born digital**" media is increasing; however, people also continue to create analog media (for instance, when they shoot on film), which is later digitized.

We can further differentiate between different types of "born digital" media. Some of this media is explicitly made for the web: for example, blogs, web sites, layers created by users for Google Earth an Googe maps. But we also now find online massive amounts of "born digital" content

---

11 http://www.neh.gov/ODH/ResourceLibrary/HumanitiesHighPerformanceComputing/tabid/62/Default.aspx.

(photography, video, music) which until the advent of "social media" was not intended to be seen by people worldwide – but which now ends up online at social media sites (Flickr, YouTube, etc.) To differentiate between these two types, we may refer to the first category as "**web native**," or "web intended." The second category can be then called "digital media proper."

As I already noted, YouTube, Flickr, and other social media sites aimed at average people are paralled by more **specialized sites which serve professional and semi-professional users**: xplsv.tv, coroflot.com, archinect.com, modelmayhem.com, deviantart.com, etc.[12] Housing projects and portfolios by hundreds of thousands of artists, media designers, and other cultural professionals, these web sites provide a live shapshot of contemporary global cultural production and sensibility - thus offering a promise of being able to analyze the global cultural trends with the level of detail unthinkable previously. For instance, as of August 20008, deviantart.com has eight million members, 62+ million submissions, and was receiving 80,000 submissions per day.[13] Importantly, in addition to the standard "professional" and "pro-ams" categories, these sites also house the content of people who are just starting out and/or are currently "pro-ams" but who aspire to be full-time professionals. I think that the portfolios (or "ports" as they are sometimes called today) of these "**aspirational non-professionals**" are particularly significant if we want to study contemporary cultural stereotypes and conventions since, in aiming to create "professional" projects and portfolios, people often inadvertently expose the codes and the templates used in the industry in a very clear way.

Another important source of contemporary cultural content – and at the same time, a window into yet another cultural world different from non-professional users and aspiring professionals - are the **web sites and wikis created by faculty** teaching in creative disciplines to post and discuss their class assignments. (Although I don't have direct statistics on how many sites and wikis for classes are out there, here is one indication: a popular wiki creation software pbwiki.com has been used by 250,000 educators.[14]) These sites often contain **student projects** – which provides yet another interesting source of content.

Finally, beyond class web sites, the sites for professionals, aspiring professionals, and non-professionals, and other centralized content repositories, we have **millions of web sites and blogs by individual cultural creators and creative industry companies**. Regardless of the industry category and the type of content people and companies produce, it is now taken for granted that you need to have a web presence with your demo reel and/or portfolio, descriptions of particular projects, a CV, and so on. All this information can be potentially used to do something that previously was un-imaginable: to create dynamic (i.e. changing in time) maps of global

---

[12] The web sites aimed at non-professionals such as Flickr.com, YouTube.com and Vimeo.com also contain large amounts of media created media professionals and students: photography portfolio, independent films, illustrations and design, etc. Often the professionals create their own groups – which makes it easier for us to find their work on these general-purpose sites. However, the sites specifically aimed at the professionals also often feature CVs, descriptions of projects, and other information not available on general social media sites.

[13] http://en.wikipedia.org/wiki/DeviantArt.

[14] http://pbwiki.com/academic.wiki, accessed December 26, 2008.

cultural developments that reflect activities, aspirations, and cultural preferences of millions of creators.

A significant part of the available media content in digital form was originally created in electronic or physical media and has been digitized since the middle of the 1990s. We can call such content "**born analog**." But it is crucial to remember that what has been digitized in many cases are only the canonical works, i.e. a tiny part of culture deemed to be significant by our cultural institutions. What remains outside of the digital universe is the rest: provincial nineteen century newspapers sitting in some small library somewhere; millions of paintings in tens of thousands of small museums in small cities around the world; millions of thousands of specialized magazines in all kinds of fields and areas which no longer even exist; millions of home moves…

This creates a problem for Cultural Analytics, which has a potential to map everything that remains outside the canon – to begin generating "art history without great names." We want to understand not only the exceptional but also the typical; not only the few "cultural sentences spoken by a few "great man" but the patterns in all cultural sentences spoken by everybody else; in short, what is outside a few great museums rather than what is inside and what has been already extensively discussed too many times. To do this, we will need as much of previous culture in digital form as possible. However, what is digitally available is surprisingly little.

Here is an example from our research. We were interested in the following question: what did people actually painted around the world in 1930 – outside of a few "isms" and a few dozen artists who entered the Western art historical canon? We did a search on artstor.org which at the time of this writing contains close to one million images of art, architecture and design which come from many important US museum and collections, as well as 200,000+ slide library of University of California, San Diego where our lab is located. (This set which at present is the largest single collection in artstor is interesting in that it reflects the biases of art history as it was taught over a few decades when color slides were the main media for teaching and studying art.) To collect the images of artworks that are outside of the usual Western art historical canon, we excluded from the search Western Europe and North America. This left the rest of the world: Eastern Europe, South-East Asia, East Asia, West Asia, Oceania, Central America, South America, etc. When we searched for paintings done in these parts of the world in 1930, we only found a few dozen images. This highly uneven distribution of cultural samples is not due to Artstor since it does not digitize images itself – it only makes available images submitted to its by museums and other cultural institutions. So what the results of our search reflect is what museums collect and what they think should be digitized first. In other words, a number of major US collections and a slide library of a major research university (which now has a large proportion of Asian students) together contain only a few dozen paintings done outside the West in 1930 which got digitized. In contrast, searching for Picasso returned around 700 images. If this example if any indication, digital depositories may be amplifying the already existed biases and filters of modern cultural canons. Instead of transforming the "top forty" into "the long tail," digitization can be producing the opposite effect.

Media content in digital form is not the only type of data that we can analyze quantitatively to potentially reveal new cultural patterns. Computers also allow us to capture and subsequently analyze many dimensions of human cultural activities that could not be recorded before. Any cultural activity – surfing the web, playing a game, etc. - which passes through a computer or a computer-based media device leaves traces: keystroke presses, cursor movements and other screen activity, controller positions (think of We controller), and so on. Combined with camera, a microphone, and other capture technologies, computers can also capture other dimensions of human behavior such as body and eye movements and speech. And web servers log yet other types of information: which pages the users visited, how much time they spend on each page, which files they downloaded, and so on. (In this respect, Google Analytics that processes and organizes this information provided a direct inspiration for the idea of Cultural Analytics.

Of course, in addition to all this information which can be captured automatically, the rise of social media since 2005 created a new social environment where people voluntarily reveal their cultural choices and preferences: rating books, movies, blog posts, software, voting for their favorites, etc. Even importantly, people discuss and debate their cultural preferences, ideas and perceptions online. They comment on Flickr photographs, post their opinions about books on amazon.com, critique movies on rottentomatoes.com, review products on epinions.com, and enthusiastically debate, argue, agree and disagree with each other on numerous social media sites, fan sites, forums, groups, and mailing lists. All these conversations, discussions and reflections which before were either invisible or simply could not take place on the same scale are now taking place in public.

To summarize this discussion: because of digitization efforts since the middle of the 1990s, and because the significant (and constantly growing) percentage of all cultural and social activities passes through, or takes place on the web or networked media devices (mobile phones, game platforms, etc.), we now have access unprecedented amounts of both “cultural data” (cultural artifacts themselves), and “data about culture.” All this data can be grouped into three broad conceptual categories:

- Cultural artifacts (“born digital” or digitized).
- Data about people’ interactions with digital media (automatically captured by computers or computer-based media devices)
- Online discourse around (or accompanying) cultural activities, cultural objects, and creation process voluntarily created by people.

There are other ways to divide this recently emerged cultural data universe. For example, we can also make a distinction between “cultural data” and “cultural information”:

- **Cultural data**: photos, art, music, design, architecture, films, motion graphics, games, web sites - i.e., actual cultural artifacts which are either born digital, or are represented through digital media (for examples, photos of architecture).
- **Cultural information**: cultural news and reviews published on the web (web sites, blogs) – i.e., a kind of “extended metadata” about these artifacts.

Another important distinction, which is useful to establish, has to do with the relationships between the original cultural artifact/activity and its digital representation:

- "Born digital" artifacts: representation = original.
- Digitized artifacts that originated in other media - therefore, their representation in digital form may not contain all the original information. For example, digital images of paintings available in online repositories and museum databases normally do not fully show their 3D texture. (This information can be captured with 3D scanning technologies – but this is not commonly done at this moment.).
- Cultural experiences (experiencing theatre, dance, performance, architecture and space design; interacting with products; playing video games; interacting with locative media applications on a GPS enabled mobile device) where the properties of material/media objects that we can record and analyze is only one part of an experience. For example, in the case of spatial experiences, architectural plans will only tell us a part of a story; we may also want to use video and motion capture of people interacting with the spaces, and other information.

The rapid explosion of "born digital" data has not passed unnoticed. In fact, the web companies themselves have played an important role in making it happen so they can benefit from it economically. Not surprisingly, out of the different categories of cultural data, born digital data is already been exploited most aggressively (because it is the easiest to access and collect), followed by digitized content. Google and other search engines analyze billions of web pages and the links between them to make their search algorithms run. Nielsen Blogpulse mines 100+ million blogs to detect trends in what people are saying about particular brands, products and other topics its clients are interested in.[15] Amazon.com analyzes the contents of the books it sells to calculate "Statistically Improbable Phrases" used to identify unique parts of the books.[16]

In terms of media types, today text receives most attention - because language is discrete and because the theoretical paradigms to describe it (linguistics, computational linguistics, discourse analysis, etc.) have already been fully developed before the explosion of "web native" text universe. Another type of cultural media, which is also starting to be systematically subjected to computer analysis in large quantities, is music. (This is also made possible by the fact that Western music used formal notation systems for a very long time.) A number of online music search engines and Internet radio stations use computation analysis to find particular songs. (Examples: Musipedia, Shazam, and other applications which use acoustic fingerprinting.[17]) In comparison, other types of media and content receive much less attention.

If we are interested in analyzing cultural patterns in other media besides text and sound, and also in asking larger theoretical questions about cultures (as opposed to more narrow pragmatic

[15] "BlogPulse Reaches 100 Million Mark" < http://blog.blogpulse.com/archives/000796.html>.
[16] http://en.wikipedia.org/wiki/Statistically_Improbable_Phrases.
[17] http://en.wikipedia.org/wiki/Acoustic_fingerprint

questions asked in professional fields such as web mining or quantitative marketing research – for instance, identifying how consumers perceive different brands in a particular market segment[18]), we need to adopt a broader perspective. Firstly, we need to develop techniques to analyze and visualize the patterns in different forms of cultural media - movies, cartoons, motion graphics, photography, video games, web sites, product and graphic design, architecture, etc. Second, while we can certainly take advantage of the "web native" cultural content, we should also work with other categories that I listed above ("digitized artifacts which originated in other media"; "cultural experiences.") Thirdly, we should be self-reflective. We need to think about the consequences of thinking of culture as data and of computers as the analytical tools: what is left outside, what types of analysis and questions get privileged, and so on. This self-reflection should be part of any Cultural Analytics study. These three points guide our Cultural Analytics research.

Cultural Image Processing

Cultural Analytics is thinkable and possible because of three developments: digitization of cultural assets and the rise of web and social media; work in computer science; and the rise of a number of fields which use computers to create new ways of representing and interacting with data. The two related fields of computer science - image processing and computer vision - provide us with the variety of techniques to automatically analyze visual media. The fields of science visualization, information visualization, media design, and digital art provide us with the techniques to visually represent patterns in data and interactively explore this data.

While people in digital humanities have been using statistical techniques to explore patterns in literary text for a long time, I believe that we are the first lab to start systematically using image processing and computer vision for automatic analysis of visual media in the humanities contest. This is what separates us from 20th century humanities disciplines that focus on visual media (art history, film studies, cultural studies) and also 20th century paradigms for quantitative media research developed within social sciences such as quantitative communication studies and certain works in sociology of culture. Similarly, while artists, designers and computer scientists have already created a number of projects to visualize cultural media, the existing projects that I am aware of rely on existing metadata such as Flickr community-contributed tags[19]. In other words, they use information about visual media – creation date, author name, tags, favorites, etc. – and do not analyze the media itself.

In contrast, Cultural Analytics uses image processing and computer vision techniques to automatically analyze large sets of visual cultural objects to generate numerical descriptions of their structure and content. These numerical descriptions can be then graphed and also analyzed statistically.

While digital media authoring programs such as Photoshop and After Effects incorporate certain image processing techniques such as blur, sharpen, and edge detecting filters, motion tracking, and so on, there are hundreds of other features that can be automatically extracted from still and

[18] http://en.wikipedia.org/wiki/Perceptual_mapping.

[19] These projects can be found at visualcomplexity.org and infosthetics.com.

moving images. Most importantly, while Photoshop and other media applications internally measure properties of images and video in order to change them - blurring, sharpening, changing contrast and colors, etc. – at this time they do not make available to users the results of these measurements. So while we can use Photoshop to highlight some dimensions of image structure (for instance, reducing an image to its edge), we can't perform more systematic analysis.

To do this, we need to turn to more specialized image processing software such as open source imageJ which has been developed for live sciences applications and which we have been using and extending in our lab. MATLAB, popular software for numerical analysis, provides many image processing applications. There are also specialized software libraries of image processing functions such as openCV. A number of high-language programming languages created by artists and designers in 2000s such as Processing and openFrameworks also provide some image processing functions. Finally, many more techniques are described in computer science publications.

While certain common techniques can be used without the knowledge of computer programming and statistics, many others require knowledge of C or Java programming. Which of the algorithms can be particularly useful for cultural analysis and visualization? Can we create (relatively) easy-to-use tools which will allow non-technical users to perform automatic analysis of visual media?
These are the questions we are currently investigating. As we are gradually discover, in spite of the fact that the fields of image processing and computer vision have existed now for approximately five decades, the analysis of cultural media often requires development of new techniques that do not yet exist.

To summarize: the key idea of Cultural Analytics is the use of computers to **automatically analyze cultural artifacts in visual media extracting large numbers of features which characterize their structure and content**. For example, in the case of a visual image, we can analyze its grayscale and color characteristics, orientations of lines, texture, composition, and so on. Therefore, we can also use another term to refer to our research method – **Quantitative Cultural Analysis** (QCA).

While we are interested in both content and structure of cultural artifacts, at present automatic analysis of structure is much further developed than the analysis of content. For example, we can ask computers to automatically measure gray tone values of each frame in a feature film, to detect shot boundaries, to analyze motion in every shot, to calculate how color palette changes throughout the film, and so on. However, if we want to annotate film's content – writing down what kind of space we see in each shot, what kinds of interactions between characters are taking place, the topics of their conversations, etc., the automatic techniques to do this are more complex (i.e., they are not available in software such as MAT LAB and imageJ) and less reliable. For many types of content analysis, at present the best way to is annotate media manually – which is obviously quite time consuming for large data sets. In the time it will take one person to produce such annotations for the content of one movie, we can use computers to automatically analyze the structure of many thousands of movies. Therefore, we started developing Cultural Analytics by developing techniques for the analysis and visualization of structures of individual cultural artifacts

and large sets of such artifacts - with the idea that once we develop these techniques we will gradually move into automatic analysis of content.

Deep Search

In November 2008 we received a grant that gives us 300,000 hr of computing time on US Department of Energy supercomputers. This is enough to analyze millions of still images and video – art, design, street fashion, feature films, anime series, etc. This scale of data is matched by the size of visual displays that we are using in our work. As I already mentioned, we are located inside one of the leading IT research centers in the U.S. - California Institute for Telecommunication and Information Technology (Calit2). This allows us to take advantage of the next-generation visual technologies - such as HIperSpace, currently one of the highest resolution displays for scientific visualization and visual analytics applications in the world. (Resolution: 35,640 by 8,000 pixels. Size: 9.7m x 2.3m.)

One of the directions we are planning to pursue in the future is the development of visual systems that would allow us to follow global cultural dynamics in real-time. Imagine a real-time traffic display (a la car navigation systems) – except that the display is wall-size, the resolution is thousands of times greater, and the traffic shown is not cars on highways, but **real-time cultural flows** around the world. Imagine the same wall-sized display divided into multiple windows, each showing different real-time and historical data about cultural, social, and economic news and trends – thus providing **a situational awareness for cultural analysts**. Imagine the same wall-sized display playing an animation of what looks like an earthquake **simulation** produced on a super-computer – except in this case the "earthquake" is the release of a new version of popular software, the announcement of an important architectural project, or any other important cultural event. What we are seeing are the effects of such "cultural earthquake" over time and space. Imagine a wall-sized computer graphic showing **the long tail** of cultural production that allows you to zoom to see each individual product together with rich data about it (à la real estate map on zillow.com) – while the graph is constantly updated in real-time by pulling data from the web. Imagine a visualization that shows how other people around the word remix new videos created in a fan community, or how a new design software gradually affects the kinds of forms being imagined today (the way Alias and Maya led to a new language in architecture). These are the kinds of tools we want to create to enable new type of cultural criticism and analysis appropriate for the era of cultural globalization and user-generated media: three hundred digital art departments in China alone; approximately 10,000 new users uploading their professional design portfolios on coroflort.com every month; billions of blogs, user-generated photographs and videos; and other cultural expressions which are similarly now created at a scale unthinkable only ten years ago.

To conclude, I would like to come back to my opening point – the rise of search as a new dominant mode for interacting with information. As I mentioned, this development is just one of many consequence of the dramatic and rapid in the scale of information and content being produced which we experienced since the middle of the 1990s. To serve the users search results, Google, Yahoo, and other search engine analyze many different types of data – including both

metadata of particular web pages (so-called “meta elements”) and their content. (According to Google, its search engine algorithm uses more than 200 input types.[20]) However, just as Photoshop and other commercial content creating software do not expose to users the features of images or videos they are internally measuring, Google and Yahoo do not reveal the measurements of web pages they analyze – they only serve their conclusions (which sites best fit the search string) which their propriety algorithms generate by combining these measures. In contrast, the goal of cultural Analytics is to enable what we may call “deep cultural search” – give users the open-source tools so they themselves can analyze any type of cultural content in detail and use the results of this analysis in new ways.

[March 2009]

[20] http://www.google.com/corporate/tech.html.

2. Internet censorship research. History and analysis

N. Villeneuve (2007). "Evasion tactics: Global online censorship is growing, but so are the means to challenge it and protect privacy." *Index on Censorship*. 36(4): 71-85.

Index on Censorship
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t714592856

Evasion tactics

Nart Villeneuve

Online Publication Date: 01 November 2007
To cite this Article: Villeneuve, Nart (2007) 'Evasion tactics', Index on Censorship, 36:4, 71 - 85
To link to this article: DOI: 10.1080/03064220701738651
URL: http://dx.doi.org/10.1080/03064220701738651

EVASION TACTICS

NART VILLENEUVE

GLOBAL ONLINE CENSORSHIP IS GROWING, BUT SO ARE THE MEANS TO CHALLENGE IT AND PROTECT PRIVACY

The number of countries that censor and monitor their citizens' use of the Internet is increasing. While it is no secret that China and Iran censor the Internet, at least 25 countries, including Pakistan, Ethiopia, Thailand and Uzbekistan, also have technical filtering regimes in place. Some of the technology is even exported by western companies: search engines, blog hosting providers and email providers have extended their existing filtering mechanisms – which usually target pornography and copyright infringement – to censor political content and gain access to lucrative markets in repressive countries.

Censorship and surveillance is not restricted to authoritarian regimes. The technology used to censor the Internet in entire countries in the Middle East and North Africa also filters access in schools and libraries in North America. An Internet service provider (ISP) in Canada blocked access to a website set up by members of its workers' union during a labour dispute. ISPs in the United States have implemented a sophisticated, and illegal, monitoring and data-mining programme, covering both Internet and telephone communications, at the behest of the National Security Agency. The problem is magnified when the concept of censorship is extended beyond just the technical aspects of filtering web content and Internet services.

There is, however, a growing resistance to Internet censorship and surveillance, although it is often characterised as a struggle confined to dissidents in a few select authoritarian regimes. There are a wide variety of awareness raising campaigns as well as academic research projects aimed at exposing and confronting censorship. Legal battles are being fought all over the globe, while the development and use of technologies that protect privacy and make it possible to circumvent censorship are rapidly increasing. The same tools helping dissidents to evade censorship in repressive countries are also being used by citizens in democratic countries – to protect themselves from unwarranted Internet surveillance.

There are three key factors to Internet censorship. First, there are formal and informal mechanisms, including laws, licensing and self-regulation, that act to create the legal, and often extra-legal, framework within which Internet censorship takes place. Second, there are a variety of technical methods

through which Internet filtering and blocking can be implemented to restrict access to content and services online. Third, Internet surveillance technologies are routinely deployed in order to monitor and track online communications. All countries use varying degrees of these to implement control, generating fear among Internet users and contributing to a climate of self-censorship that is creating alarming challenges to freedom of expression online.

The legal basis for technical filtering is murky and rarely explicit, and can vary significantly from country to country. It is often a combination of press law, telecommunications regulations and laws protecting state security.

Uzbekistan online
Credit: Sean Sprague/Panos, 2005

Regulation and oversight is most often conducted by the Telecommunication Ministry or by the often state-controlled telecommunications companies.

In South Korea, the Ministry of Information and Communication instructed Internet service providers to block access to content deemed to be 'North Korean propaganda' and thus illegal under the vague, and often abused, national security law. The Korean Internet Safety Commission (KISCOM) has also been set up to advise the government's Internet censorship policies and its logo is prominently featured, along with the National Police Agency's logo, on the 'block page' users see when they try to access censored websites. South Korea received a 'high' transparency rating from the OpenNet Initiative – a research project documenting

Internet censorship. This was based on the country's open acknowledgment of filtering, along with the presence of a 'block page' that informs users when attempts are made to access censored content.

In contrast, Uzbekistan received a 'low' transparency rating because the country's filtering regime is based on a combination of self-censorship by ISPs and pressure from the country's intelligence service – the National Security Service (SNB). In addition to occasionally ordering ISPs to block specific sites, the SNB monitoring also encourages them to self-censor or risk having their licences revoked. In a way, the practice is symbolic of the censorship regime as a whole. The ISPs attempt to conceal their filtering by redirecting users to innocuous sites when they try to access blocked content.

In some countries, there is no technical filtering in place; it is the legal system itself which acts as the primary mechanism of Internet censorship. Threatening ISPs, or content providers such as search engines, with 'takedown' requests is one of the most undocumented methods of censoring Internet content. In some cases these can be formal legal requests for removal due to copyright violation or claims of libel/defamation or informal requests due to allegations of supporting terrorism. ISPs are not required to report such 'takedowns' and most happen in complete silence. In these cases, ISPs act as judge, jury and enforcer at the same time and will act to remove content rather than fully investigate the claim, in order to avoid liability.

The questions surrounding the lack of transparency and accountability led Christian Ahlert, Chris Marsden and Chester Yung, from the Oxford Centre for Socio-Legal Studies, to investigate what they termed the 'privatisation of censorship'. In 2003, they conducted an experiment, known as 'Liberty', to test notice and takedown procedures in the US and Europe. They created a web page containing text that was clearly in the public domain and uploaded it to ISPs in the US and the UK. The uploaded text was an excerpt from Chapter 2 of J S Mill's *On Liberty*, which discusses freedom of the press and censorship. They then created an email account with a free service for a mythical organisation called the 'John Stuart Mill Heritage Foundation' and sent takedown notices to the ISPs claiming copyright infringement. In the UK, ISPs took the information down, but in the US, they asked for more details, including a declaration 'under penalty of perjury' that the claim was valid. At this point, the researchers terminated the experiment. However, they noted that if they had supplied the language required by the ISPs, the takedown process could have continued.

In 2004, the group 'Bits of Freedom' conducted a similar experiment using Dutch ISPs. They uploaded text that was clearly in the public domain – the text even stated that it was in the public domain – and then sent takedown notices

from free email accounts. Of the ten ISPs tested, only three did not remove the content. One provider even forwarded the account details of the customer to the complainant. 'Bits of Freedom' went further than the 'Liberty' experiment by filling out a form sent by the ISPs that asked for additional details including name and address and to 'indemnify the provider from any liability for acting upon the request to take down'. This led 'Bits of Freedom' to conclude that the 'penalty of perjury' test which worked in the 'Liberty' experiment was clearly not enough of a check against abuse.

These studies exposed the flawed process through which takedown and notice are being implemented. It is clearly being exploited to silence online critics. The Church of Scientology has used takedown notices alleging copyright violations with great success, even forcing Google to remove links from its search engine to particular sites. In addition to copyright, threats of law suits for defamation and libel are increasingly being used to stifle criticism. Singapore and Malaysia have often been accused of using such tactics. The new targets for libel and defamation cases are bloggers. While many blogs are about personal interests and read more like a diary, the blogging platform is also being used by citizen journalists, who publish without the filters of the traditional media.

While there have been documented cases where bloggers have been prosecuted for libel or defamation, many never make it to court. In August 2007, the website of the Iranian blogger Hossein Derakhshan was shut down. Derakhshan's blog has long been censored in Iran. Despite being filtered, it remained popular and Iranians used technology to bypass the filters and access the site. However, after criticising an Iranian intellectual, Mehdi Khalaji, for working for a conservative think-tank in Washington DC, Derakhshan, his web hosting company, Hosting Matters, and domain registrar, GoDaddy, were served with a takedown notice. The notice, alleging libel and defamation, led to the deletion of some of Derakhshan's blog posts by his hosting company and ultimately to the termination of his blog's hosting service. Exemplifying just how flawed the notice and takedown process is, the notice claimed that in addition to Derakhshan, both the domain registrar and the web hosting company were implicated in and/or liable for activities conducted on Derakhshan's blog. The notice implied that each of the three named in the notice (the registrar, the hosting company and Derakhshan) 'published' defamatory information and were therefore liable for damages.

The chilling effect of notice and takedown is well illustrated in this case. Faced with legal threats, Derakshan's web-hosting company ordered him to remove 'all' references to Mr Khalaji or they would remove his entire website, even though the company recognised that the claims fell into a 'grey area'. After taking down the offending posts, but refusing to remove all references to Mr Khalaji, Hosting Matters asked Mr Derakhshan to remove additional posts about Mr Khalaji.

> Please remove the latest post you have made referencing Mehdi Khalaji. This person continues to insist that everything and anything you post about him is defamatory. While we do not agree with the assessment as it relates to the latest post you have made, we do not have the time, interest, or resources to invest in continually dealing with his complaints and to review your site.
>
> (Source: http://hodertemp.blogspot.com/2007/08/accounts-and-billing-hosting-matters.html)

This exchange clearly shows why ISPs are not equipped or qualified to make judgments on content and will always default to the lowest common denominator, with serious repercussions for freedom of speech and expression.

Content removed for allegedly supporting terrorism is one of the least documented forms of takedown. With copyright and defamation there is at least some element of a legal procedure, however flawed, but when it comes to terrorism, individuals and groups simply contact ISPs and have content removed. The Internet Haganah, which calls for the removal of sites which allegedly support terrorism, had counted 600 successful takedowns by 2005. These include websites, groups hosted by Yahoo! and storefronts at Cafe Press. In 2005, the Toronto-based Friends of Simon Wiesenthal Center had several sites removed by their ISPs, one of which only contained a flag that carried the inscription, 'There is no other God but Allah'. There was no hateful text or material advocating suicide bombing. The issue, as noted in the press release, was that the flag appeared to be the same one used by Hizb-ut-Tahrir, a group that, at the time, was not on the US State Department's or Canada's list of terrorist organisations.

While content removal remains largely undocumented, it is possible to interrogate the technical infrastructure through which countries block access. There is a variety of methods through which content on the Internet can be blocked that falls into three general categories: domain name server (DNS) tampering, Internet protocol (IP) address blocking, uniform resource locator (URL) filtering and keyword filtering.

DNS is the system that translates a domain name into a numerical IP address. By tampering with their DNS server, ISPs can force domain names to resolve to invalid or 'spoofed' IP addresses. The South Korean ISP, Kornet, resolves censored domains to an IP address which displays a police block page, indicating to the user that illegal content is being accessed. One of India's leading ISPs, Videsh Sanchar Nigam Ltd, uses DNS tampering to block websites, forcing domains to resolve to the invalid address 1.2.3.4 India focuses its filtering on Hindu extremists and some American right-wing sites, as well as sites advocating

a Dalit homeland. DNS tampering is easy to circumvent, as a user can simply configure their computer to use an alternate DNS server, but it is often used by ISPs to avoid problems with over-blocking.

Countries new to filtering will generally start with blocking by IP address, before moving on to more expensive URL filtering solutions. Most ISPs do not have the capacity to filter by URL and the ones that do would need to purchase a significant amount of equipment to implement URL filtering without a significant drop in performance. ISPs must often respond quickly and effectively to blocking orders from the government or national security and intelligence services. So they block material in the cheapest way, using technology already integrated into their normal network environment. Blocking by IP is effective (the target site is blocked) and no new equipment needs to be purchased. It can be implemented in an instant, as all the required technology and expertise is readily available. Many ISPs already block IP addresses to combat spam and viruses.

But blocking by IP address comes with a significant cost: over-blocking. Many unrelated websites may be hosted on a single IP address, so, when blocked, all other content hosted on the server will also be inaccessible. Pakistan is an interesting case, because it is one of the few countries in which the blocking lists have become public. Internet traffic routes through a gateway operated by the Pakistan Telecommunications Company Limited. Officially, Pakistan only blocks 17 sites, although the list contains dead sites and typographical errors. The OpenNet Initiative tested 11 of these designated sites. It found that, in total, nearly 3.5 million are actually blocked. This total does not, however, include the hundreds of thousands of individual blogs hosted on Google's blogspot service. Pakistan has blocked access to the IP addresses of key hosting providers including GoDaddy and Yahoo! In the past, Pakistan has also blocked IP addresses associated with the mirroring company Akamai, causing hundreds of thousands of sites to become inaccessible.

This is the same technique that the Canadian ISP Telus used to block access to a union-affiliated site during a labour dispute. In the process, it blocked access to over 700 unrelated sites. This generated a considerable amount of criticism and clearly demonstrated the unintended consequences of filtering technologies.

Over-blocking tends to create a significant backlash, especially from non-activist Internet users. While people will often tolerate the blocking of extremist or offensive sites, when their own regular browsing and blogging is interrupted they quickly become aware of censorship's impact and campaign against it. An excellent example has been the 'Don't Block the Blog' campaign which was started after Pakistan blocked access to Blogspot; pkblogs.com now offers an alternate means of accessing Blogspot, bypassing Pakistan's filtering.

However, in response, the authorities will often seek to implement filtering techniques that better target the specific sites they want to block.

As the complexities of implementing an effective filtering system are recognised, countries are beginning to move towards the use of commercial filtering technology. In addition to the issue of over-blocking, filtering systems suffer from another inherent problem: under-blocking. Alongside the maintenance of blocking lists – which can be considerable for categories such as pornography – other forms of content need to be blocked in order to have a reasonably effective filtering system. This primarily involves finding and blocking sites that enable users to get around the filtering. Commercial technologies have enabled the expansion of Internet censorship, providing a fine-grain control over the filtering and monitoring process. They are equipped with easy-to-use graphical interfaces for management of the filtering system, as well as pre-configured blocking categories which include 'anonymisers' – sites that allow one to bypass censorship.

There are a growing number of countries that use commercial filtering technology. However it is often difficult to determine the exact technology being used. To date, the OpenNet Initiative has identified the use of SmartFilter, produced by the US company Secure Computing, in Saudi Arabia, Tunisia, Oman, Sudan, United Arab Emirates, and possibly in Iran, while Websense and Fortinet are being used in Yemen and Burma respectively.

Commercial filtering technologies can be configured to block very specific content as well. In Saudi Arabia, for example, the websites of the Arab Human Rights Information Network and Humum are mostly accessible. Only specific pages about Saudi Arabia are blocked. They can also be used to avoid network degradation associated with other methods of filtering. Saudi Arabia claims that its system actually improves performance.

But commercial filtering technologies introduce additional concerns. The way in which these companies categorise websites affects access to the Internet more widely. SmartFilter, for example, is configured to block predefined categories of content: anonymisers, nudity, pornography, and sexual materials. Recently, the video-sharing website dailymotion.com was blocked in Tunisia. SmartFilter had temporarily categorised the site as pornography, and, since Tunisia blocks the pornography category, the website was blocked. Several days later, SmartFilter removed dailymotion.com from the pornography category and it became accessible.

In effect, governments are ceding the decision on what precisely to filter to unaccountable commercial entities. Due to the categorisation choices made by these companies, content may become inaccessible to entire populations, even if the government never intended to block the content. This situation is exacerbated by the intellectual property protections afforded to the companies. The block lists

used by commercial filtering software are secret; decrypting and analysing them is considered to be illegal.

The chilling effect of legislation, such as the United States' Digital Millennium Copyright Act (DMCA), has resulted in researchers stopping work on the impact of commercial filtering software. This is especially relevant because the software is increasingly turning up in undemocratic countries and is being used to filter all sorts of content – including political speech.

The work of two high-profile researchers was cut short in this field due to mounting legal risks. Ben Edelman sought to obtain a court judgment in order to protect himself from liability for decrypting the blocking lists of commercial filtering technologies, but his case was dismissed. Seth Finkelstein was forced to abandon work decrypting the blocking lists of filtering software products because of the associated legal risks.

Despite the obstacles, there are growing efforts to resist and challenge the spread of Internet censorship. These range from research projects designed to document and expose current censorship practices, to legal challenges to the development and use of technologies. Combined, these efforts seek to challenge the norms surrounding the practice of filtering, change the policies of governments and ISPs and empower users to protect their privacy and exercise the right of free expression online.

There are numerous human rights organisations investigating and highlighting egregious cases of Internet censorship, including Amnesty International, Reporters Without Borders and Human Rights Watch. These groups collect and analyse reports of blocked content, as well as create campaigns to highlight egregious cases of censorship and make that information available to a wide audience. They also seek to influence public policy and engage in lobbying and advocacy, targeting governments and corporations. Amnesty International started the irrepressible.info campaign that seeks to highlight Internet censorship by allowing website owners to display fragments of text taken from censored sites around the world. More than 70,000 people have signed the pledge calling for an end to 'unwarranted restriction of freedom of expression on the Internet'. The signatures from this pledge were delivered at the 2006 Internet Governance Forum before an audience of governments and companies involved in censoring the Internet.

Reporters Without Borders maintains a list of imprisoned cyberdissidents and has also created the *Handbook for Bloggers and Cyber-dissidents* which provides information on how to secure one's communications and bypass Internet censorship. Human Rights Watch has released detailed reports that not only document the technical aspects of filtering, but also the cases of individuals who have been directly affected by state censorship. The reports contain detailed

recommendations for governments, corporations and activists to promote policies that enhance freedom of expression online.

In addition to major international organisations, there are coalitions such as the Global Voices Advocacy project and the Society Against Internet Censorship in Pakistan that seek to build alliances among bloggers and free expression advocates worldwide. There are also numerous grass-roots campaigns to free imprisoned bloggers around the world. The groups not only raise awareness about violations of freedom of expression, but also provide information on how to bypass Internet censorship and on strategies to maintain anonymity online.

While advocacy is an extremely important component in challenging censorship, there also exists the need to technically uncover exactly the methods and targets of state censorship. Research projects have been pivotal in establishing a body of credible evidence, exposing practices that are most often secretive and forcing governments and corporations to account for their censorship practices. Faced with accurate, empirical evidence, it becomes increasingly difficult for states to continue denying the fact that they are censoring the Internet.

The chillingeffects.org project, a collaboration between leading law schools and universities across the US, tracks notice and takedown requests. The majority of complaints relate to copyright and trademark infringement, but increasingly also cover libel and defamation. The project has tracked over 2,000 such notices. It also provides 'weather reports', which are a great resource for investigating the use of the law to remove content.

The OpenNet Initiative (ONI) has developed a set of tests that interrogate the Internet to identify filtered content. To date, ONI has tested in over 40 countries worldwide and has uncovered the techniques employed by states, usually at the ISP level, to filter the Internet. Moreover, ONI has begun to develop methods to monitor Internet access during key time periods, such as elections, in order to collect evidence of the temporary tampering with Internet access and in some cases denial of service to opposition websites. ONI has also identified technologies created by American companies, which are used to censor political speech in repressive countries. This work has informed a US Congressional committee that brought representatives from leading companies to explain their actions. ONI work has also been widely cited and used by human rights and press freedom groups around the world.

But while ONI has done excellent work in interrogating systems of Internet filtering, surveillance has proven to be much more elusive: it can be conducted in a passive manner and is thus extremely difficult, if not impossible, to document technically. Therefore, the majority of the work done in uncovering systems

of surveillance has been through leaks, freedom of information requests and legal process.

The United States maintains the most sophisticated surveillance programme in the world. The American Civil Liberties Union (ACLU) created the 'Surveillance Society Clock', modelled after the doomsday clock, to symbolise just how much of a threat the current levels of surveillance in the US are to a free society. The clock is currently at six minutes to midnight.

Surveillance practices in the US are being challenged in the courts. The Electronic Frontier Foundation (EFF) and the Electronic Privacy Information Center (EPIC) have been extremely active in bringing legal challenges to uncover the vast surveillance programme. The EFF filed a lawsuit on behalf of AT&T customers to challenge the company's participation in the National Security Agency's (NSA) illegal domestic surveillance. The challenge was made after it was revealed that the NSA had been data-mining Internet and telephone logs from various telecommunications companies in the US without the proper legal authority. In response, the Bush administration is seeking to shield participating companies behind vaguely worded 'state secrets' protection. The Department of Homeland Security (DHS) and the Pentagon also maintain surveillance programmes. As a result of investigative reporting and the threat of legal challenges, two of these programmes have been suspended. The DHS suspended ADVISE (Analysis, Dissemination, Visualisation, Insight and Semantic Enhancement) after it was found to violate privacy laws. The Pentagon suspended its TALON database – which monitored peace activists amongst others – and the infamous Total Information Awareness project after similar concerns were mounted.

Legal challenges against Internet censorship are also being mounted worldwide. In Iran, the conservative website Baztab was filtered after several articles critical of President Ahmadinejad were published, but access to the site was restored following a successful legal challenge. The unblocking of one website – run by well-connected people – is a small victory, but it could be very significant. If the procedures for blocking content become transparent, if there is an appeals process and some level of accountability, it then becomes increasingly difficult for governments to justify censorship. Human rights groups have long called for a legally transparent process through which censorship can be challenged.

China has also been the site of a legal challenge – once largely thought to be impossible. A Chinese blogger known as Yetaai [see pp161–164] brought a case against China Telecom for blocking his website. It is seen as a landmark case because it may force the company or the government to admit that Internet censorship actually takes place. Although many believe that Yetaai will not be successful, his case has inspired others to use the legal system to

Internet cafe, China
Credit: Gemma Kate Thorpe

challenge Internet censorship in China. Another blogger, Liu Xiaoyuan, has attempted to sue the Chinese company Sohu for censoring several posts on his blog, while a website, www.bullog.cn, is calling for public hearings to protect it from being shut down.

In another case that is emblematic of the global resistance to censorship, the family of Wang Xiaoning, an activist who was arrested and tortured in China, is suing Yahoo! in an American court because Yahoo! provided information to the Chinese government that was used in the prosecution. Yahoo! has filed a motion to dismiss the case.

This is not the first case in which Yahoo! has provided evidence to the Chinese government resulting in the conviction of dissidents. Chinese journalist Shi Tao was sentenced to ten years in prison in China, after distributing the Chinese government's instructions to domestic journalists on how to cover the anniversary of the Tiananmen Square massacre. Shi Tao sent the information to a foreign-hosted dissident website from his Yahoo! email account. The Chinese government asked Yahoo! to provide information on the account details and this information was used in the case against Shi Tao.

The case illustrates that while many people assume that there is anonymity online, users have to protect themselves to keep their identity hidden. Technologies that make it possible to circumvent censorship and enhance the individual's right to communicate and access information are also an important means for challenging censorship and surveillance. Filtering and monitoring communications online make it possible for hostile actors to find identifying information that may be used to arrest and imprison political dissidents.

In order to combat these growing threats, technologies are being developed to evade censorship and protect privacy. These same technologies are used by dissidents in politically repressive countries as well as activists in democratic countries. Peacefire, for example, is an organisation that develops and provides technology to evade censorship. It was formed to advocate on behalf of children who were being subjected to filtering in schools and libraries throughout the US. Peacefire now also focuses on providing these same censorship circumvention methods to users in China and Iran.

The technology allows a user in a censored location to connect to an unblocked, intermediary computer, in an uncensored location, to access content through the computer's Internet connection. The user in the censored country does not directly access a blocked website, but asks the intermediary computer to do so. The intermediary computer retrieves the requested website and displays it back to the user.

While there are a variety of technologies available that can be used to circumvent censorship, there is a fundamental challenge: how to disclose the location of the uncensored intermediary to users who want to bypass censorship, while keeping it secret from agents who seek to find and censor these intermediaries. There are two main approaches to this problem: public and private. The public approach is to create numerous intermediary locations, through which users can

bypass censorship and simply reveal more, through email lists, instant messaging and so on, as each becomes blocked. Censors who are slow to act will find more and more people using these circumvention systems. However, since many countries now use commercial filtering applications, the list of 'proxy and anonymiser' sites that these companies maintain are updated frequently, resulting in a situation where the lifetime of a new circumvention intermediary can last between one day and one week before being blocked.

Private circumvention solutions focus on distributing the location of the intermediary computer to people who know and trust one another. By leveraging these relationships of trust, a circumvention provider can slowly develop a network and provide stable circumvention services to a few – with a greatly reduced risk of being blocked by censors. Psiphon is a personal circumvention system that was designed and developed by the Citizen Lab at the University of Toronto. It allows users in uncensored locations to turn their own home computer into a circumvention server and allow their friends and family members in censored locations to surf freely. One of the goals of the project was to make the software extremely simple, so that those with limited technical abilities could make use of the technology.

There is an important distinction to be made between circumvention and anonymity technologies. Circumvention technologies focus, with varying degrees of security, on allowing users to bypass censorship, while anonymity technologies focus on protecting the users' identity from outside observers, such as government surveillance, as well as from the anonymity system itself. Circumvention systems that use encryption can protect users in some surveillance scenarios, but are not anonymous because owners of the circumvention system can see everything that the user does. They also cannot protect users from traffic analysis attacks in the same way that anonymity systems can. Anonymity systems protect privacy by shielding the identity of the requesting user from the content provider. In addition, they employ routing techniques to ensure that the user's identity is shielded from the anonymous communications system itself. In addition to providing anonymity, these technologies are also used in many countries to bypass Internet censorship. Anonymity systems are increasingly being recommended by privacy advocates. The Privacy Commissioner of Canada, for example, recommends that Internet users protect themselves online by using anonymity technologies, as well as anonymous remailers.

The most widely known anonymity system is Tor (see p143). It is promoted by the Electronic Frontier Foundation as software to protect privacy and civil liberties online and is used by bloggers who want anonymity, as well as by government embassies around the world. Tor works by routing a user's request through at least three Tor servers. As the request hops from one Tor

server to another, a layer of encryption is removed, so no individual server knows both the original source and destination of the request. The last server in the chain of hops, known as a circuit, actually connects to the requested content and then sends that information back through the circuit to the user. However, anonymity technologies are currently not difficult to block. Tor's developers are working on building in blocking resistance to the anonymity system.

The Internet is a tool, like any other, that can be both used and abused. We know that governments around the world, much like companies, schools, libraries, and parents, restrict access to Internet content they do not want their citizens, employees, students, patrons and children to see. However, there is a failure to recognise Internet censorship and surveillance as a growing global concern. There is a tendency instead to criticise the most infamous offenders – notably China and Iran – and to overlook repressive practices elsewhere. Focusing on the global character of both the practice of Internet censorship and surveillance, as well as the resistance to it, provides for both a better understanding of this important trend as well as for the possibility of creating global alliances to combat its spread. ❑

Nart Villeneuve *is a PhD student in Political Science at the University of Toronto. As Director of Technical Research for the Citizen Lab he has developed and conducted censorship testing in over 40 countries worldwide as part of the OpenNet Initiative and participated in the Psiphon circumvention project*

J. Zittrain, and B. Edelman (2002). "Documentation of Internet filtering in Saudi Arabia." Working Paper, Berkman Center for Internet & Society, Harvard Law School.

Documentation of Internet Filtering in Saudi Arabia

Jonathan Zittrain* and Benjamin Edelman**
Berkman Center for Internet & Society
Harvard Law School

[Overview - Specific Blocked Pages - Analysis & Summary Statistics - Conclusions]

Abstract: The authors connected to the Internet through proxy servers in Saudi Arabia and attempted to access approximately 60,000 Web pages as a means of empirically determining the scope and pervasiveness of Internet filtering there. Saudi-installed filtering systems prevented access to certain requested Web pages; the authors tracked 2,038 blocked pages. Such pages contained information about religion, health, education, reference, humor, and entertainment. See highlights of blocked pages. The authors conclude (1) that the Saudi government maintains an active interest in filtering non-sexually explicit Web content for users within the Kingdom; (2) that substantial amounts of non-sexually explicit Web content is in fact effectively inaccessible to most Saudi Arabians; and (3) that much of this content consists of sites that are popular elsewhere in the world.

Overview

A 2001 Council of Ministers Resolution prohibits users within the Kingdom of Saudi Arabia from publishing or accessing certain content on the Internet. The government's Internet Services Unit (ISU) operates the high-speed data links that connect the country to the international Internet; while Saudi internet users may subscribe to any of a number of local internet service providers, all Web traffic is apparently forwarded through a central array of proxy servers at the ISU, which implements Internet content filtering roughly in line with parts of the Resolution. If a user's requested URL is found on the Saudi blacklist, the user is directed to a page that explicitly informs him or her that access to the site has been denied. The ISU administrative web site explains the implementation of the government's content filtering regime, presents the reasoning behind it, and lets Saudi internet users request that a particular site or URL be blocked or unblocked. Citing to the Qur'an as a basis, the government describes its task with filtering as "preserv[ing] our Islamic values, filtering the Internet content to prevent the materials that contradict with our beliefs or may influence our culture."

In addition to detailing Saudi blocking of sexually explicit content, the ISU web site lists as bannable "pages related to drugs, bombs, alcohol, gambling and pages insulting the Islamic religion or the Saudi laws and regulations." Such non-sexually explicit sites are said to be blocked only upon the direction of security bodies within the Saudi government. The ISU describes its policy as filtering only the "absolute minimum possible number of web pages possible to fulfill its duties."

As with most filtering regimes, whether implemented at the client, ISP, or government level, no list is made available of the sites blocked. We therefore sought to collect and distribute a list of blocked sites and pages -- a list that is large in absolute terms even if small relative to the size of the Internet and to the total amount of blocked content, and a list that is diverse even if not perfectly representative of all blocked content. Such a list allows us and others to begin to assess the nature and scope of filtering within Saudi Arabia, with particular attention to non-sexually explicit Web sites rendered inaccessible there. Having requested some 64,557 distinct web pages and found 2,038 to be blocked, we conclude that Saudi Arabia indeed blocks a range of web content beyond that which is sexually explicit. For example, we found blocking of at least 246 pages indexed by Yahoo as Religion (including 67 about Christianity, 45 about Islam, 22 about Paganism, 20 about Judaism, and 12 about Hinduism). We also found blocking of 76 pages within Yahoo's humor categories, 70 within music categories, and 43 within movies, and we found 13 blocked pages about homosexuality. Taken as a whole, the Saudi government's stated blocking criteria are quite broad, making it difficult to assess whether the blocking of a given site is consistent with the criteria. However, a look at the list beyond sexually explicit content yields some insight into the particular areas the Saudi government appears to find most sensitive.

In future work, the authors intend to expand analysis to Internet filtering systems in other countries and to generate URLs to test based on queries invoked in the local language. Sign up to receive updates. The authors are also developing a distributed application for use by Internet users worldwide in testing, analyzing, and documenting respective Internet filtering regimes. Get more information and sign up to get involved.

Specific Pages Found to be Blocked

With the permission and cooperation of ISU staff, we obtained access to the ISU's proxy servers from May 14 to May 27, 2002. During that time we requested 64,557 distinct URLs drawn from various web indices, and we were

able to determine which specific Web pages among them were blocked from within Saudi Arabia. We found that entire sites could be filtered, or individual pages within them.

Filtering of Sexually Explicit Content

A preliminary round of testing examined 795 distinct URLs containing sexually explicit images. These URLs had been used as the basis for a portion of one author's expert testimony in American Library Association v. United States, 201 F.Supp.2d 401 (E.D.Pa., 2002). An expert for the plaintiffs had generated this list by collecting all 797 results from Google in response to an October 2001 Web search using the search criteria "free adult sex," less two pages removed because they were found not to include sexually explicit images. Of these 795 pages, 685 (86.2%) were blocked while 110 (13.8%) were accessible.

Filtering of Other Content

Our main testing examined 63,762 web pages drawn from categories other than sexually explicit content. These pages were extracted from selected areas of the Yahoo Directory (detailed below); from Google's "Similar Pages" feature (requesting pages similar to pages in certain Yahoo categories); and from ordinary Google searches. Of the tested pages, a total of 1,353 were found to be blocked. Some of these blocked pages may fit the second half of Saudi Arabia's stated blocking profile ("related to drugs, bombs, alcohol, gambling, and pages insulting the Islamic religion or the Saudi laws and regulations"), a small number may actually be sexually explicit, while still others may be examples of overblocking, i.e. blocking of pages beyond Saudi Arabia's stated blocking criteria.

Given the large number of pages blocked, we have organized our listing of specific blocked pages into highlights (a subset of blocked pages that are well known or otherwise of possible interest) followed by the full list. Where available, each page's listing includes its HTML title as well as META keywords and description, its Yahoo Directory and Google Directory classifications, and information about past snapshots of the page available in the Internet library archive.org. These details are as retrieved in June 2002.

Specific web pages blocked in Saudi Arabia

Highlights of blocked pages - pages that are well known or otherwise of particular interest

Complete list of 1,353 pages, sorted alphabetically by URL
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z numbers

Raw testing results (.ZIP file, >800KB)

For the duration of our limited access to the ISU proxy server system, we retested pages initially found to be blocked in order to determine whether blocking continued over time and whether ISU staff ever reversed decisions to block certain content. Our testing indicated that four blocked pages became unblocked during the course of testing: swim-n-sport.com (an online swimsuit catalog) was blocked on May 14, 19, and 22, but was accessible in testing of May 24 and 27. The front page and one additional page of theonion.com (an online humor magazine) were found to be blocked on May 19 and May 22, but they too were accessible in testing of May 24 and 27. Finally, warfarerecords.net (a pay-per-click search engine) was blocked in testing of May 14, 19, and 22, but it was also accessible on May 24 and 27. Our inference from these results is that ISU staff may periodically revisit blocked site logs to restore access to certain blocked sites; however, given the small number of sites unblocked during the sampled time period, we are uncertain of the prevalence of this procedure.

Analysis & Summary Statistics

The blocked web pages cover a wide variety of substantive areas. To get a better sense of the types of pages blocked, we have organized the blocked sites within the Yahoo hierarchy where possible. For each blocked URL, Yahoo categories were obtained by entering the blocked URL into Yahoo's ordinary search interface.

Of the 884 web pages with at least one listing in Yahoo's web directory, pages were included in the Yahoo categories as reported in the following tables:

Blocked pages by Yahoo category - collapsible outline (requires Internet Explorer)

Printer-friendly version:
Count of pages found to be blocked in Saudi Arabia

by Yahoo top-level category, second-level category, third-level category, and fourth-level category

Among the specific blocked pages are the following categories of content:

- *Religion*. A total of 246 pages were blocked from Yahoo Religion categories, including Christianity (67 pages blocked), Islam (45), Paganism (22), Judaism (20), and Hinduism (12). An additional 11 pages placed by Yahoo within the Religion section of Business and Economy were also found to be blocked. Specific blocked pages included substantial portions (including the home pages) of religioustolerance.org ("an agency promoting religious tolerance as a human right"), answering-islam.org ("A Christian-Muslim Dialog" [sic.]), and al-bushra.org (a Web site calling for "brotherhood and love" between religions).
- *Health*. Blocked health pages included information about specific diseases, treatments, and prevention methods. 8 blocked pages describe mental health specifically, 3 describe abortion, and 2 describe other aspects of women's health. 18 additional pages described illegal drugs, the war on drugs, and their effects and risks.
- *Education and reference*. Specific blocked web pages providing education and reference content include women.eb.com (the Women in American History section of Encyclopedia Britannica Online), home.bip.net/hyla (the Islamic Cultural Library), and channels.nl/amsterdam/annefran.html (the Anne Frank House).
- *Sites providing information specifically to and about women*. Blocked pages include ivillage.com ("The Women's Network - Busy women sharing solutions and advice"), skirtmag.com ("Skirt Magazine for Women Online"), teenwire.com ("Sexuality and relationship info you can trust from Planned Parenthood Federation of America"), and the previously-mentioned Women in American History section of Encyclopedia Britannica Online.
- *Humor sites*. A total of 81 blocked pages were categorized, by their own authors or by Yahoo, as providing humor content; some were, by their own descriptions, "off-color" or "offensive." Example sites include createafart.com, poopreport.com, and jokes about Monica Lewinsky.
- *Entertainment, music, and movies*. Blocked content includes 251 distinct pages classified, by their authors or by Yahoo, as providing music, movies, or other forms of entertainment. Specific blocked pages include foxsearchlight.com (Fox Searchlight Pictures), rollingstone.com (the Rolling Stone magazine), and wbr.com (Warner Brothers Records).
- *Sites providing information to the gay community*. 13 pages were blocked from Yahoo's Society - Cultures and Groups - Lesbians, Gays, and Bisexuals category. Blocked pages included listings of regional organizations, support groups, and news coverage, as well as pages providing information of specific interest to Muslim gays and/or to gays living in Muslim countries.
- *Pages perceived to be hostile to Saudi Arabia*. Among the specific pages blocked were numerous Amnesty International pages about Saudi Arabia and the saudiinstitute.org reports on Human Rights in Saudi Arabia. While blocked content on these sites seemed to be restricted to that portion of content specific to Saudi Arabia -- top-level site home pages were not found to be blocked -- these pages are nonetheless of particular interest since they are produced by well-known international human rights organizations.
- *Pages about Middle Eastern politics, organizations, or groups*. Various blocked pages provided content likely to be controversial in the context of modern Middle Eastern politics. Example sites include hizbollah.org and idf.il (the Israel Defense Force).
- *Services allowing circumvention of filtering restrictions*. Certain web sites allow a user to view other web sites; such sites include translation services, proxies, and archives. Numerous such pages were blocked, including translators provided by systransoft.com, Altavista/Babelfish, and dictionary.com, as well as the anonymizer.com and megaproxy.com proxy servers.
- *Swimsuits, lingerie, modeling, and other non-pornographic human images*. Pages were blocked from Yahoo categories that suggest the display of images of people wearing less clothes than is typical in Saudi Arabia. For example, 28 pages were blocked from Yahoo's Swimming & Diving category.
- *Pornography*. The majority of the Google "free adult sex" pages were blocked by Saudi Arabia's filtering system. It is likely that blocking of all 795 Google "free adult sex" pages would be consistent with Saudi Arabia's desire to block pornography. Accordingly, the 110 such pages that were not blocked were likely examples of underblocking; within this sample of relatively well-known sexually-explicit pages, Saudi Arabia's correct blocking rate is about 86% and its underblocking rate is about 14%. In addition, Saudi Arabia was found to block sites recently reregistered after prior domain registrants allowed their domain registrations to

lapse; many such sites come to provide sexually-explicit content, as documented in one author's prior Domains Reregistered for Distribution of Unrelated Content: A Case Study of "Tina's Free Live Webcam".

Among the pages tested were many thousands not affected by the Saudi filtering system. We attempted to access many sites based on our initial knowledge of what content is blocked in other countries worldwide and of what content might be of particular concern to the Saudi Arabian government. We found that news sites, US government sites, and Israeli government sites (excluding the Israel Defense Force) could all be viewed as usual. We also found that the overwhelming majority of education sites remained accessible.

Conclusions and Future Work

Since our listing of blocked pages is not and cannot be perfectly representative of content blocked in Saudi Arabia, it is difficult to draw sweeping conclusions about the Saudi blocking system. On the basis of the blocked sites we have found, we do conclude (1) that the Saudi government maintains an active interest in filtering non-sexually explicit Web content from users within the Kingdom; (2) that substantial amounts of non-sexually explicit Web content is in fact effectively inaccessible to most Saudi Arabians; and (3) that much of this content consists of sites that are popular elsewhere in the world.

Use of others' work to assist filtering. The ISU reports that it delegates to its filtering software provider the preparation of a list of pornographic sites to be blocked. Should the ISU choose to "farm out" such work, those reviewing sites or creating filtering lists can be anywhere in the world and still, from a technical perspective, effectively implement their blocks within Saudi Arabia. Such delegation also accords with a New York Times account from November 2001 which described the competition among nearly a dozen mostly American software companies to provide content filters and reported that Secure Computing's Smartfilter was currently in place. ("Companies Compete to Provide Internet Veil for the Saudis," New York Times, November 19, 2001. Archived at websense.com.) Accordingly, it is likely that the Saudi Arabian blocking system inherits whatever categorization errors are made by the current provider of proxy and filtering software; some such errors are documented in one author's previous Sites Blocked by Internet Filtering Programs. While ISU's "filtering procedure" page reports that Saudi Arabia blocks sexually-explicit content on the basis of determinations made by its filtering software provider, reviewing the list of specific blocked pages suggests that the ISU may also have engaged categories of the filtering program that pertain to drugs and to personal home pages. Smartfilter includes both of these categories in its control list.

Indeed, the Yahoo categories that provided the basis for a portion of our queries to the Saudi proxy servers could themselves be used to help determine sites and pages for blocking. However, review of Yahoo-listed sites blocked suggests that there has been no wholesale adoption by the Saudi filterers of Yahoo categories listing Web pages within sensitive substantive areas.

Effectiveness of the Web filtering regime. The significance of the contents of the Saudi filters depends in part on the robustness of the filtering system against those who seek to bypass it. One common method of bypassing a filtering system is via independent, non-filtered proxy servers that can intermediate access requests. For example, a Saudi user might request from megaproxy.com that megaproxy.com give the user a copy of some blocked page; if the Saudi user can access megaproxy, this approach ordinarily bypasses Saudi filtering since megaproxy's Internet access is unfettered by Saudi network policy. However, the Saudi filtering system blocks access to megaproxy.com as well as a large number of other well-known proxy servers, suggesting that Saudi filtering administrators are well aware of this loophole and have sought to close it. Such "loophole" sites include not just proxy servers but also privacy protection systems and web page translators; further testing shows that such services are also blocked in Saudi Arabia.

Since the best-known methods of circumventing filters are blocked in Saudi Arabia, our sense to date is that the Saudi filtering system is likely relatively effective in constraining the information accessed by most Saudis. At the same time, we expect that the tech-savvy users can devise new methods to circumvent blocking. However, should savvy users share their methods with many additional users, Saudi network staff would likely work to close newly-exposed loopholes; we therefore conclude that filtering is likely to remain effective over time. In addition, since Saudi network staff can review access logs of accepted web requests, even expert Internet users can never fully know whether a given circumvention method will yet yield an investigation or even criminal sanctions by Saudi network staff. It remains unknown whether other methods of circumventing filtering -- peer-to-peer applications, for example -- are successful or even usable on the Saudi network. The authors' tests were limited to ordinary http requests lodged on default port 80 of the desired Web pages.

Popularity of sites blocked. The significance of the Saudi blocking system depends in part on the relative popularity of blocked sites; if blocked sites would be frequently accessed by Saudis (if accessible to them at all), the blocking is in a certain sense more constraining than if the blocked sites would be of little interest. Certain of the sites found to be blocked seem to be quite popular without specific reference to localized surfing variations, as measured by the number of inbound links from other Web pages. Google reports that 48,700 distinct links point to pages at the ivillage.com Women's Network (all of which appears to be blocked); 18,100 to the cards.webshots.com eCards site; 15,300 to the terra.es Spanish-language portal; 13,100 to the theonion.com humor magazine; and 9,470 to the systransoft.com translator. Furthermore, archive.org change-tracking histories report that many of the blocked sites change frequently; the rollingstone.com magazine site was found to offer at least 461 distinct front pages between 1997 and 2001; the hecklers.com comedy site, 263; the brutal.com news site, 150. While Saudi Internet users may seek access to sites other than those most linked by Internet authors worldwide and other than those that change most frequently, these link and change counts suggest that at least some of the blocked sites are of substantial interest to Internet users.

Future work might seek to investigate some or all of the following issues:

- Documentation of additional blocked sites, including sites indexed under Arabic language searches.
- Changes in blocking over a more extended time period, with possible relation to official Saudi government shifts in views on particular issues or foreign countries.
- Nature and timeliness of responses to requests for blocking and unblocking of specified pages and sites.
- Effectiveness of circumvention methods.
- Whether Saudi filtering systems make the same blocking errors (i.e. overblocking) as ordinary installations of commercial filtering systems.

---

* Jack N. and Lillian R. Berkman Assistant Professor of Entrepreneurial Legal Studies, Harvard Law School.
** J.D. Candidate, Harvard Law School, 2005.

Support for this project was provided by the Berkman Center for Internet & Society at Harvard Law School.

Last Updated: September 12, 2002 - Sign up for notification of major updates and related work.

R. Faris and N. Villeneuve (2008). "Measuring Global Internet Filtering." in R. Deibert et al. (eds.), *Access denied: The practice and policy of global Internet filtering.* Cambridge, MA: MIT Press: 5-27.

1
Measuring Global Internet Filtering

Robert Faris and Nart Villeneuve

The Scope and Depth of Global Internet Filtering

In this chapter, we set out to provide an overview of the data regarding Internet filtering that the OpenNet Initiative[1] has gathered over the past year. Empirical testing for Internet blocking was carried out in forty countries in 2006. Of these forty countries, we found evidence of technical filtering in twenty-six (see table 1.1). This does not imply that only these countries filter the Internet. The testing we carried out in 2006 constitutes the first step toward a comprehensive global assessment. Not only do we expect to find more countries that filter the Internet as we expand our testing, but we also expect that some of the countries that did not show signs of filtering in 2006 will institute filtering in subsequent years.[2]

Conceptually, the methodology we employ is simple. We start by compiling lists of Web sites that cover a wide range of topics targeted by Internet filtering. The topics are organized into a taxonomy of categories that have been subject to blocking, ranging from gambling, pornography, and crude humor to political satire and Web sites that document human rights abuses and corruption. (See table 1.2.) Researchers then test these lists to see which Web sites are available from different locations within each country.[3]

The states that filter the Internet must choose which topics to block (the scope of filtering) and how much of each topic to filter (the depth of filtering). The results of these decisions are summarized in figure 1.1, comparing the breadth and depth of filtering for the countries where evidence of filtering was found.

The number of different categories in which Internet filtering was found to occur is shown on the horizontal axis. We put this forward as a measure of the scope of Internet filtering in each country. (The categories are shown in table 1.2.)

The vertical axis depicts the comprehensiveness of filtering efforts as measured by the highest degree of content blocked in any of the topical categories. This captures a markedly different angle on filtering. If the breadth of filtering represents the ambition of censors to limit information related to a range of topics, the depth of filtering measures the success in actually blocking content. This might correspond to the level of sophistication of the filtering regime

and amount of resources devoted to the endeavor, or it may be a reflection of the resolve and political will to shut down large sections of the Internet.

The countries occupying the upper right of figure 1.1, including Iran, China, and Saudi Arabia, are those that not only intercede on a wide range of topics but also block a large amount of content relating to those topics. Myanmar and Yemen cover a similarly broad scope, though with less comprehensiveness in each category. South Korea is in a league of its own. It has opted to filter very little, targeting North Korean sites, many of which are hosted in Japan. Yet South Korea's thoroughness in blocking these sites manifests a strong desire to eliminate access to them. There is a cluster of states occupying the center of the plot that

Table 1.1
Filtering by state

Evidence of filtering	Suspected filtering	No evidence of filtering
Azerbaijan	Belarus	Afghanistan
Bahrain	Kazakhstan	Algeria
China		Egypt
Ethiopia		Iraq
India		Israel
Iran		Kyrgyzstan
Jordan		Malaysia
Libya		Moldova
Morocco		Nepal
Myanmar		Russia*
Oman		Ukraine
Pakistan		Venezuela
Saudi Arabia		West Bank/Gaza
Singapore		Zimbabwe
South Korea		
Sudan		
Syria		
Tajikistan		
Thailand		
Tunisia		
United Arab Emirates		
Uzbekistan		
Vietnam		
Yemen		

*Testing in Russia was limited to a selection of ISPs in Moscow; these preliminary results may not extend beyond this sample.

Table 1.2
Categories subject to Internet filtering

Free expression and media freedom
Political transformation and opposition parties
Political reform, legal reform, and governance
Militants, extremists, and separatists
Human rights
Foreign relations and military
Minority rights and ethnic content
Women's rights
Environmental issues
Economic development
Sensitive or controversial history, arts, and literature
Hate speech
Sex education and family planning
Public health
Gay/lesbian content
Pornography
Provocative attire
Dating
Gambling
Gaming
Alcohol and drugs
Minority faiths
Religious conversion, commentary, and criticism
Anonymizers and circumvention
Hacking
Blogging domains and blogging services
Web hosting sites and portals
Voice over Internet Protocol (VOIP)
Free e-mail
Search engines
Translation
Multimedia sharing
P2P
Groups and social networking
Commercial sites

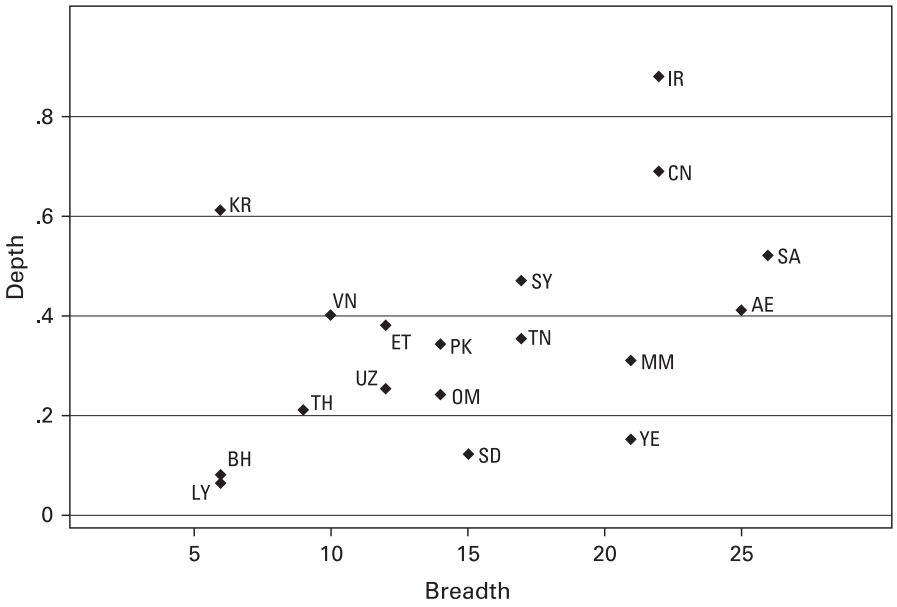


Figure 1.1
Comparing the breadth and depth of filtering. AE—United Arab Emirates; BH—Bahrain; CN—China; ET—Ethiopia; IR—Iran; JO—Jordan; KR—South Korea; LY—Libya; MM—Burma/Myanmar; OM—Oman; PK—Pakistan; SA—Saudi Arabia; SD—Sudan; SY—Syria; TH—Thailand; TH—Tunisia; UZ—Uzbekistan; VN—Vietnam; YE—Yemen. A number of countries that filter a small number of sites are omitted from this diagram, including Azerbaijan, Belarus, India, Jordan, Kazakhstan, Morocco, Singapore, and Tajikistan.

are widely known to practice filtering. These countries, which include Syria, Tunisia, Vietnam, Uzbekistan, Oman, and Pakistan, block an expansive range of topics with considerable depth. Ethiopia is a more recent entrant into this category, having extended its censorship of political opposition into cyberspace.

Azerbaijan, Jordan, Morocco, Singapore, and Tajikistan filter sparingly, in some cases as little as one Web site or a handful of sites. The evidence for Belarus and Kazakhstan remains inconclusive at the time of this writing, though blocking is suspected in these countries.

Of equal interest are the states included in testing in 2006 in which no evidence of filtering was uncovered (see table 1.1). We make no claims to have proven the absence of filtering in these countries. However, our background research supports the conclusion drawn from the technical testing that none of these states are currently filtering Internet content.[4]

Later in the book we turn our attention to the question of why some countries filter and others do not, even under similar political and cultural circumstances.

The Principal Motives and Targets of Filtering

On September 19, 2006, a military-led coup in Thailand overthrew the democratically elected government headed by Prime Minister Thaksin Shinawatra. Thailand is not unfamiliar with such upheavals. There have been seventeen coups in the past sixty years. This time, however, Internet users noticed a marked increase in the number of Web sites that were not accessible, including several sites critical of the military coup.[5] A year earlier in Nepal, the king shut down the Internet along with international telephone lines and cellular communication networks when he seized power from the parliament and prime minister. In Bahrain, during the run-up to the fall 2006 election, the government chose to block access to a number of key opposition sites. These events are part of a growing global trend. Claiming control of the Internet has become an essential element in any government strategy to rein in dissent—the twenty-first century parallel to taking over television and radio stations.

In contrast to these exceptional events, the constant blocking of a swath of the Internet has become part of the everyday political and cultural reality of many states. A growing number of countries are blocking access to pornography, led by a handful of states in the Persian Gulf region. Other countries, including South Korea and Pakistan, block Web sites that are perceived as a threat to national security.

Notwithstanding the wide range of topics filtered around the world, there are essentially three motives or rationales for Internet filtering: politics and power, social norms and morals, and security concerns. Accordingly, most of the topics subject to filtering (see table 1.2) fall under one of three thematic headings: political, social, and security. A fourth theme—Internet tools—encompasses the networking tools and applications that allow the sharing of information relating to the first three themes. Included here are translation tools, anonymizers, blogging services, and other Web-based applications categorized in table 1.2.

Protecting intellectual property rights is another important driver of Internet content regulation, particularly in western Europe and North America. However, in the forty countries that were tested in 2006, this is not a major objective of filtering.[6]

Figure 1.2 compares the political and social filtering practices of these same twenty-seven countries. On one extreme is Saudi Arabia, which heavily censors social content. While there is also substantial political filtering carried out in Saudi Arabia, it is done with less scope and depth. On the other fringe are Syria and China, focusing much more of their extensive filtering on political topics. Myanmar and Vietnam are also notable for their primary focus on political issues, which in the case of Vietnam contradicts the stated reason for filtering the Internet.[7] Iran stands out for its pervasive filtering of both political and social material.

Filtering directed at political opposition to the ruling government is a common type of blocking that spans many countries. Politically motivated filtering is characteristic of authoritarian and repressive regimes. The list of countries that engage in substantial political blocking includes Bahrain, China, Libya, Iran, Myanmar, Pakistan, Saudi Arabia, Syria, Tunisia,

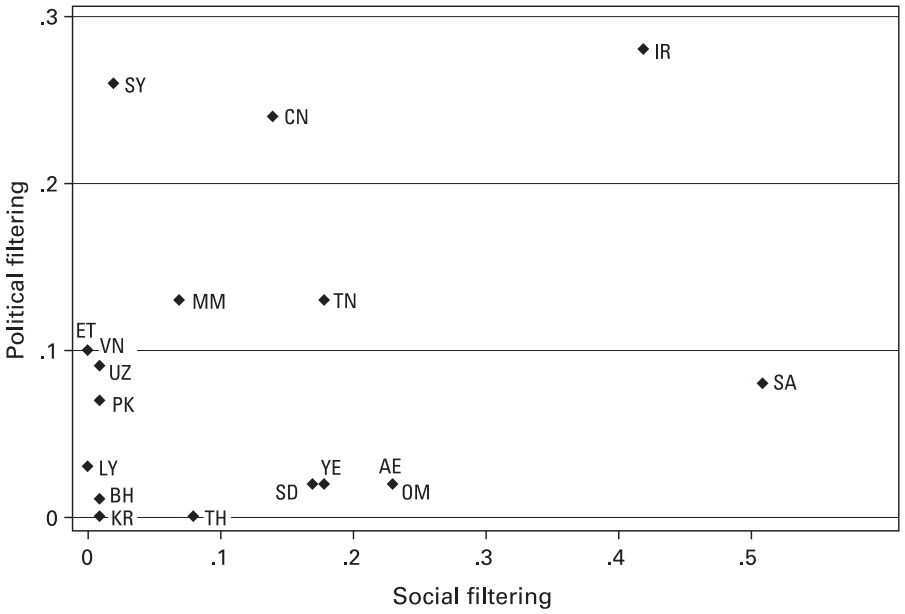


Figure 1.2
Political and social filtering. AE—United Arab Emirates; BH—Bahrain; CN—China; ET—Ethiopia; IR—Iran; JO—Jordan; KR—South Korea; LY—Libya; MM—Burma/Myanmar; OM—Oman; PK—Pakistan; SA—Saudi Arabia; SD—Sudan; SY—Syria; TH—Thailand; TH—Tunisia; UZ—Uzbekistan; VN—Vietnam; YE—Yemen. A number of countries that filter a small number of sites are omitted from this diagram, including Azerbaijan, Belarus, India, Jordan, Kazakhstan, Morocco, Singapore, and Tajikistan.

Uzbekistan, and Vietnam.[8] Thailand and Ethiopia are the most recent additions to this group of countries that filter Web sites associated with political opposition groups. Yet in other countries with an authoritarian bent, such as Russia and Algeria, we have not uncovered filtering of the Internet.

The perceived threat to national security is a common rationale used for blocking content. Internet filtering that targets the Web sites of insurgents, extremists, terrorists, and other threats generally garners wide public support. This is best typified by South Korea where pro–North Korean sites are blocked, or by India where militant and extremist sites associated with groups that foment domestic conflict are censored. In Pakistan, Web sites devoted to the Balochi independence movement are blocked. Similarly, the Web sites of separatist or radical groups such as the Muslim Brotherhood are blocked in some countries in the Middle East.

Social filtering is focused on those topics that are held to be antithetical to accepted societal norms. Pornographic, gay and lesbian, and gambling-related content are prime examples

Box 1.1
Identifying and documenting Internet filtering

Measuring and describing Internet filtering defies simple metrics. Ideally, we would like to know how Internet censorship reduces the availability of information, how it hampers the development of online communities, and how it inhibits the ability of civic groups to monitor and report on the activities of the government, as these answers impact governance and ultimately economic growth. However, this is much easier to conceptualize at an abstract level than to measure empirically. Even if we were able to identify all the Web sites that have been put out of reach due to government action, the impact of blocking access to each Web site is far from obvious, particularly in this networked world where information has a habit of propagating itself and reappearing in multiple locations. Nevertheless, every obstacle thrown into the path of citizens seeking out information bears a cost or, depending on how one views the contribution of a particular Web site to society, a benefit. With this recognition of the inherent complexity of evaluating Internet censorship, we set out with modest goals—to identify and document filtering.

Two lists of Web sites are checked in each of the countries tested: a global list and a local list. The global list is a standardized list of Web sites that cover the categories listed in table 1.1. The global list of Web sites is comprised principally of internationally relevant Web sites with English content. The same global list is checked in each of the countries in which we have tested. A separate local list is created for each of the countries tested; it includes Web sites related to the specific issues and context of the study country.

These testing lists encompass a wide variety of content including political topics such as human rights, political commentary and news, religion, health and sex education, and Web sites sponsored by separatists and militant organizations. Pornography, gambling, drugs, and alcohol are also represented in the testing lists. The lists embody portions of the Web space that would be subject to Internet filtering in each of the countries being tested. They are designed to unearth filtering and blocking behavior where it exists. Background research is focused on finding sites that are likely to be blocked. In countries where Internet censorship has been reported, the lists include those sites that were alleged to have been blocked. These are not intended to be exhaustive lists of the relevant subject matter, nor do we presume to have identified all the Web sites that are subject to blocking.

The actual tests are run from within each country using software specifically designed for this purpose. Where appropriate, the tests are run from different locations to capture the differences in blocking behavior across Internet service providers (ISPs). The tests are run across multiple days and weeks to control for normal connectivity problems.

The completion of the initial accessibility testing is just the first step in the evaluation process. Additional diagnostic work is required to separate normal connectivity errors from intentional tampering. As described in further detail later, there are a number of technical alternatives for filtering the Internet, some of which are relatively easy to discover. Others are difficult to detect and require extensive diagnostic work to confirm.

of what is filtered for social and cultural reasons. Hate speech and political satire are also the target of Internet filtering in some countries. Web sites that deny the Holocaust or promote Nazism are blocked in France and Germany. Web sites that provide unflattering details related to the life of the king of Thailand are censored in his country.

An emergent impetus for filtering is the protection of existing economic interests. Perhaps the best example is the blocking of low-cost international telephone services that use Voice-over Internet Protocol (VoIP) and thereby reduce the customer base of large telecommunications companies, many of which enjoy entrenched monopoly positions. Skype, a popular and low-cost Internet-based telephone service, has been blocked in Myanmar and United Arab Emirates, which heavily block VoIP sites. The Web sites of many VoIP companies are also blocked in Syria and Vietnam.

Many countries seek to block the intermediaries: the tools and applications of the Internet that assist users in accessing sensitive material on the Internet. These tools include translation sites, e-mail providers, Weblog hosting sites, and Web sites that allow users to circumvent standard blocking strategies. Blogging services such as Blogspot are often targeted; eight countries blocked blogs hosted there, while Syria, Ethiopia, and Pakistan blocked the entire domain, denying access to all the blogs hosted on Blogspot. Fourteen countries blocked access to anonymity and censorship circumvention sites. Both SmartFilter, used in Sudan, Tunisia, Saudi Arabia, and UAE, and Websense, used in Yemen, have filtering categories—called ''Anonymizers'' and ''Proxy Avoidance,'' respectively—used to block such sites.

A handful of countries, including China, Vietnam, and states in the MENA region (the Middle East and North Africa), block Web sites related to religion and minority groups. In China, Web sites that represent the Falun Gong and the Tibetan exile groups are widely blocked. In Vietnam, religious and ethnic sites associated with Buddhism, the Cao Dai faith, and indigenous hill tribes are subject to blocking. Web sites that are aimed at religious conversion from Islam to Christianity are often blocked in the MENA region. Decisively identifying the motives of filtering activity is often impossible, particularly as the impact of filtering can simultaneously touch a host of social and political processes. That being said, it probably would be a mistake to attribute the filtering of religious and ethnic content solely to biases against minority groups, as these movements also represent a political threat to the ruling regimes.

A Survey of Global Filtering Strategies, Transparency, and Consistency

There are many techniques used to block access to Internet content. Each of these techniques can be used at different levels of Internet access within a country. Internet filtering is most commonly implemented at two levels: at the ISPs within the country and on the Internet backbone at the international gateway. These methods may overlap; an ISP may filter content using one particular technique while another technique is used at the international gateway.

Pakistan is an example of a country that blocks at both the international gateway and at the ISP level.

There are a few principal techniques used for Internet filtering including IP blocking, DNS tampering, and proxy-based blocking methods. (For blocking behavior by country, see table 1.3.) These techniques are presented in further detail by Anderson and Murdoch in chapter 3.

IP blocking is effective in blocking the intended target and no new equipment needs to be purchased. It can be implemented in an instant; all the required technology and expertise is

Table 1.3
Blocking techniques

	IP blocking	DNS tampering	Blockpage	Keyword
Azerbaijan	X		X	
Bahrain		X	X	
China	X			X
Ethiopia	X			
India	X	X		
Iran			X	X
Jordan	X			
Libya	X			
Myanmar			X	
Oman			X	
Pakistan	X	X		
Saudi Arabia			X	
Singapore			X	
South Korea	X	X	X	
Sudan			X	
Syria			X	
Thailand			X	
Tunisia			X	
United Arab Emirates			X	
Uzbekistan*			X	
Vietnam		X	X	
Yemen			X	X

Blocking behavior included in this table may include international gateway level filtering, and filtering techniques used by different ISPs.

* In Uzbekistan, the blockpage does not clearly indicate that filtering is occurring but rather redirects users to a third-party Web site.

readily available. Depending on the network infrastructure within the country it may also be possible to block at or near the international gateways so that the blocking is uniform across ISPs.

Countries new to filtering will generally start with IP blocking before moving on to more expensive filtering solutions. ISPs most often respond quickly and effectively to blocking orders from the government or national security and intelligence services. Therefore they block what is requested in the cheapest way using technology already integrated into their normal network environment. Blocking by IP can result in significant overblocking as all other (unrelated) Web sites hosted on that server will also be blocked.

China uses IP blocking to obstruct access to at least three hundred IP addresses. This blocking is done at the international gateway level affecting all users of the network regardless of ISP. The IPs blocked among the two backbone providers, China Netcom and ChinaTelecom, are remarkably similar.[9]

The ISP ETC-MC in Ethiopia uses IP blocking to block, among other sites, Google's Blogspot blogging service. This results in all Blogspot blogs being blocked in Ethiopia. Pakistan implements IP blocking at the international gateway level. In addition to blocking the IP for Blogspot, they also block Yahoo's hosting service, which results in major overblocking. For example, in targeting www.balochvoice.com they are actually blocking more than 52,000 other Web sites hosted on that same server.

DNS tampering is achieved by purposefully disrupting DNS servers, which resolve domain names into IP addresses. Generally, each ISP maintains its own DNS server for use by its customers. To block access to particular Web sites, the DNS servers are configured to return the wrong IP address. While this allows the blocking of specific domain names, it also can be easily circumvented by simple means such as accessing an IP address directly or by configuring your computer to use a different DNS server.

In Vietnam, the ISP FPT configures DNS to not resolve certain domain names, as if the site does not exist. The ISP Cybernet in Pakistan also uses this technique. The ISP Batelco in Bahrain uses this technique for some specific opposition sites. Batelco did not, however, completely remove the entry (the MX record for e-mail still remains). In India, the ISP BHARTI resolves blocked sites to the invalid IP address 0.0.0.0 while the ISP VSNL resolves blocked sites to the invalid IP address 1.2.3.4. The South Korean ISP, Hananet, uses this technique but makes the blocked Web site resolve to 127.0.0.1. This is the IP address for the ''localhost.'' Another South Korean ISP, KORNET, makes blocked sites resolve to an ominous police Web site. This represents an unusual case in which DNS tampering resolves to a blockpage.[10]

Our tests revealed that there is often a combination of IP blocking and DNS tampering. It may be a signal that countries are responding to the outcry concerning the overblocking associated with IP blocking and moving to the targeting of specific domain names with DNS tam-

pering. In India, for example, the Internet Service Providers Association of India reportedly has sent instructions to ISPs showing how to block by DNS instead of by IP.[11]

In proxy-based filtering strategies, Internet traffic passing through the filtering system is reassembled and the specific HTTP address being accessed is checked against a list of blocked Web sites. These can be individual domains, subdomains, specific long URL paths, or keywords in the domain or URL path. When users attempt to access blocked content they are subsequently blocked. An option in this method of filtering is to return a *blockpage* that informs the user that the content requested has been blocked.

Saudi Arabia uses SmartFilter as a filtering proxy and displays a blockpage to users when they try to access a site on the country's block list. The blockpage also contains information on how to request that a block be lifted. Saudi Arabia blocks access to specific long URLs. For example, www.humum.net/ is accessible, while www.humum.net/country/saudi.shtml is blocked. United Arab Emirates, Oman, Sudan, and Tunisia also use SmartFilter. Tunisia uses SmartFilter as a proxy to filter the Internet. But instead of showing users a blockpage indicating that the site has been blocked, they have created a blockpage that looks like the Internet Explorer browser's default error page (in French), presumably to disguise the fact that they are blocking Web sites.

A proxy-based filtering system can also be programmed such that Internet traffic passing through the filtering system is reassembled and the specific HTTP address requested is checked against a list of blocked keywords. No country that ONI tested blocked access to a Web site as a result of a keyword appearing in the body content of the page, however, there are a number of countries that block by keyword in the domain or URL path, including China, Iran, and Yemen.

China filters by keywords that appear in the host header (domain name) or URL path. For example, while the site http://archives.cnn.com/ is accessible, the URL http://archives.cnn.com/2001/ASIANOW/east/01/11/falun.gong.factbox/ is not. When this URL is requested, reset (RST) packets are sent that disrupt the connection, presumably because of the keyword *falun.gong*. Iran uses a filtering proxy that displays a blockpage when a blocked Web site is requested. On some ISPs in Iran, such as Shatel and Datak, keywords in URL paths are blocked. This most often affects search queries in search engines. For example, here is a query run on Google for *naked* in Arabic (www.google.com/search?hl=fa&q=%D9%84%D8%AE%D8%AA&btnG=%D8%A8%D9%8A%D8%A7%D8%A8) that was blocked. Ynet in Yemen blocks any URL containing the word *sex*. The domain www.arabtimes.com is blocked in Oman and the UAE but the URL for the Google cached version (http://72.14.235.104/search?q=cache:8utpDVLa1yYJ:www.arabtimes.com/+arabtimes&hl=en&ct=clnk&cd=1) is also blocked because *www.arabtimes.com* appears in the URL path.

Filtering systems can also be configured to redirect users to another Web site. In most cases, redirection is identical to blockpage filtering, the only difference being the route used

to produce the blockpage. ISPs in Iran, Singapore, Thailand, and Yemen all use redirection to a blockpage. Uzbekistan uses redirection but instead of redirecting to a blockpage the filters send users to Microsoft's search engine at www.live.com, suggesting that the government wishes to conceal that fact that blocking has taken place.

There are thus various degrees of transparency in Internet filtering. Where blockpages are used, it is clearly apparent to users when a requested Web site has been intentionally blocked. Other countries give no indication that a Web site is blocked. In some cases, this is a function of the blocking technique being used. Some countries, such as Tunisia and Uzbekistan, appear to deliberately disguise the fact that they are filtering Internet content, going a step farther to conceal filtering activity beyond the failure to inform users that they are being filtered.

Another subset of countries, including Bahrain and United Arab Emirates, employ a hybrid strategy, indicating clearly to users that certain sites are blocked while obscuring the blocking of other sites behind the uncertainty of connection errors that could have numerous other explanations. In Bahrain, users normally receive a blockpage. However, for the specific site www.vob.org, Bahrain uses DNS tampering that results in an error. In United Arab Emirates all blocked sites with the exception of www.skype.com returned a blockpage. There is an apparent two-tiered system in place. They are willing to go on the record as blocking some sites, and not for others.

Providing a blockpage informing a user that their choice of Web site is not available by action of the government is still short of providing a rationale for the blocking of that particular site, or providing a means for appealing this decision. Very few countries go this far. A small group of countries, including Saudi Arabia, Oman, and United Arab Emirates, and some ISPs in Iran, allow Internet users to write to authorities to register a complaint that a given Web site has been blocked erroneously.

Centralized filtering regimes require all Internet traffic to pass through the same filters. This results in a consistent view of the Internet for users within the country; all users experience the same degree of filtering. This is most commonly implemented at the international gateway. When filtering is delegated to the ISP level, and hence decentralized, there may be significant differences among ISPs regarding the filtering techniques used and the content that is filtered. In this case, access to Web sites may vary substantially depending on the blocking choices of individual ISPs. (Table 1.4 presents the use of centralized and/or decentralized filtering strategies across the countries in the study, and the resulting consistency in filtering within each country.) In Iran there is considerable variation in the blocking among ISPs. For example, one ISP blocks considerably less political content than the other six ISPs tested. Only one ISP out of the five tested in Azerbaijan, AzNet, blocks access to a considerable amount of social content, most of which is pornographic, while the others block access to only a single IP address. In Myanmar, there is substantial variation in the filtering between the two ISPs tested. One filters much more pornography, while the other blocks a significantly greater portion of politically oriented Web sites. In the United Arab Emirates, an ISP that serves primarily the free-trade

Table 1.4
Comparing filtering regimes

	Locus	Consistency	Concealed filtering	Transparency and accountability
Azerbaijan	D	Low		Medium
Bahrain	C	High	Yes	Low
China	C and D	Medium	Yes	Low
Ethiopia	C	High	Yes	Low
India	D	Medium		High
Iran	D	Medium		Medium
Jordan	D	High		Low
Libya	C	High	Yes	Low
Morocco	C	High	Yes	Low
Myanmar	D	Low		Medium
Oman	C	High		High
Pakistan	C and D	Medium	Yes	High
Saudi Arabia	C	High		High
Singapore	D	High		High
South Korea	D	High		High
Sudan	C	High		High
Syria	D	High		Medium
Tajikistan	D	Low		Medium
Thailand	D	Medium		Medium
Tunisia	C	High	Yes	Low
United Arab Emirates	D	Low		Medium
Uzbekistan	C and D	High	Yes	Low
Vietnam	D	Low	Yes	Low
Yemen	D	High		Medium

The **Locus** of filtering indicates where Internet traffic is blocked. **C** indicates that traffic is blocked from a central location, normally the Internet backbone, and affects the entire state equally. **D** indicates that blocking is decentralized, typically implemented by ISPs. (Note that this study does not include filtering at the institutional level, for example, cybercafés, universities, or businesses.)

Consistency measures the variation in filtering within a country across different ISPs where applicable.

Concealed filtering reflects either efforts to conceal the fact that filtering is occurring or the failure to clearly indicate filtering when it occurs.

Transparency and accountability corresponds to the overall level of openness in regard to the practice of filtering. It also considers the presence of concealed filtering, the type of notice given to users regarding blocking, provisions to appeal or report instances of inappropriate blocking, and public acknowledgement of filtering policies.

zone has not historically filtered the Internet, while the predominant ISP for the rest of the country has consistently filtered the Internet.

Modifications can be made to the blocking efforts of a country by the authorities at any time. Sites can be added or removed at their discretion. For example, during our tests in Iran the Web site of the *New York Times* was blocked, but for only one day. Some countries have also been suspected of introducing temporary filtering around key time periods such as elections.

Hosting modifications can also be made to a blocked site resulting in it becoming accessible or inaccessible. For example, while Blogspot blogs were blocked in Pakistan due to IP blocking, the interface to update one's blog was still accessible. However, Blogspot has since upgraded its service and the new interface is hosted on the blocked IP, making the interface to update one's blog inaccessible in Pakistan. The reverse is also possible. For example, if the IP address of a Web site is blocked, the Web site may change its hosting arrangement in order to receive a new IP address, leaving it unblocked until the new IP address is discovered and blocked.

Summary Measures of Internet Filtering

To summarize the results of our work, we have assigned a score to each of the countries we studied. This score is designed to reflect the degree of filtering in each of the four major thematic areas: 1) the filtering of political content, 2) social content, 3) conflict- and security-related content, and 4) Internet tools and applications. Each country is given a score on a four-point scale that captures both the breadth and depth of filtering for content of each thematic type (see table 1.5).

- Pervasive filtering is defined as blocking that spans a number of categories while blocking access to a large portion of related content.
- Substantial filtering is assigned where either a number of categories are subject to a medium level of filtering in at least a few categories or a low level of filtering is carried out across many categories.
- Selective filtering is either narrowly defined filtering that blocks a small number of specific sites across a few categories, or filtering that targets a single category or issue.
- Suspected filtering is assigned where there is information that suggests that filtering is occurring, but we are unable to conclusively confirm that inaccessible Web sites are the result of deliberate tampering.

The scores in table 1.5 are subjective evaluations based upon the quantitative information gathered during a year of testing and research. In 2006, we tested thousands of Web sites across more than 120 ISPs in 40 countries, creating a database with close to 200,000 observations. Each observation is in turn based on the conclusion of an average of ten accessibility tests. Despite the breadth of this data, a purely quantitative reporting might be

Table 1.5
Summary of filtering

	Political	Social	Conflict and security	Internet tools
Azerbaijan	●	—	—	—
Bahrain	●●	●	—	●
Belarus	○	○	—	—
China	●●●	●●	●●●	●●
Ethiopia	●●	●	●	●
India	—	—	●	●
Iran	●●●	●●●	●●	●●●
Jordan	●	—	—	—
Kazakhstan	○	—	—	—
Libya	●●	—	—	—
Morocco	—	—	●	●
Myanmar	●●●	●●	●●	●●
Oman	—	●●●	—	●●
Pakistan	●	●●	●●●	●
Saudi Arabia	●●	●●●	●	●●
Singapore	—	●	—	—
South Korea	—	●	●●●	—
Sudan	—	●●●	—	●●
Syria	●●●	●	●	●●
Tajikistan	●	—	—	—
Thailand	●	●●	—	●
Tunisia	●●●	●●●	●	●●
United Arab Emirates	●	●●●	●	●●
Uzbekistan	●●	●	—	●
Vietnam	●●●	●	—	●●
Yemen	●	●●●	●	●●

●●● Pervasive filtering; ●● Substantial filtering; ● Selective filtering; ○ Suspected filtering; — No evidence of filtering.

misleading unless we were able to effectively measure the relative importance of each Web site. For example, the blocking of BBC or Wikipedia represents far more than the blocking of a less prominent Web site. Similarly, blocking a social networking site or a blogging server could have a profound impact on the formation of online communities and on the publication of user-generated content. While Internet users will eventually provide alternatives to recreate these communities on other sites hosted on servers that are not blocked, the transition of a wide community is unlikely given the time, effort, and coordination required to reconstitute a community in another location. At the other extreme, the blocking of one pornographic site will have a minor impact on Internet life if access to thousands of similar sites remains unimpeded. For these reasons, we have decided to summarize the results of testing categorically, considering both the scope and depth of the quantitative testing results, in conjunction with expert opinion regarding the significance of the blocking of individual Web sites.

It is tempting to aggregate the results by summing up the scores in each category. Yet doing so would imply that the blocking of political opposition is equivalent to filtering that supports conservative social values or the fear of national security risks. These competing sets of values suggest that a number of different weighting schemes might be appropriate. In any case, the results are generally quite clear, as the most pervasive filtering regimes tend to filter across all categories.

Country-specific and Global Filtering

A comparison between the blocking of country-specific sites and the blocking of internationally relevant Web sites provides another view of global filtering. Not surprisingly, we found that

Box 1.2
Where we tested

> The decision where to test was a simple pragmatic one—where were we able to safely test and where did we have the most to learn? Two countries did not make the list this year because of security concerns: North Korea and Cuba. Learning more about the filtering practices in these countries is certainly of great interest to us. However, we were not confident that we could adequately mitigate the risks to those who would collaborate with us in these countries.
>
> A number of other countries in Europe and North America that are known to engage in filtering to varying degrees were not tested this year. This decision again was a fairly easy practical choice. The filtering practices in these countries are better understood than in other parts of the world and we therefore had less to contribute here. Many of the countries in Europe focus their Internet filtering activity on child pornography. This is not a topic that we will test for ethical and legal reasons.

the incidence of blocking Web sites in our testing lists was approximately twice as high for Web sites available in a local language compared to sites available only in English or other international languages. Figure 1.3 shows that many countries focus their efforts on filtering locally relevant Web content. Ethiopia, Pakistan, Syria, Uzbekistan, and Vietnam are examples of countries that extensively block local content while blocking relatively few international Web sites. China and Myanmar also concentrate more of their filtering efforts on country-specific Internet content, though they block somewhat more global content. Middle Eastern filtering regimes tend to augment local filtering with considerably more global content. This balance mirrors the use of commercial software, generally developed in the West, to identify and block Internet content.

Table 1.6 shows an alternative view of filtering behavior, looking at the blocking of different types of content providers rather than content. The apparent prime targets of filtering are blogs, political parties, local NGOs, and individuals. In the case of blogs, a number

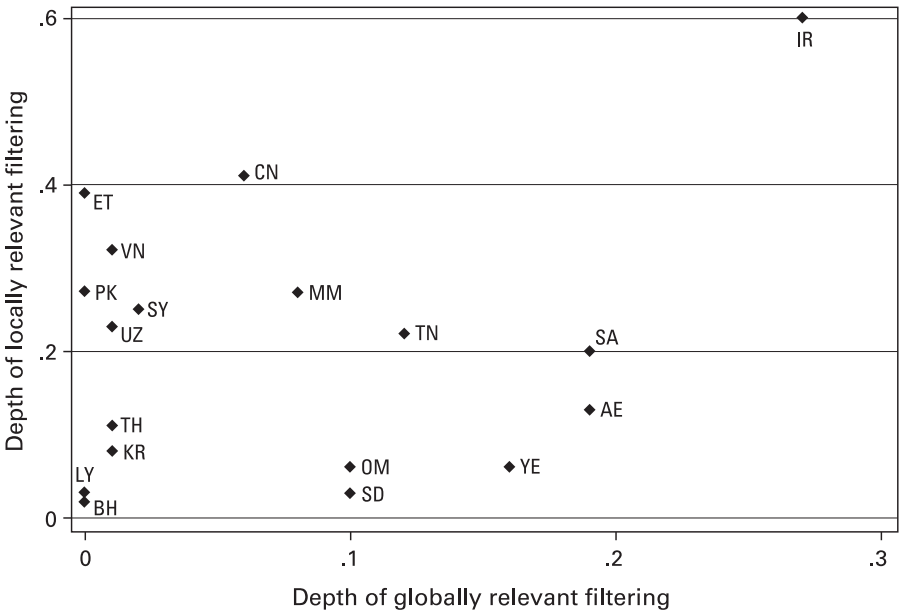


Figure 1.3

Filtering targeted at local sites and global sites. AE—United Arab Emirates; BH—Bahrain; CN—China; ET—Ethiopia; IR—Iran; JO—Jordan; KR—South Korea; LY—Libya; MM—Burma/Myanmar; OM—Oman; PK—Pakistan; SA—Saudi Arabia; SD—Sudan; SY—Syria; TH—Thailand; TH—Tunisia; UZ—Uzbekistan; VN—Vietnam; YE—Yemen. A number of countries that filter a small number of sites are omitted from this diagram, including Azerbaijan, Belarus, India, Jordan, Kazakhstan, Morocco, Singapore, and Tajikistan.

of countries, including Pakistan and Ethiopia, have blocked entire blogging domains, which inflates these figures. Logically, these assessments represent more accurately the result of filtering rather than the intention. Establishing the intention of blocking is never as clear. The blocking of this wide array of blogs could be the result of a lack of technical sophistication or a desire to simultaneously silence the entire collection of blogs hosted on the site.

The other prominent target of filtering is political parties, followed by NGOs focused on a particular region or country, and Web sites run by individuals. The implications of targeting civic groups and individual blogs are addressed by Deibert and Rohozinski in chapter 6 of this volume.

First Steps Toward Understanding Internet Filtering

In this chapter, we summarize what we have learned over the past year regarding the incidence of global Internet filtering. Taking an inventory of filtering practices and strategies is a necessary and logical first step, though still far from a thorough understanding of the issue. The study of Internet filtering can be approached by asking why some states filter the Internet or by asking why others do not. The latter question is particularly apt in countries that maintain a repressive general media environment while leaving the Internet relatively open. This is not

Table 1.6
Blocking by content provider

Content provider type	Portion of content filtered
Academic	0.02
Blogs	0.20
Chat and discussion boards	0.05
Government	0.03
Government media	0.02
International governmental organizations	0.00
Independent media	0.06
Individual	0.09
International NGOs	0.02
Labor groups	0.05
Locally focused NGOs	0.09
Militant groups	0.01
Political parties	0.19
Private businesses	0.06
Religious groups	0.02
Regional NGOs	0.04

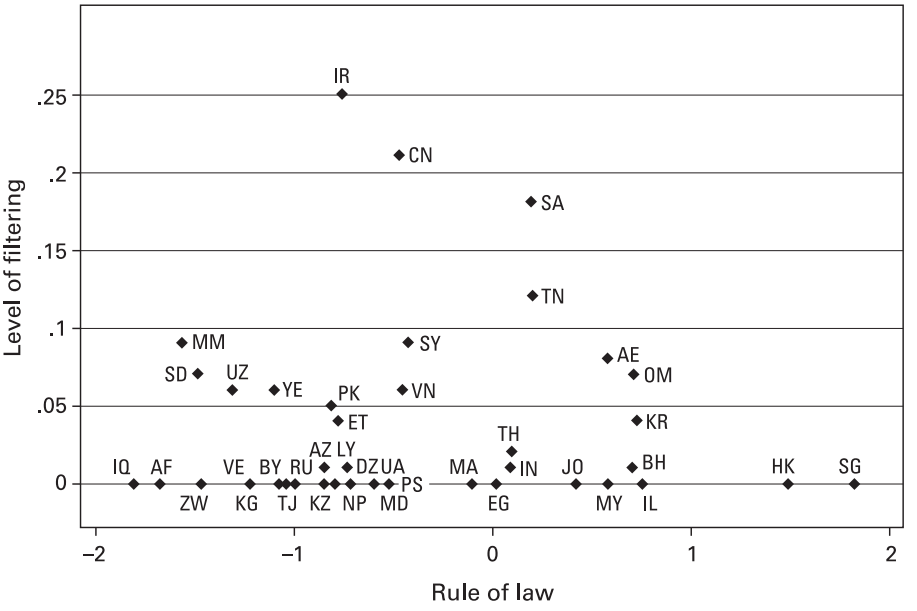


Figure 1.4

Filtering and the rule of law. AE—United Arab Emirates; AF—Afghanistan; AZ—Azerbaijan; BH—Bahrain; BY—Belarus; CN—China; DZ—Algeria; EG—Egypt; ET—Ethiopia; HK—Hong Kong; IL—Israel; IN—India; IR—Iran; IQ—Iraq; JO—Jordan; KG—Kyrgyzstan; KR—South Korea; KZ—Kazakhstan; LY—Libya; MA—Morocco; MD—Moldova; MM—Burma/Myanmar; MY—Malaysia; NP—Nepal; OM—Oman; PK—Pakistan; PS—Gaza/West Bank; RU—Russia; SA—Saudi Arabia; SD—Sudan; SG—Singapore; SY—Syria; TH—Thailand; TH—Tunisia; TN—Tunisia; TJ—Tajikistan; UA—Ukraine; UZ—Uzbekistan; VE—Venezuela; VN—Vietnam; YE—Yemen; ZW—Zimbabwe.

an uncommon circumstance. Pointing simply toward the absence of a solid rule of law does not seem promising. As seen in figure 1.4, there is no simple relationship between the rule of law and filtering, at least not as rule of law is defined and measured by the World Bank.[12] A country can maintain a better-than-average rule of law record and still filter the Internet. Similarly, many countries suffer from a substandard legal situation while maintaining an open Internet.

Comparing filtering practices with measures of voice and accountability is more telling. The countries that actively engage in the substantial filtering of political content also score poorly on measures of voice and accountability. This is true for both political and social Internet blocking, as shown in figures 1.5 and 1.6. Yet many of the anomalies persist. We are still far from explaining why some countries resort to filtering while others refrain from taking this step. This does stress the diversity of strategies and approaches that are being taken to regulate

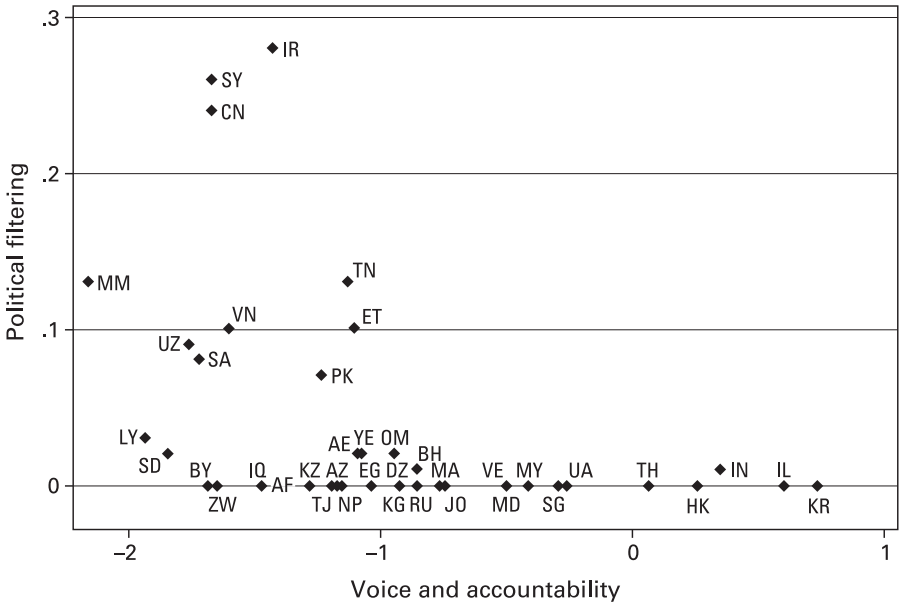


Figure 1.5
Political filtering and voice and accountability. AE—United Arab Emirates; AF—Afghanistan; AZ—Azerbaijan; BH—Bahrain; BY—Belarus; CN—China; DZ—Algeria; EG—Egypt; ET—Ethiopia; HK—Hong Kong; IL—Israel; IN—India; IR—Iran; IQ—Iraq; JO—Jordan; KG—Kyrgyzstan; KR—South Korea; KZ—Kazakhstan; LY—Libya; MA—Morocco; MD—Moldova; MM—Burma/Myanmar; MY—Malaysia; NP—Nepal; OM—Oman; PK—Pakistan; PS—Gaza/West Bank; RU—Russia; SA—Saudi Arabia; SD—Sudan; SG—Singapore; SY—Syria; TH—Thailand; TH—Tunisia; TN—Tunisia; TJ—Tajikistan; UA—Ukraine; UZ—Uzbekistan; VE—Venezuela; VN—Vietnam; YE—Yemen; ZW—Zimbabwe.

the Internet. We are also observing a recent and tremendously dynamic process. The view we have now may change dramatically in the coming years.

The link between repressive regimes and political filtering follows a clear logic. However, the link between regimes that suppress free expression and social filtering activity is less obvious. Part of the answer may reside in that regimes that tend to filter political content also filter social content.

Figure 1.7 demonstrates that few states restrict their activities to one or two types of content. Once filtering is implemented, it is applied to a broad range of content. These different types of filtering activities are often correlated with each other, and can be used as a pretense for expanding government control of cyberspace.

Vietnam, for example, uses pornography as its publicly stated reason for filtering, yet blocks little pornography. It does, however, filter political Internet content that opposes one-party rule

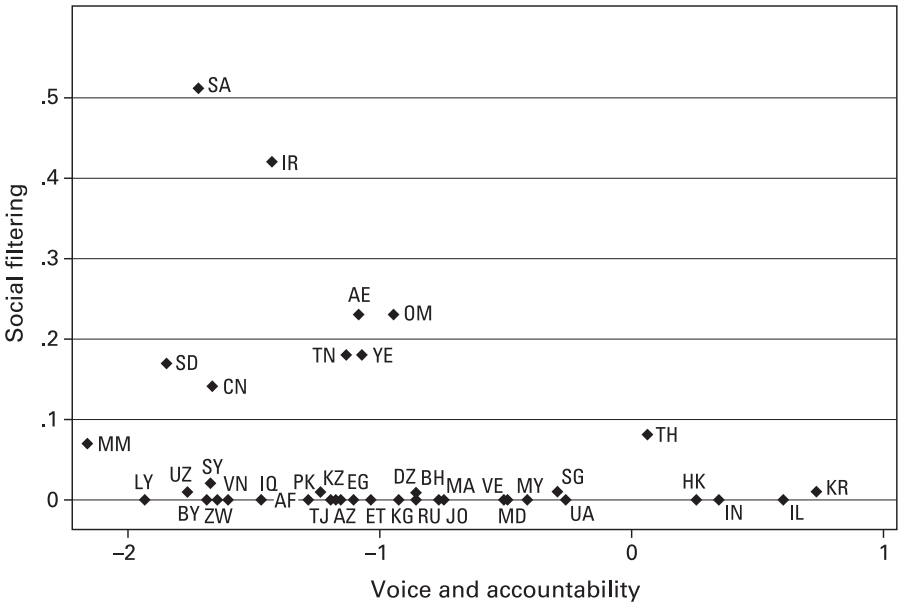


Figure 1.6
Social filtering and voice and accountability. AE—United Arab Emirates; AF—Afghanistan; AZ—Azerbaijan; BH—Bahrain; BY—Belarus; CN—China; DZ—Algeria; EG—Egypt; ET—Ethiopia; HK—Hong Kong; IL—Israel; IN—India; IR—Iran; IQ—Iraq; JO—Jordan; KG—Kyrgyzstan; KR—South Korea; KZ—Kazakhstan; LY—Libya; MA—Morocco; MD—Moldova; MM—Burma/Myanmar; MY—Malaysia; NP—Nepal; OM—Oman; PK—Pakistan; PS—Gaza/West Bank; RU—Russia; SA—Saudi Arabia; SD—Sudan; SG—Singapore; SY—Syria; TH—Thailand; TH—Tunisia; TN—Tunisia; TJ—Tajikistan; UA—Ukraine; UZ—Uzbekistan; VE—Venezuela; VN—Vietnam; YE—Yemen; ZW—Zimbabwe.

in Vietnam. In Saudi Arabia and Bahrain, filtering does not end with socially sensitive material such as pornography and gambling but expands into the political realm.

Once the technical and administrative mechanisms for blocking Internet content have been put into place, it is a trivial matter to expand the scope of Internet censorship. As discussed in subsequent chapters, the implementation of filtering is often carried by private sector actors—normally the ISPs—using software developed in the United States. Filtering decisions are thus often made by selecting categories for blocking within software applications, which may also contain categorization errors resulting in unintended blocking. The temptation and potential for mission creep is obvious. This slope is made ever more slippery by the fact that transparency and accountability are the exception in Internet filtering decisions, not the norm.

In the following chapter, Zittrain and Palfrey probe in further detail the political motives and implications of this growing global phenomenon, with subsequent chapters elaborating on technical, legal, and ethical considerations.

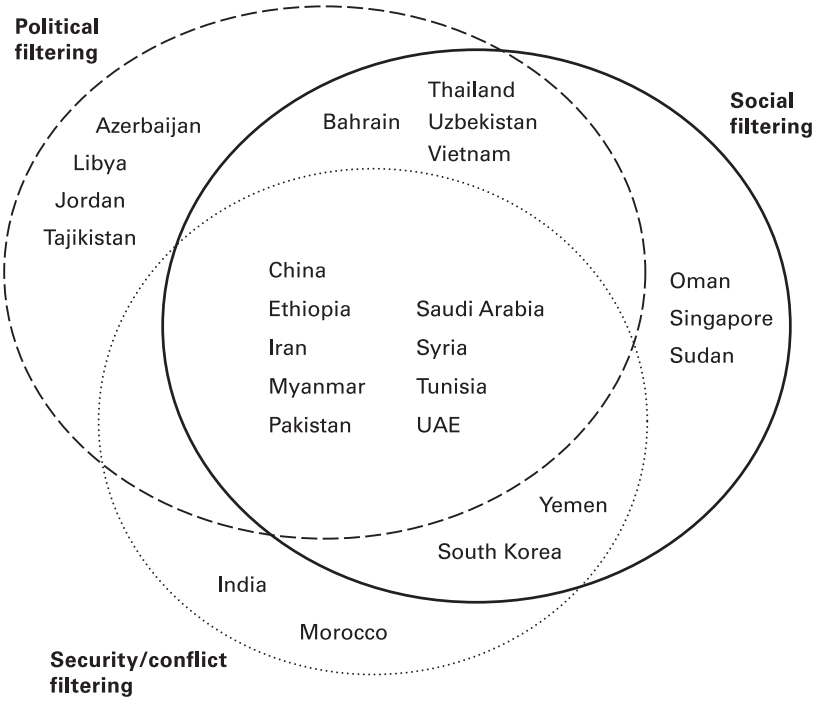


Figure 1.7
Content filtering choices.

Notes

1. The OpenNet Initiative is a collaboration of four institutions: the Citizen Lab at the University of Toronto, the Oxford Internet Institute at Oxford University, the Berkman Center for Internet & Society at Harvard Law School, and the University of Cambridge. More information is available at http://www.opennetinitiative.net.
2. A number of countries are currently debating strategies and legislation to filter the Internet, including Norway, Russia, and many countries in Latin America.
3. Each set of tests is performed on different Internet service providers within the country.
4. The Internet filtering tests carried out in Russia in 2006 were limited to ISPs accessible in Moscow. These results therefore do not necessarily reflect the situation in other areas of the country.
5. The blocking of two sites garnered most of the attention: one devoted to opposition to the September 19 coup (http://www.19sep.com/) and another hosted by Thai academics (http://www.midnightuniv.org/).
6. The strategies for addressing alleged intellectual property rights violations can vary significantly from standard Internet filtering. Rather than blocking Web sites that continue to be available from other locations, efforts generally focus on taking down the content from the Web sites that have posted the material and on removing the sites from the results of search engines. Moreover, takedown efforts are often instigated by private parties with the threat of subsequent legal action rather than being initiated by government action. See www.chillingeffects.org for more information.

7. The ONI Vietnam report is available at http://www.opennetinitiative.net/studies/vietnam/ONI_Vietnam_Country_Study.pdf.
8. We were not able to test in Cuba or North Korea. Both countries are reported to engage in pervasive filtering in addition to curtailing access to the Internet. See ''Going Online in Cuba: Internet under Surveillance,'' http://www.rsf.org/IMG/pdf/rapport_gb_md_1.pdf, and Tom Zeller, ''The Internet Black Hole That Is North Korea,'' *New York Times*, 23 October 2006.
9. There are two principal ISPs in China—one that covers the north and one the south. The smaller ISPs in China that serve Internet users connect to the Internet backbone through one of these large ISPs.
10. It also demonstrates that the use of DNS tampering does not necessitate a lack of transparency in filtering. If it were deemed important, users could be informed that the Web site they were seeking was being intentionally blocked.
11. See Shivam Vij, ''Blog Blockade Will Be Lifted in 48 Hours,'' Rediff India Abroad, http://www.rediff.com/news/2006/jul/19blogs.htm.
12. Information on the compilation and estimation of the ''rule of law'' and ''voice and accountability'' measures are available at the World Bank Governance and Anti-Corruption Web site, www.worldbank.org/wbi/governance. Their definitions of these indicators are: ''Voice and Accountability includes in it a number of indicators measuring various aspects of the political process, civil liberties, political and human rights, measuring the extent to which citizens of a country are able to participate in the selection of governments.'' ''Rule of Law includes several indicators which measure the extent to which agents have confidence in and abide by the rules of society. These include perceptions of the incidence of crime, the effectiveness and predictability of the judiciary, and the enforceability of contracts.''

R. Rogers (2009). "The Internet treats censorship as a malfunction and routes around it? A new media approach to the study of state Internet censorship," in J. Parikka and T. Sampson (eds.), *The spam book: On viruses, porn, and other anomalies from the dark side of digital culture*. Cresskill, NJ: Hampton Press, 229-247.

12

THE INTERNET TREATS CENSORSHIP AS A MALFUNCTION AND ROUTES AROUND IT?

A New Media Approach to the Study of State Internet Censorship

Richard Rogers

THE WEB AS A SET OF DISCRETE SITES?

The research approach described here is a contribution to the study of state Internet censorship. It seeks to move beyond the dominant treatment of the Web as a set of discrete sites, which are blocked or accessible. Here the Web is considered to be an information-circulation space. In a sense, a conceptualization of the Web as circulation space as opposed to a set of discrete sites is more of a new media than old media starting point.

In an old media way of thinking, there are, say, single books that are censored, just as there are now single sites. There may be types of books, or types of sites, that are censored (e.g., dating, religious conversion, or human rights). But if censorship research work is considered from a new media perspective, the methods, techniques, as well as the research output may change.

On the Internet, part of a single site may have circulated, and that content may be available elsewhere. The information on sites that are censored may be syndicated, and fed by RSS, or it may have been scraped, in an automated or semi-automated form of copying and pasting. Additionally, snippets of censored content may also have been grabbed, and subsequently annotated, commented on, or similar, for example in the blogosphere. That

"new media apparatus" may be available. Finally, there may be "related sites" and "related content"—related because they are in surfers' topical paths. (Alexa provides such "related sites.") Thus, single sites may be censored but portions of the same or related content, and its apparatus, may be unblocked.

Revealing the unblocked content shifts the focus of the work from the analysis of single sites to that of information circulation. It also shifts the research away from the policies of the censor to the Web knowledge and skills of the censored site owner. For example, site owners cognizant of censorship have been known to change their domain names repeatedly, striving to keep a step ahead of filtering software and censor's blacklists. The day-to-day competition between the censor and the censored is not so unlike that between search engine companies and search engine optimizers. The optimizer, like the censored, is striving to find out whether the new sleight of hand that keeps the information in the right space has been discovered.

Demonstrating the techniques of circulatory forms of censorship circumvention has implications for both censors as well as the censored. For example, the filtering software companies subscribe to proxy list providers' notifications. Proxies are machines serving as gateways, and are used by surfers in censored countries (among others) to have a different geographical (Internet provider [IP]) point of entry to the net. (They also are used by censorship researchers to check sites in countries known to censor the Internet. One connects to the Internet in Iran (through an Iranian proxy), and fetches sites in order to see the connection statistics, and/or to capture screen grabs of blocked sites. Censors and filtering software companies also make use of proxy lists, adding them to their blacklists. Just as filtering companies may subscribe to alerts from proxy list providers, censors could pull in site feeds, query them in engines, and refresh the blacklist according to the engine returns.

URL LISTS AND INTERNET CENSORSHIP RESEARCH

One of the more comprehensive (and open source) blacklists of sites is coupled with the Dans Guardian filtering software, listing some 56 categories of sites blocked (at urlblacklist.com) from "kids time-wasting" to "weapons" (see Table 12.1). There is also the ability to register both suggestions for blocking as well as complaints about blocked sites. A well-known filtering application in the proprietary arena, SmartFilter by the Secure Computing Corporation, advertises 73 categories of blocked sites (see Table 12.2). In the past filtering companies' lists have been cracked, and circulated, leading to great consternation about the editorial skills and orientations of the list-makers.

TABLE 12.1
URL Black List Categories and Descriptions for the Dans Guardian Open Source Filtering Software, 23 March 2007

CATEGORY	DESCRIPTION
Ads	Advert servers and banned URLs
Adult	Sites containing adult material such as swearing but not pornography
Aggressive	Similar to violence but more promoting than depicting
Anti-spyware	Sites that remove spyware
Artnudes	Art sites containing artistic nudity
Banking	Banking Web sites
Beer liquor info	Sites with information only on beer or liquors
Beer liquor sale	Sites with beer or liquors for sale
Cell phones	stuff for mobile/cell phones
Chat	Sites with chat rooms, etc.
Child care	Sites to do with child care
Clothing	Sites about and selling clothing
Culnary	Sites about cooking et al.
Dating	Sites about dating
Dialers	Sites with dialers such as those for pornography or trojans
Drugs	Drug-related sites
E-commerce	Sites that provide online shopping
Entertainment	Sites that promote movies, books, magazine, humor
French education	Sites to do with French education
Gambling	Gambling sites, including stocks and shares
Gardening	Gardening sites
Government	Military and schools, etc.
Hacking	Hacking/cracking information
Home repair	Sites about home repair
Hygiene	Sites about hygiene and other personal grooming-related information
Instant messaging	Sites that contain messenger client download and Web-based messaging sites
Jewelry	Sites about and for selling jewelry
Job search	Sites for finding jobs
Kids time wasting	Sites kids often waste time on
Mail	Web mail and e-mail sites
Naturism	Sites that contain nude pictures and/or promote a nude lifestyle

TABLE 12.1
URL Black List Categories and Descriptions for the Dans Guardian Open Source Filtering Software, 23 March 2007 (*continued*)

CATEGORY	DESCRIPTION
News	News sites
Online auctions	Online auctions
Online games	Online gaming sites
Online payment	Online payment sites
Personal finance	Personal finance sites
Pets	Pet sites
Phishing	Sites attempting to trick people into giving out private information
Porn	Pornography
Proxy	Sites with proxies to bypass filters
Radio	Non–news-related radio and television
Religion	Sites promoting religion
Ring tones	Sites containing ring tones, games, picture, etc.
Search engines	Search engines such as Google
Sexuality	Sites dedicated to sexuality, possibly including adult material
Sports news	Sports news sites
Sports	All sports sites
Spyware	Sites that run or have spyware software to download
Update sites	Sites where software updates are downloaded from, including virus sigs
Vacation	Sites about going on vacation
Violence	Sites containing violence
Virus infected	Sites that host virus-infected files
Warez	Sites with illegal pirate software
Weather	Weather news sites and weather-related
Weapons	Sites detailing with or selling weapons
Web mail	Just Web mail sites
White list	Contains site specifically 100% suitable for kids

Source: urlblacklist.com

TABLE 12.2
SmartFilter (Rich feature-set, March 23, 2007

FILTERING OPTIONS
73 individual categories of Web sites
Both URL and IP addresses
http and https traffic
File type (jpg, MP3, etc.)
Granular key word searches/search engine key word blocking
Time of day
Day of week
Default filtering policies available
FILTERING ACTIONS
Group users or workstations under a common policy
Deny, allow, warn, but allow, exempt, delay, or report only
Authorized override—authorized users can bypass the filter for a specified amount of time
Global block/allow
FILTERING CUSTOMIZATION
500 user-defined categories
Create unique filtering response message for end users
Add, delete, or exempt sites from categories
Pattern matching: build dynamic rules for granular custom filtering

Leading researchers of Internet censorship have had a similar point of departure. Until recently, the work has been devoted to building a global list of URLs, with some 37 categories in all. Once the lists are in place, the censorship researchers fetch the URLs through a browser in each of the countries under study (see Tables 12.3 and 12.4). As an initial check, proxy servers located in countries that censor the Internet may be used. If the http return codes are 403 (forbidden) or 504 (server gateway time out), the sites are tagged as suspected blocks. (Other http return codes may provide indications of censorship.) Researchers on the ground subsequently check each URL (suspected or otherwise). Lists are made of blocked sites, per category, across the set of countries under study. Country levels of censorship by site category (with specific lists of blocked URLs) constitute a main research

output. State censorship policy is described, as are the censorship techniques (e.g., gateway time outs in China), including the identification of particular software packages in use (e.g., SmartFilter in Saudi Arabia).

So far the main thrusts of Internet censorship research have been described, also in the context of filtering software more generally—list creation, URL fetching, and http return code monitoring. Now, I describe the means by which one may contribute to the creation of URL lists, and gradually fill in the notion of new media Internet censorship research, with its emphasis on the Web as a circulation space. In particular, I describe three Internet censorship research techniques: related site dynamic URL sampling (URL list-making with hyperlink analysis), redistributed content discovery (through key word searching, key phrase parsing, and additional searching), and surfer re-routing (through route map-making).

TABLE 12.3
Open Net Initiative's Categories in the Global URL List for State Internet Censorship Research

Alcohol	Humor
Anonymizers	Major events
Blogging domains	Medical
Drugs	Miscellaneous
Dating	News outlets
E-mail	Person-to-person
Encryption	Porn
Entertainment	Provocative attire
Environment	Religion (fanatical)
Famous bloggers	Religion (normal)
Filtering sites	Religious conversion
Free webspace	Search engines
Gambling	Sexual education
Gay/lesbian/bisexual/transgender/ queer issues	Translation sites
Government	Terrorism
Hacking	Universities
Hate speech	Weapons/violence
Human rights	Women's rights
	Voice over Internet protocol

TABLE 12.4
Open Net Initiative's Country List for State Internet Censorship Research

ASIA AND SOUTH ASIA	LATIN AMERICA
Burma	Cuba
China, Hong Kong	Venezuela
India	
Malaysia	MIDDLE EAST AND AFRICA
Maldives	Afghanistan
Nepal	Algeria
North Korea *	Bahrain
Pakistan	Egypt
Singapore	Eritrea
South Korea	Ethiopia
Thailand	Iran
Vietnam	Iraq
	Israel
EASTERN AND CENTRAL ASIA	Jordan
Azerbaijan	Libya
Belarus	Morocco
Kazakhstan	Oman
Kyrgyzstan	Saudi Arabia
Moldova	Sudan
Russia	Syria
Tajikistan	Tunisia
Turkmenistan	United Arab Emirates
Ukraine	Yemen
Uzbekistan	Zimbabwe

Related Site Dynamic URL Sampling

The current method in Internet censorship research for compiling the global list of URLs is editorial. For an initial URL trawl, directories may be used, such as Yahoo's, Google's or Dmoz.org's. Subsequently, country experts are consulted, and URLs of interest only for one or more particular country are collected. These are the so-called high-impact sites, such as opposition par-

ties. Generally speaking, between 1,000 and 2,000 URLs are checked per country. However, Julian Pain, head of the Internet Freedom desk at Reporters Without Borders, has indicated that the quantity of sites censored in particular countries may be much greater. In Saudi Arabia, "the official Internet Service Unit (ISU) is proud to tell you it's barred access to nearly 400,000 sites and has even posted a form online for users to suggest new websites that could be blocked." [625]

In its own form of a new media style (user-generated content), Saudi Arabia, like urlblacklist.com, "crowd-sources" URLs to bring to the attention of the ISU, using the many-eyes approach over the assumingly few eyes of the censors. If there are 400,000 sites being censored, however they are all sourced, and the Internet censorship researchers are checking only some 2,000, questions arise. How should URLs be sourced? How should the list be made more sizeable? An important consideration concerns the people on the ground in each of the countries who fetch the URLs on the lists through browsers. The time it takes to run the lists may be considerable; care also needs to be taken for personal security reasons. Thus the additional URLs put on the list to be checked should be vetted for relevance.

In a post-directory era, where in Google the directory is no longer a main tab (and three clicks away) and in Yahoo no longer the default search engine, relevance follows from counting links, and boosting sites either through freshness (in a pagerank style) or through votes (in a user ratings style). Here, initially, the link-counting strategy is employed, where a set of sites point to other sites to which they collectively link. Using the URL and site-type data furnished by the Internet censorship researchers, I crawled one category of sites in one country—the "political, social and religious" sites on the Iranian list. The sites' hyperlinks (external links) are harvested, and co-link analysis is performed, where those sites with two links from the initial list of sites are retained. Once the network of interlinked sites is found, all the sites are cross-checked with the Internet censorship researchers' lists of known blocked sites, ascertaining which sites are already known blocks. All newly discovered sites are fetched through proxies in Iran, in order to ascertain their status. The result is a map showing political, religious, and social sites blocked and unblocked in Iran, with pins indicating newly dis - covered blocks (see Fig. 12.1). Of particular interest is the case of the British Broadcasting Service (BBC). The Internet censorship researchers had the BBC news homepage on its list of sites to check (http://news.bbc.co.uk). The link analysis turned up a deep page on the site, the BBC's Persian language page (http://www.bbc.co.uk/persian). In Iran, the BBC news page is accessible, as the researchers had found, but the Persian-language page is not. In all some 37 censored sites were newly discovered through what we termed a dynamic URL sampling method , which relied on an analysis of hyperlinking for related site relevance as opposed to the editorial process—directories and experts.

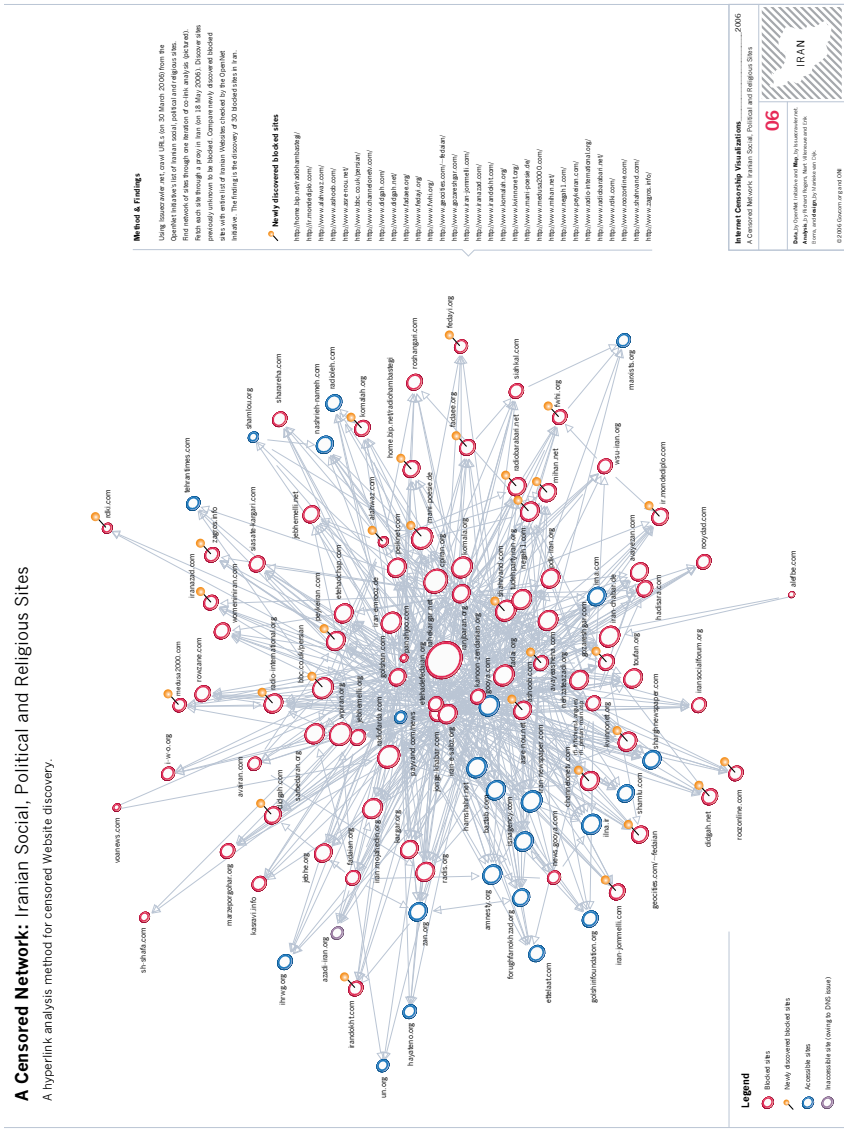

A Censored Network: Iranian Social, Political and Religious Sites
A hyperlink analysis method for censored Website discovery.
Method & Findings
Newly discovered blocked sites
Internet Censorship Visualizations
A Censored Network: Iranian Social, Political and Religious Sites
2006
06
IRAN
Legend
Blocked sites
Newly discovered blocked sites
Accessible sites
Inaccessible site (owing to DNS issue)

FIGURE 12.1.

REDISTRIBUTED CONTENT DISCOVERY

Research into state censorship in Pakistan has found, among other things, that two groups seeking autonomy (the Balochi and the Sindhi) have their sites routinely blocked. The Internet censorship researchers have lists of blocked sites for the two groups, one of which (the Balochi) served as starting points for the URL discovery method just described—the crawling of sites, the link analysis, and the proxy checking. With two newly discovered censored sites added to the list through a hyperlink analysis, the overall question concerns the extent to which the blocked content has been redistributed to sites that are not blocked in Pakistan. The case study concerns the killing by the Pakistan military of the Baloch tribal leader, Nawab Akbar Khan Bugti. A special Google query for "Nawab Akbar Khan Bugti," which excludes known blocked sites in Pakistan, shows some 900 results. (Google only serves up to 1,000 results per query.) The teaser texts of the returns are analyzed for unique phrases, and sorted by date (see Fig. 12.2). When listed chronologically, from June to October 2006, the parsed phrases appearing before and after Nawab Akbar Khan Bugti tell a story.

The following "story" describes of the death of "Nawab Akbar Khan Bugti," the Baloch tribal leader, from parsed Google (teaser text) returns, June 26 to October 12, 2006. Baloch-authored content, not blocked by Pakistan Internet censorship, appears in italics.

> He' s 80 years old, but Nawab Akbar Khan Bugti, a feudal lord in Pakistan' s rugged Baluchistan province, wants to fight to the death.
>
> The irony was that Nawab Akbar Khan Bugti served to help the federal government when he was appointed as Governor of Balochistan by Mr. Zulfikar Ali Bhutto
>
> "Nawab Akbar Khan Bugti was directly attacked. Luckily he survived all attacks and is safe," said Khan, rejecting rumours that Akbar Bugti's grandson
>
> have claimed to have killed Nawab Akbar Khan Bugti, one of the founding fathers of the Baloch independence struggle, and 36 other freedom-fighters
>
> The martyrdom of Nawab Akbar Khan Bugti is a loss for Pakistan and a gain for Baloch nationalist movement
>
> It was the third attempt on the life of Nawab Akbar Khan Bugti. After the interception of satellite phone communication, the Nawab' s location was pin

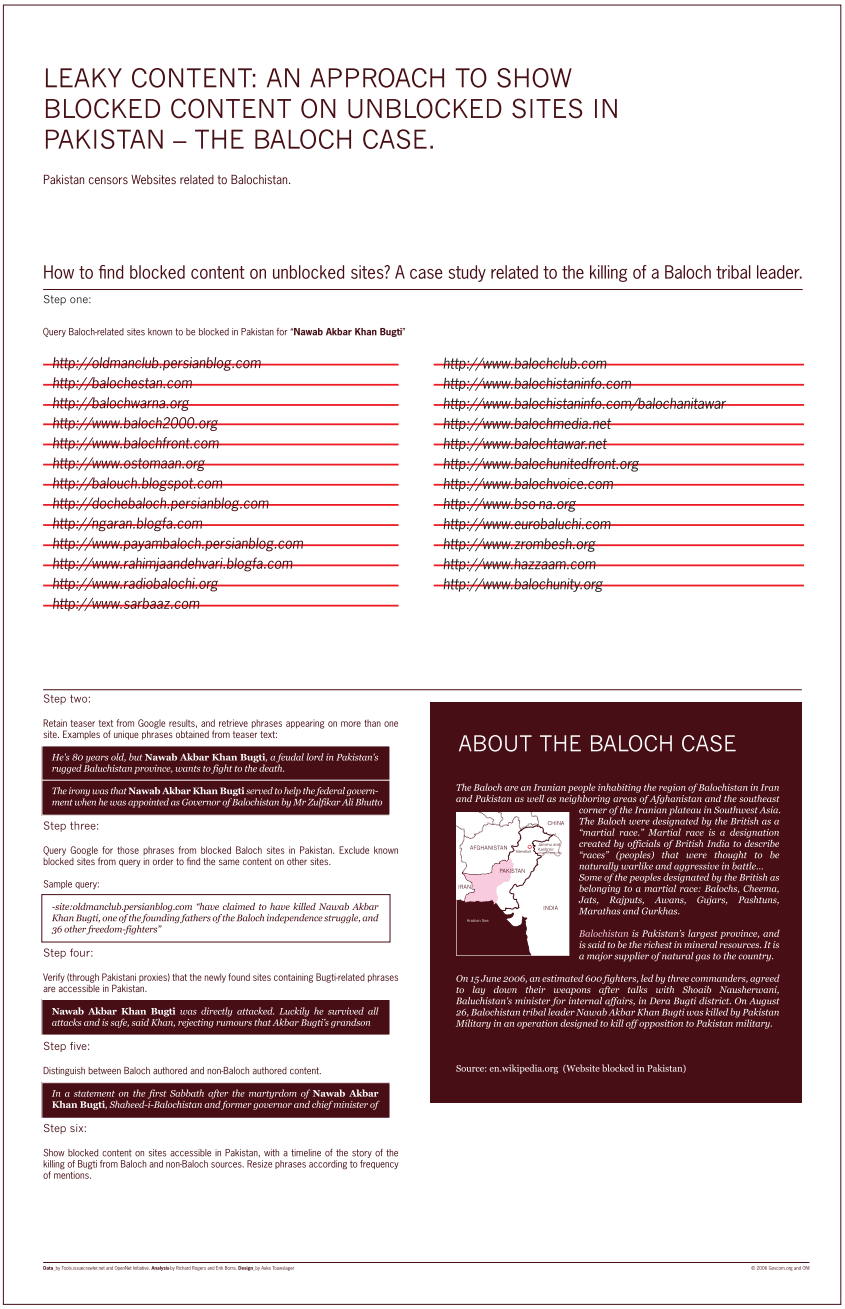

LEAKY CONTENT: AN APPROACH TO SHOW BLOCKED CONTENT ON UNBLOCKED SITES IN PAKISTAN – THE BALOCH CASE.
Pakistan censors Websites related to Balochistan.
How to find blocked content on unblocked sites? A case study related to the killing of a Baloch tribal leader.
Step one:
Query Baloch-related sites known to be blocked in Pakistan for "Nawab Akbar Khan Bugti"
http://oldmanclub.persianblog.com
http://balochestan.com
http://balochwarna.org
http://www.baloch2000.org
http://www.balochfront.com
http://www.ostomaan.org
http://balouch.blogspot.com
http://dochebaloch.persianblog.com
http://ngaran.blogfa.com
http://www.payambaloch.persianblog.com
http://www.rahimjaandehvari.blogfa.com
http://www.radiobalochi.org
http://www.sarbaaz.com
http://www.balochclub.com
http://www.balochistaninfo.com
http://www.balochistaninfo.com/balochanitawar
http://www.balochmedia.net
http://www.balochtawar.net
http://www.balochunitedfront.org
http://www.balochvoice.com
http://www.bso-na.org
http://www.eurobaluchi.com
http://www.zrombesh.org
http://www.hazzaam.com
http://www.balochunity.org
Step two:
Retain teaser text from Google results, and retrieve phrases appearing on more than one site. Examples of unique phrases obtained from teaser text:
He's 80 years old, but Nawab Akbar Khan Bugti, a feudal lord in Pakistan's rugged Baluchistan province, wants to fight to the death.
The irony was that Nawab Akbar Khan Bugti served to help the federal government when he was appointed as Governor of Balochistan by Mr Zulfikar Ali Bhutto
Step three:
Query Google for those phrases from blocked Baloch sites in Pakistan. Exclude known blocked sites from query in order to find the same content on other sites.
Sample query:
-site:oldmanclub.persianblog.com "have claimed to have killed Nawab Akbar Khan Bugti, one of the founding fathers of the Baloch independence struggle, and 36 other freedom-fighters"
Step four:
Verify (through Pakistani proxies) that the newly found sites containing Bugti-related phrases are accessible in Pakistan.
Nawab Akbar Khan Bugti was directly attacked. Luckily he survived all attacks and is safe, said Khan, rejecting rumours that Akbar Bugti's grandson
Step five:
Distinguish between Baloch authored and non-Baloch authored content.
In a statement on the first Sabbath after the martyrdom of Nawab Akbar Khan Bugti, Shaheed-i-Balochistan and former governor and chief minister of
Step six:
Show blocked content on sites accessible in Pakistan, with a timeline of the story of the killing of Bugti from Baloch and non-Baloch sources. Resize phrases according to frequency of mentions.
ABOUT THE BALOCH CASE
The Baloch are an Iranian people inhabiting the region of Balochistan in Iran and Pakistan as well as neighboring areas of Afghanistan and the southeast corner of the Iranian plateau in Southwest Asia. The Baloch were designated by the British as a "martial race." Martial race is a designation created by officials of British India to describe "races" (peoples) that were thought to be naturally warlike and aggressive in battle. . . Some of the peoples designated by the British as belonging to a martial race: Balochs, Cheema, Jats, Rajputs, Awans, Gujars, Pashtuns, Marathas and Gurkhas.
CHINA
AFGHANISTAN
PAKISTAN
IRAN
INDIA
Balochistan is Pakistan's largest province, and is said to be the richest in mineral resources. It is a major supplier of natural gas to the country.
On 15 June 2006, an estimated 600 fighters, led by three commanders, agreed to lay down their weapons after talks with Shoaib Nausherwani, Baluchistan's minister for internal affairs, in Dera Bugti district. On August 26, Balochistan tribal leader Nawab Akbar Khan Bugti was killed by Pakistan Military in an operation designed to kill off opposition to Pakistan military.
Source: en.wikipedia.org (Website blocked in Pakistan)

FIGURE 12.2.

Nawabzada Hyrbair Marri on Monday rejected government's claims that Nawab Akbar Khan Bugti had died because of the collapse of his cave hideout

Nawab Akbar Khan Bugti buried in Balochistan without the presence of his family

Nawab Akbar Khan Bugti in a military operation, prominent Baloch leaders and Pakistani human rights activists said it spelt doom for the country's unity and

In a statement on the first Sabbath after the martyrdom of Nawab Akbar Khan Bugti, Shaheed-i-Balochistan and former governor and chief minister of

Baloch Nationalist leader Nawab Akbar Khan Bugti, who was murdered by Pakistani military

Balochistan, Nawab Akbar Khan Bugti, the highest elected official to be killed by the Pakistan Army. Since March 27, 1948 when Balochistan was forcibly

The killing of Baloch leader Nawab Akbar Khan Bugti in August 2006 sparked riots and will likely lead to more confrontation. The conflict could escalate if

In fact, Nawab Akbar Khan Bugti has lighted the candle

But I wonder why journalists, brought in on a military helicopter to witness Nawab Akbar Khan Bugti being buried by a dozen common labourers, couldn't ask

blooded murder of their great leader Nawab Akbar Khan Bugti

Nawab Akbar Khan Bugti had played a significant and controversial role in Pakistani

Speaking at a condolence reference for the late Nawab Akbar Khan Bugti at the Hyderabad press club under the aegis of Sindh National Party (SNP),

tensions have increased since the killing of a veteran nationalist politician, Nawab Akbar Khan Bugti, in a military offensive in August.

The status of Nawab Akbar Khan Bugti, the octogenarian chieftain of a tribe in the restive southwestern province of Balochistan, almost reached the mythical

Note there is Baloch-authored (in italics) and non–Baloch-authored content. The research questions relate to the amount of Baloch-authored content accessible in Pakistan, as well as the level of distinctiveness of the story of his death from Baloch-related sites vis-à-vis non-Baloch. Where the first question is concerned, it is remarkable, in some sense, how "well" Pakistan appears to be blocking Baloch-authored content, for so little is redistributed. Phrased differently, the content circulation is relatively low. In the depiction of the "leaky content," where the Baloch and non-Baloch content are resized according to frequency of returns, the Baloch-authored story size is small (see Fig. 12.3). Among the scant number of sites carrying a Baloch-authored story, often with redistributed content from blocked sites in Pakistan, are Gedrosia.blogspot.com, Intellibriefs.blogspot.com, Ezboard.com, Thechosenpeople.blogspot.com, Thebalochpeople.org, Dc.indymedia.org, and Baltimore.indymedia.org—blogs, forums, and indymedia sites. (Later, the account used on Ezboard.com by "hinduunity" was "locked down" under the site's terms of use, after a threatened lawsuit, recounted on intellibriefs.blogspot.com. [626] Hinduunity.org now has its forum hosted on its own site.) To take up the second question, the difference in the Baloch and non-Baloch versions of the story of the death of Nawab Akbar Kan Bugti is stark for the Baloch reference to murder as opposed to killing.

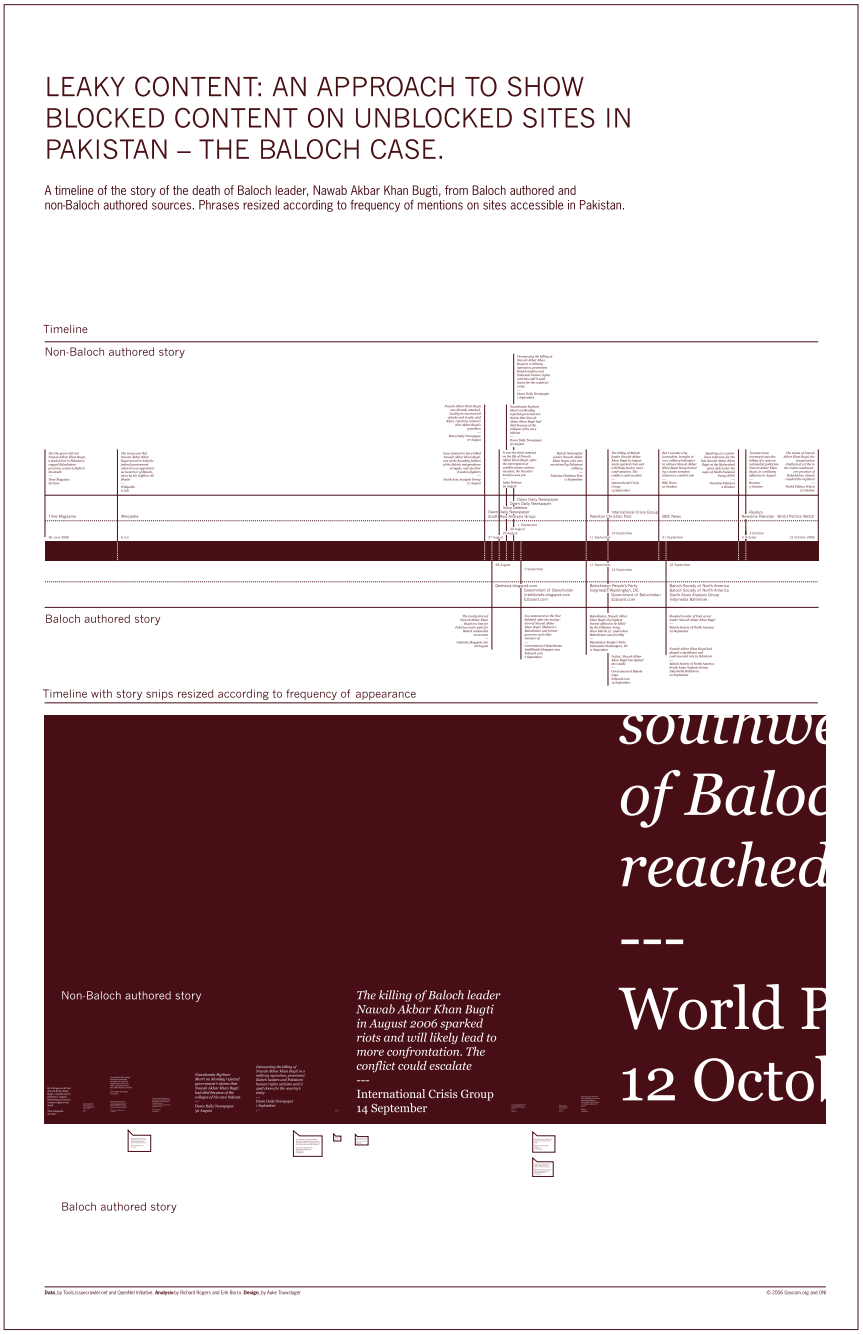

LEAKY CONTENT: AN APPROACH TO SHOW BLOCKED CONTENT ON UNBLOCKED SITES IN PAKISTAN – THE BALOCH CASE.
A timeline of the story of the death of Baloch leader, Nawab Akbar Khan Bugti, from Baloch authored and non-Baloch authored sources. Phrases resized according to frequency of mentions on sites accessible in Pakistan.
Timeline
Non-Baloch authored story
Baloch authored story
Timeline with story snips resized according to frequency of appearance
Non-Baloch authored story
The killing of Baloch leader Nawab Akbar Khan Bugti in August 2006 sparked riots and will likely lead to more confrontation. The conflict could escalate
International Crisis Group
14 September
Baloch authored story

FIGURE 12.3.

Surfer Re-Routing

The famous quotation about how the Internet treats censorship—a version of which is the title of this chapter—is attributed to John Gilmore, co-founder of the Electronic Frontier Foundation. In the notes to his 1998 paper, "Why the Internet is Good," Internet law scholar, Joseph Reagle, has the following annotations for the original quotation (in bold):

> "The Net interprets censorship as damage and routes around it." John Gilmore (EFF). [source: Gilmore states: "I have never found where I first said this. But everyone believes it was me, as do I. If you find an appearance of this quote from before March '94, please let me know." Also in NYT 1/15/96, quoted in CACM 39(7):13. Later, Russell Nelson comments (and is confirmed by Gilmore) that on December 05 1993 Nelson sent Gilmore an email stating, "Great quote of you in Time magazine: 'The net treats censorship as a defect and routes around it.'"] [627]

The technical thought behind the quotation refers to packet switching, as another legal scholar, James Boyle wrote in 1997:

> The distributed architecture and its technique of packet switching were built around the problem of getting messages delivered despite blockages, holes and malfunctions. Imagine the poor censor faced with such a system. There is no central exchange to seize and hold; messages actively 'seek out' alternative routes so that even if one path is blocked another may open up. Here is the civil libertarian's dream. [628]

There are now technical means to route around censorship, such as the circumventor by peacefire.org, a proxy service. Lists of proxy servers are updated frequently, in the ongoing race to stay a day or two ahead of the updates furnished by the content filtering software companies to their clients. Peacefire.org claims that filtering companies are routinely three to four days behind in updating their blacklists of proxies, so peacefire's fresh proxy lists are useable on any given day. The intensive censorship and anti-censorship work behind the scenes is telling for how the discourse has changed for "route arounds." Rather than being built into the infrastructure of the Internet, routing around should be described as labor-intensive and semi-manual work—proxy detection, list updating, alert sending. Thus, the discourse of routing around censorship is changing from the reverence of the Internet architecture and the far-sighted architects of the end-to-end principle to governance as well as to artful technique.

In an effort to show the routes, not for packets, but for content surfers, the Internet researchers' global list of women's rights sites was employed to make a surfer route map, as it may be called initially. As in the URL discovery method described earlier, the sites' outlinks were captured, and a network graph generated, showing the clusters of women's rights sites disclosed by inter-linking. The route map has sites and paths annotated in red and green, with red indicating known blockage (see Fig. 12.4).

The map plays on reworked ideas from hypertext theory whereby a surfer "authors" a path through Web space, one that eventually may be retrieved in the browser history. It also harkens to the art of surfing as opposed to mere searching. If one were to think of a surfer in China moving through a women's rights space (largely in English owing to the URLs on the Internet censorship researchers' global list), and authoring some sense of a story, first, from the seed list, hrw.org/women, ifeminists.com and womenofarabia.com (now offline) would not figure among the sources, for they are blocked.[629]

Which issues and stories about women's rights in China are discussed on hrw.org/women and ifeminists.com? Is there a path to similar or related content on unblocked sites? Ifeminists.com have 10 entries on China: 3 of the 10 deal with the one-child policy, and the disproportion of boys born. Another follows from the "shortage of women," and reports the trafficking of North Korean women, sold to Chinese "husbands." A syphillus epidemic is discussed in two further stories, and the others deal with sexual harrassment, online porn, easier divorces and AIDS, respectively. In discussing South Asia, China, and South Korea, Human Rights Watch, whose entire site is censored in China, writes about preferences for boys, "sex-selective abortions" as well female infanticide.[630]

In order to find surfer content routes, the actor sites on the women's rights map are queried initially for China-related topics discussed by Ifeminists and Human Rights W atch: "one-child policy" China, syphilis China, "shortage of women" China, AIDS China, "online porn" China, divorce China, and "sexual harassment" China. (Queries are made in English, for there is less censorship for English-language terms than for Mandarin.) Of the 88 nodes in the women's rights network, approximately one-third of unblocked sites return content on those key-word issues. The map organizes a women's rights-related content space, and through the choice of a map, as opposed to a list or a site tag cloud, suggests pathways. Because China has search engines delist sites and also performs key-word blocking, it is also important to cross-check known blocked words. A search through the known key words blocked in China did not turn up any of the above words.[631]

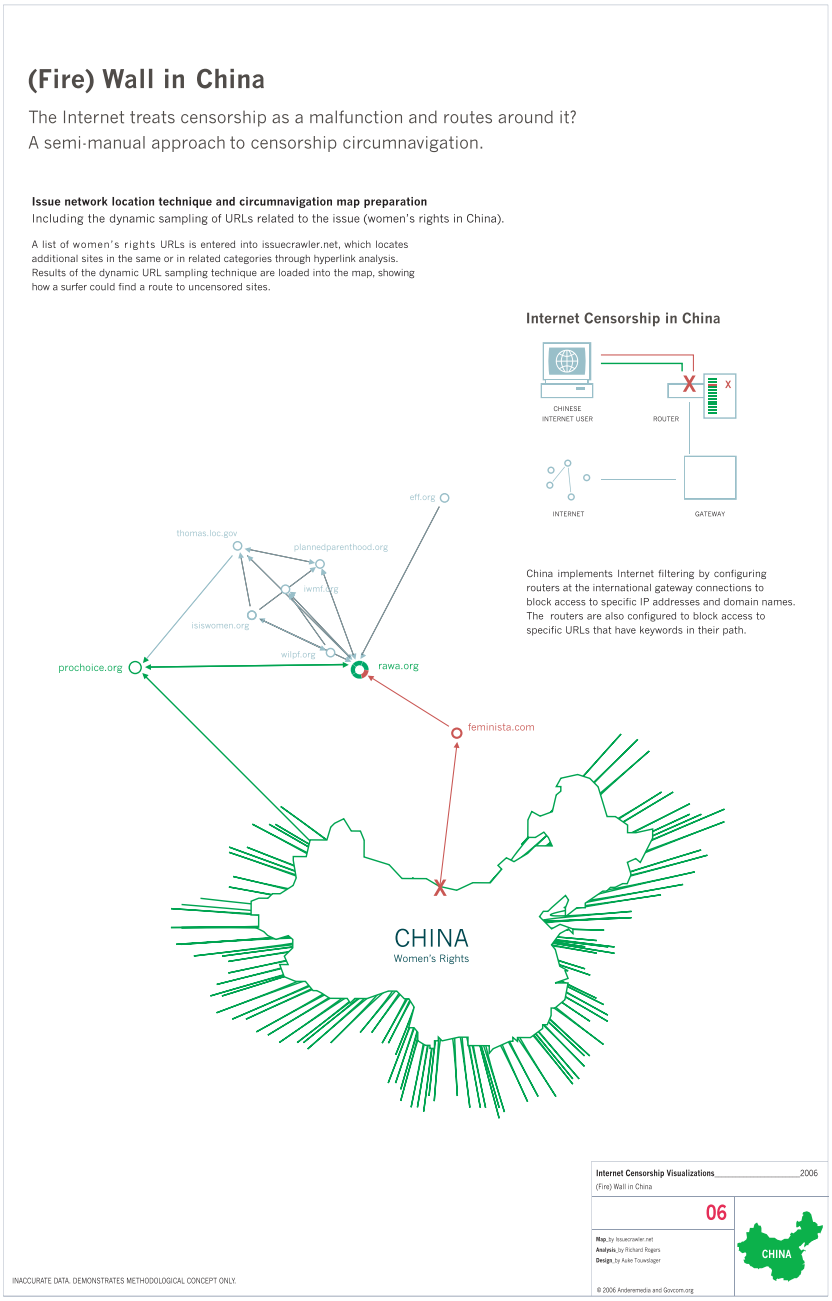

(Fire) Wall in China
The Internet treats censorship as a malfunction and routes around it?
A semi-manual approach to censorship circumnavigation.
Issue network location technique and circumnavigation map preparation
Including the dynamic sampling of URLs related to the issue (women's rights in China).
A list of women's rights URLs is entered into issuecrawler.net, which locates additional sites in the same or in related categories through hyperlink analysis. Results of the dynamic URL sampling technique are loaded into the map, showing how a surfer could find a route to uncensored sites.
Internet Censorship in China
CHINESE INTERNET USER
ROUTER
INTERNET
GATEWAY
China implements Internet filtering by configuring routers at the international gateway connections to block access to specific IP addresses and domain names. The routers are also configured to block access to specific URLs that have keywords in their path.
eff.org
thomas.loc.gov
plannedparenthood.org
iwmf.org
isiswomen.org
wilpf.org
prochoice.org
rawa.org
feminista.com
CHINA
Women's Rights
INACCURATE DATA. DEMONSTRATES METHODOLOGICAL CONCEPT ONLY.
Internet Censorship Visualizations 2006
(Fire) Wall in China
06
Map, by Issuecrawler.net
Analysis, by Richard Rogers
Design, by Auke Touwslager
© 2006 Anderemedia and Govcom.org
CHINA

FIGURE 12.4.

CONCLUSION

State Internet Censorship is evolving from the directory-editor model, described earlier as old media-style for it assumes a Web constituted of institutions or actors operating single sites. The new media style, conversely, follows the movement of content around the Web ("circulation space"), and concentrates less on the policies of the censor than on the skills of the content movers, and how the results of those skills may be captured.

The new media style of Internet censorship research concentrates on describing both specific skills of the content movers as well as the techniques to measure the extent of the content movement. Importantly, the idea that the Web 2.0-style of content redistribution (scraping, feeding) is the new infrastructure of the Internet for routing around censorship appears to be in its infancy, however. The Baloch-authored story of the death of the tribal leader, it was found, is underredistributed on sites accessible in Pakistan. Although present and available (in English) information about the consequences of female infanticide in China (e.g., the "shortage of women" and the trafficking of North Korean "wives" to Chinese "husbands") should not be considered abundant.

To date, digital journalism studies have focused on such subjects as newspapers going online and whether gatekeeping will be lessened owing to "interactivity." Also treated are the relationship between blogs and mainstream news (who's following whom) as well as the challenges of the amateur, where the Saddam Hussein hanging video appears to have greater claims to veracity owing to its mobile-phone graininess than news accounts of it filmed in a studio with anchorpersons. There is less emphasis on how information may become separated from its sources, and the consequences of the untethering for the distribution of attention.

When researchers and others consider the Web as circulation space, often there are particular connotations—Web as rumor mill or blogosphere as echo chamber, for example. Working with these assumptions, the "good journalist" would then be asked to trace the story back to a source. Source tracing, whether thought of in an archeological or genealogical sense, becomes the techno-epistemological practice, with an emphasis on source page date stamps. Here the practice is just as technical, however much the commitment changes to the expanse of the spread or "sharing," as it's sometimes called in participatory culture studies.

ACKNOWLEDGMENT

I acknowledge the support and assistance of the OpenNet Initiative, the project by the Citizen Lab at the University of Toronto, the Berkman

Center at Harvard University and the Advanced Research Group, Security Studies, University of Cambridge. The figures are by the Govcom.org Foundation, Amsterdam, with special appreciation extended to Erik Borra, Marieke van Dijk and Auke Touwslager. Laura van der Vlies provided valuable assistance.

R. Deibert and R. Rohozinski (2010). "Control and Subversion in Russian Cyberspace," *Access Controlled*. Cambridge, MA: MIT Press, 15-34.

2 Control and Subversion in Russian Cyberspace

Ronald Deibert and Rafal Rohozinski

Introduction

It has become a truism to link censorship in cyberspace to the practices of authoritarian regimes. Around the world, the most repressive governments—China, Burma, North Korea, Cuba, Saudi Arabia—are the ones that erect digital firewalls that restrict citizens' access to information, filter political content, and stymie freedom of speech online. When we turn to the countries of the former Soviet Union—Russia and the Commonwealth of Independent States (CIS)—we should expect no different. The Economist *index of democracy* paints a bleak picture of political freedoms in the CIS (see Table 2.1; numbers represent the country's rank in the world).[1] Only two countries, Ukraine and Moldova, rank as *flawed democracies*, with the remaining 10 countries of the region described as either *hybrid regimes* or *authoritarian*.

Throughout the CIS, this creeping authoritarianism is evident in just about every facet of social and political life. Independent media are stifled, journalists intimidated, and opposition parties and civil society groups harassed and subject to a variety of suffocating regulations. And yet, in spite of this increasingly constrained environment, the Internet remains accessible and relatively free from filtering. The ONI has tested extensively through the CIS region, far deeper and more regularly in fact than in any other region in the world. To date we have documented traditional ''Chinese-style'' Internet filtering—the deliberate and static blocking of Internet content and services by state sanction—only in Uzbekistan and Turkmenistan. For the rest of the region, while connectivity may be poor and unreliable, and suffer from the usual rent-seeking distortions found in other developing country environments, the same basic content is available there as in the most open country contexts.

In our chapter, we explore this seeming disjuncture between authoritarianism in the CIS and the relative freedom enjoyed in Russian cyberspace, commonly known as RUNET. We argue that attempts to regulate and impose controls over cyberspace in the CIS are not necessarily absent (as ONI testing results may suggest) but are *different* than in other regions of the world. We hypothesize that CIS control strategies have

Table 2.1

INDEX OF DEMOCRACY		
Less Authoritarian	**World Ranking**	
Ukraine	53	Flawed democracy
Moldova	62	
Georgia	104	Hybrid regime
Russia	107	
Armenia	113	
Kyrgyzstan	114	
Kazakhstan	127	Authoritarian
Belarus	132	
Azerbaijan	135	
Tajikistan	150	
Uzbekistan	164	
Turkmenistan	165	
More Authoritarian		

Source: The Economist Intelligence Unit, ''The Economist Intelligence Unit's Index of Democracy 2008,'' 2008, http://graphics.eiu.com/PDF/Democracy%20Index%202008.pdf.

evolved several generations ahead of those used in other regions of the world (including China and the Middle East). In RUNET, control strategies tend to be more subtle and sophisticated and designed to *shape and affect* when and how information is received by users, rather than denying access outright.

One reason for this difference may be the prior experiences of governments and opposition groups in the region. State authorities are aware of the Internet's potential for mobilizing opposition and protest that goes far beyond the nature of content that can be downloaded from Web sites, chat rooms, and blogs. These technologies have the potential to enable *regime change*, as demonstrated by the eponymous color revolutions in Ukraine, Georgia, and Kyrgyzstan. By the same token, state actors have also come to recognize that these technologies make opposition movements vulnerable, and that disruption, intimidation, and disinformation can also cause these movements to fragment and fail. The failure of opposition movements in Belarus and Azerbaijan to ignite a wider social mobilization, along with the role that targeted information controls played in fragmenting and limiting the effectiveness of these movements, also points to the possible trajectory in which controls aimed at Russian cyberspace may be moving.

Our chapter unfolds in several steps. We begin by describing some of the unique characteristics of the ''hidden'' information revolution that has taken place in Russian cyberspace since the end of the cold war. Contrary to widespread perceptions outside of

the region, Russian cyberspace is a thriving and dynamic space, vital to economics, society, and politics. Second, we outline *three generations* of cyberspace controls that emerge from the research conducted by the ONI in this region. *First-generation controls*—so-called Chinese-style filtering—are unpopular and infrequently applied. While instances of filtering have been identified in just about all CIS countries, wide-scale national filtering is only pursued as a matter of state policy in two of the CIS states. Rather, information control seems to be exercised by way of more subtle, hidden, and temporally specific forms of denial. These controls can involve legal and normative pressures and regulations designed to inculcate an environment of self-censorship. Others, like denial-of-service attacks, result in Web sites and services becoming unavailable, often during times of heightened political activity. Still others, like mass blogging by political activists on opposition Web sites, cannot be characterized as an attack per se, although the outcome of silencing these Web sites is as effective as traditional filtering (if not more so).

These *second-* and *third-generation* controls are increasingly widespread, and they are elusive to traditional ONI testing methods. They are difficult to measure and often require in-depth fieldwork to verify. Consequently, many of the examples in this chapter are based on field investigations carried out by our ONI regional partners where technical testing was used to establish the characteristics of controls, rather than measure the extent of them. We hypothesize that, although these next-generation controls emerged in the CIS, they may in fact be increasingly practiced elsewhere. In the next section of the chapter we turn our lens beyond the CIS to find examples of second- and third-generation controls.

We conclude by arguing that, contrary to initial expectations, first-generation filtering techniques may become increasingly rare outside of a few select content categories, raising serious public policy issues around accountability and transparency of information controls in cyberspace. The future of cyberspace controls, we argue, can be found in RUNET.

RUNET

On July 6, 2006, Russian President Vladimir Putin fielded questions from the Internet at an event organized by the leading Russian Web portal Yandex.[2] It was the first time a Russian leader directly engaged and interacted with an Internet audience. The event itself made few headlines in the international media, but in Russia it marked an important milestone. The Internet had graduated to the mainstream of Russian politics and was being treated by the highest levels of state authority as equal in importance to television, radio, and newspapers. The question put to President Putin by the Internet audience also revealed a sense of the informal, irreverent culture of Russian cyberspace. Over 5,640 *netizens* wrote in to ask when the President first had sex. More surprising, perhaps, was that Putin replied.[3]

The rise of the Internet to the center of Russian culture and politics remains poorly understood and insufficiently studied. With the end of the cold war and the demise of the USSR, Russia and the CIS entered into a long period of decline. Economies stagnated, political systems languished, and the pillars of superpower status—military capacities and advanced scientific and technological potential—rapidly ebbed away. Overnight, the CIS become less relevant and dynamic. The precipitously declining population rates in the Slavic heartland, a wholesale free-for-all of *mafiya*-led privatization, growing impoverishment, and failing public infrastructure, all made the distant promise of a knowledge revolution led by information technologies seem highly improbable.

Moreover, the prospects for Russia and the CIS keeping up with the Internet and telecom boom of the late 1990s and early 2000s seemed, for many, a distant reality. By the time the USSR finally collapsed in 1991, it had the lowest teledensity of any industrialized country. Its capacity for scientific development, particularly in the field of PCs (which the USSR had failed to develop) and computer networking (which was based on reverse-engineered systems pirated from European countries) was weak to nonexistent. Moreover, Russian seemed to be a declining culture and language as newly independent CIS countries adopted national languages and scripts, and preferred to send their youth to study at Western institutions. In almost every major indicator of economic progress, political reform, scientific research, and telecommunications capacity, the countries of the CIS seemed headed for the dustheap of history. Not surprisingly, scholarly and policy interest in the effects and impact of the information revolution in the CIS waned, as attention focused on the rising behemoths in Asia (particularly China and India), and the need and potential of bridging the *digital divide* in Africa and the Middle East. And yet, during the last decade the CIS has undergone a largely unnoticed information revolution. Between 2000 and 2008 the Russian portion of cyberspace, or RUNET, which encompasses the countries of the CIS, grew at an average rate of 7,208 percent, or over five times the rate of the next faster region (Middle East) and 15 times faster than Asia (see Table 2.2).

More than 55 million people are online in the CIS, and Russia is now the ninth-largest Internet country in terms of its percentage of world users, just ahead of South Korea.[4] By latest official estimates, 38 million Russians, or a third of the population of the Russian Federation, are connected, with over 60 percent of those surfing the Internet from home on broadband connections. And these figures may be low. Russian cyberspace also embraces the global Russian diaspora that, through successive waves of emigration, is estimated at above 27 million worldwide. Many Russian émigrés reside in developed countries, but tend to live *online* in the RUNET. Statistics to back this claim are methodologically problematic, but anecdotal evidence suggests that this is the case. The popular free mail service mail.ru, for example, boasts over 50 million user accounts, suggesting that the number of inhabitants in Russia cyberspace may be significantly above the 57 million users resident in the CIS. And these figures are

Table 2.2

PROFILE OF INTERNET USE, PENETRATION, AND GROWTH IN THE CIS

Country	Population (2008)	Number of Internet Users	Internet Penetration (2008)	Internet Growth (2000–2008)
Armenia	2,968,586	172,800	5.8%	476%
Azerbaijan	8,177,717	1,500,000	18.3%	12,400%
Belarus	9,685,768	2,809,800	29%	1,461%
Georgia	4,630,841	360,000	7.8%	1,700%
Kazakhstan	15,340,533	1,900,600	12.4%	2,615.1%
Kyrgyzstan	5,356,869	750,000	14%	1,353.5%
Moldova	4,324,450	700,000	16.2%	2,700%
Tajikistan	7,211,884	484,200	6.7%	24,110%
Turkmenistan	4,829,332	70,000	1.4%	3,400%
Russia	140,702,094	38,000,000	27%	1,125.8%
Ukraine	45,994,287	6,700,000	14.6%	3,250%
Uzbekistan	27,345,026	2,400,000	8.8%	31,900%
Totals	267,567,387	55,847,400	20% (average 13.5%)	7,208%

Source: Miniwatts Marketing Group, ''Internet World Statistics, 2009,'' http://www.internetworldstats.com.

set to rise—dramatically. By official predictions, Russia's Internet population is set to double to over 80 million users by 2012.[5]

Paradoxically, the very *Russianness* of the RUNET may have contributed to hiding this ''cyber revolution.'' Unlike much of the Internet, which remains dominated by English and dependent on popular applications and services that are provided by U.S.-based companies (such as Google, Yahoo, and Hotmail), RUNET is a self-contained linguistic and cultural environment with well developed and highly popular search engines, Web portals, social network sites, and free e-mail services. These sites and services are modeled on services available in the United States and the English-speaking world but are completely separate, independent, and only available in Russian.[6] In a recent ranking of Internet search engines, the Russian Web portal Yandex was one of only three non-English portals to make the top ten, and was only beaten out by a Baidu (China) and NHK (Korea), both of which have much larger absolute user base.[7] Within RUNET, Russian search engines dominate with Yandex (often called the Google of Russia), beating out Google with 70 percent of the market (Google has between 18 and 20 percent).[8]

The RUNET is also increasingly central to politics. Elections across the CIS are now fought online, as the Internet has eclipsed all the mass media in terms of its reach, readership, and especially in the degree of free speech and opportunity to mobilize that it provides. By 2008, Yandex could claim a readership larger than that of the

popular mainstream newspapers *Izvestia, Komsomolskaya Pravda*, and *Moskovsky Komsomolets* combined.[9] The Russian-language *blogosphere*—which currently makes up 3 percent of the world's 3.1 million blogs—grows by more than 7,000 new blogs per day.[10] There are currently more Russian-language blogs than there are French, German, or Portuguese, and only marginally fewer than Spanish,[11] which is spoken by a larger percentage of the world population.[12]

This shift has been fueled as much by the growing state control over the traditional mass media as it has been by the draw of what the new online environment has to offer. Well-known journalists, commentators, and political figures have all turned to the RUNET as the off-line environment suffers through more severe restrictions and sanctions. Across the CIS, especially in the increasingly authoritarian countries of Uzbekistan, Belarus, and Kazakhstan, the RUNET has become the last and only refuge of public debate. Given its rapid ascent to the popular mainstream, it is paradoxical—and certainly a puzzle—that RUNET has elided filtering controls of the kind imposed by China on its Internet in all but a few countries. In the next section, we explore why that is the case.

Next-Generation Information Controls in the CIS

Although RUNET is a wild hive of buzzing online activity, it is not completely unregulated. Since its emergence in the early 1990s, RUNET has been subject to a variety of controls. Some controls have been commercial in motivation and represent crude attempts to use formal authority to create what amounts to a monopoly over secure communications and as means to seek rents.[13] This form of control has not been unique to RUNET and has extended to every other facet of post-Soviet life, from car registration through to the supply of gasoline, as an aspect of the great scramble to *prihvatizatsia* public assets that occurred during the early to mid 1990s.[14] Other controls have emerged from a legal system inherited from the Soviet era, which criminalized activities without necessarily seeking prosecution, except selectively. These forms of control effectively form the rules of the game for all informal networks. Their emergence in the virtual online world of the RUNET is transparent and natural.

But during the late 1990s, and especially following the color revolutions that swept through the CIS region, states began to think seriously about the security implications of RUNET, and in particular its potential to enable mobilization of mass social unrest. The first attempts at formally controlling cyberspace were legal, beginning with legislation enabling surveillance (SORM-II),[15] and later in 2011 with the publication of Russia's *Doctrine of Information Security*. While the doctrine addressed mass media and did not focus on RUNET specifically, it declared the information sphere to be a vital national asset that required state protection and policing. The doctrine used strong language to describe the state's right to guide the development of this space, as well as its responsibility to ensure that information space respects ''the stability of the constitu-

tional order, sovereignty, and the territorial integrity of Russian political, economic and social stability, the unconditional ensuring of legality, law and order, and the development of equal and mutually beneficial international cooperation.''[16]

The intent of the doctrine was as much international as it was domestic, establishing demarcated borders in cyberspace, at least in principle. The international intent of the doctrine appears to have been driven by a growing concern that Russia was falling behind its major adversaries in developing a military capability in cyberspace; efforts by countries such as the United States, China, India, and others to develop covert computer network attack capabilities risked creating a strategic imbalance.[17] Domestically, the doctrine was aimed at the use of the Internet by militant groups to conduct information operations, specifically the Chechen insurgency. Within a few years, most other CIS countries had followed suit, adopting variations of the Russian doctrine.

ONI Tests for Internet Controls in RUNET

The controls outlined previously are qualitatively different from the usual types of controls for which the ONI tests. Establishing empirical evidence of the effects of policies like SORM and the Doctrine of Information Security is challenging, since their application is largely contextual, their impact at times almost metaphysical. Such controls do not yield a technological ''fingerprint'' in the way that a filtering system blocking access to Internet content does. However, they may be just as effective, if not more so, in achieving the same outcomes. In its 2007 study of the policy and practice of Internet filtering, the ONI found that substantial and pervasive attempts to technically filter content on RUNET did not begin until 2004, and even then were isolated to Turkmenistan and Uzbekistan, with lesser attempts at filtering found in most other CIS countries (see Table 2.3)[18]

These reports have remained consistent in more recent rounds of ONI tests. And yet persistent *anecdotal* reports, as well as special monitoring efforts mounted by the ONI, reveal in the majority of CIS countries that *information denial* and *access shaping* is occurring, and on a significant scale, especially around critical events such as elections. The ONI carried out a number of special investigations, including mounting monitoring efforts during the 2005 parliamentary elections in Kyrgyzstan[19] and the March 2006 Belarus presidential elections.[20] These efforts yielded the first technically verified results that the RUNET was being deliberately tampered with to achieve a political effect.

The results obtained by ONI in the CIS are unique, and they differ significantly from the results obtained in ONI's global survey. They demonstrate that information controls in the CIS have developed in different ways and using different techniques than those found in other areas of the world. They suggest a much more sophisticated approach to managing networks through denial that is highly selective and event based, and that *shapes* access to the sources of information and means of

Table 2.3

SUMMARY RESULTS FOR ONI TESTING FOR INTERNET FILTERING, 2007–2008					
	No Evidence	Suspected	Selective	Substantial	Pervasive
Armenia			•		
Azerbaijan			•		
Belarus			•		
Georgia			•		
Kazakhstan			•		
Kyrgyzstan			•		
Moldova	•				
Tajikistan			•		
Turkmenistan					•
Russia			•		
Ukraine	•				
Uzbekistan			•	•	

communication in a manner that could plausibly be explained by errant technical failures or other random network effects. In the following sections, we define the three different generations of cyberspace controls and provide examples for each from our research in the CIS region. The three generations of controls are also summarized in Table 2.4.

First-Generation Controls

First-generation controls focus on denying access to specific Internet resources by directly blocking access to servers, domains, keywords, and IP addresses. This type of filtering is typically achieved by the use of specialized software or by implementing instructions manually into routers at key Internet choke points. First-generation filtering is found throughout the world, in particular among authoritarian countries, and is the phenomenon targeted for monitoring by the ONI's methodology. In some countries, compliance with first-generation filtering is checked manually by security forces, who physically police cybercafés and ISPs.

In the CIS, first-generation controls are practiced on a wide scale only in Uzbekistan and Turkmenistan. In Uzbekistan, a special department of the SNB (KGB) monitors the Internet and develops block lists that are then conveyed to individual ISPs who in turn implement blocking against the specific resources or domain names. The filtering is universal across all ISPs, and the SNB spot-checks ISPs for compliance. In Turkmenistan, filtering is centralized on the country's sole ISP (operated by Turkmentelekom), and access is heavily filtered. Up until late 2007, Internet access in Turkmenistan was severely restricted and expensive, limiting its access and impact.

Table 2.4

SPECTRUM OF CYBERSPACE CONTENT CONTROLS IN THE CIS

	First Generation		Second Generation				Third Generation			
	Internet Filtering	Policing Cybercafés	Legal Environment for Information Control[1]	Informal Removal Requests	Technical Shutdowns	Computer Network Attack	Warrantless Surveillance	National Cyberzones	State-Sponsored Information Campaigns	Direct Action
Armenia			•	•	•	•[2]				
Azerbaijan			•	•		•[2]				•
Belarus	•		•	•	•	•	•	•	•	•
Georgia			•							
Kazakhstan	•		•	•				•	•	•
Kyrgyzstan			•	•		•[3]				
Moldova			•				•	•		
Tajikistan			•	•				•		
Turkmenistan	•	•	•				•	•		
Russia			•	•		•	•		•	•
Ukraine			•							
Uzbekistan	•	•	•				•			•

1. Legal and Normative Environment for Information Control includes the following:

a. Compelling Internet sites to register with authorities and using noncompliance as grounds for filtering ''illegal'' content.

b. Strict criteria pertaining to what is ''acceptable'' within the national media space, leading to the de-registration of sites that do not comply.

c. Expanded use of defamation, slander, and ''veracity'' laws to deter bloggers and independent media from posting material critical of the government or specific government officials.

d. Evoking national security concerns, especially at times of civic unrest, as the justification for blocking specific Internet content and services.

e. Legal regime for Internet surveillance.

2. CNA has been used by both Azeri and Armenian hackers in an ongoing series of attacks. It is unclear whether these are the actions of individual hackers, or whether these groups receive tacit or direct support from the state. Attacks are directed against the Web sites of the opposing country, so are not a content control mechanism.

3. The DDoS attacks were outsourced to commercial ''black hat'' hackers in Ukraine. The party ordering attacks is unknown, but suspicion falls on rogue elements inside the security services.

A second practice associated with first-generation blocking is policing and surveillance of Internet cafés. In Uzbekistan, SNB officers monitor Internet cafés, often enlisting café owners to notify them of individual users who try to access ''banned'' sites. Many Uzbek Internet cafés now openly post notices that viewing illegal sites is subject to fine and arrest. On several occasions, ONI researchers have manually verified the surveillance.

Second-Generation Controls

Second-generation controls aim to create a legal and normative environment and technical capabilities that enable state actors to deny access to information resources as and when needed, while reducing the possibility of blowback or discovery. Second-generation controls have an overt and a covert track. The overt track aims to legalize content controls by specifying the conditions under which access can be denied. Instruments here include the doctrine of information security as well as the application of existent laws, such as slander and defamation, to the online environment. The covert track establishes procedures and technical capabilities that allow content controls to be applied ''just in time,'' when the information being targeted has the highest value (e.g., during elections or public demonstrations), and to be applied in ways that assure plausible deniability.

The legal mechanisms used by the overt track vary from country to country, but most share the characteristic of establishing double jeopardy for RUNET users, making requirements such that compliance sets the grounds for prosecution, and noncompliance establishes a legal basis for sanction.

The following are among the more common legal mechanisms being applied:

Compelling Internet sites to register with authorities and to use noncompliance as grounds for taking down or filtering ''illegal'' content, and possibly revoking service providers' licenses. This tack is effectively used in Kazakhstan and Belarus, and it is currently being considered in Russia. The mechanism is particularly effective because it creates multiple disincentives for potential Web site owners who must go through the hassle of registering with authorities, which leaves them open to legal sanction should their site be deemed to be carrying illegal content. It also creates double jeopardy for international content providers (such as the BBC, CNN, and others) and opens the question whether they should register their services locally. In practice, the registration requirement applies to them so long as their audience is local, and a failure to comply leaves open the option to filter their content for ''noncompliance'' with local registration requirements. On the other hand, registering would make the content they carry subject to local laws, which may deem their content ''unacceptable'' or ''slanderous'' and could lead to legally sanctioned filtering.

Strict criteria pertaining to what is ''acceptable'' within the national media space, leading to the de-registration of sites that do not comply. In Kazakhstan, opposition Web sites or Web sites carrying material critical of the government are regularly de-registered from the national domain. This includes a large number of opposition sites and, notably, the *Borat* Web site, ostensibly because the owners of the site were not resident in Kazakhstan as required by the Kazakh domain authority. In Belarus, the popular portal tut.by refused to put up banners advertising opposition Web sites, possibly for fear of reprisals (although those fears were not made explicit).[21]

Expanded use of defamation, slander, and ''veracity'' laws, to deter bloggers and independent media from posting material critical of the government or specific government officials, however benignly (including humor). In Belarus, slander laws were used to prosecute an owner of a Web site posting cartoons of the president. In both Belarus and Uzbekistan, the law on mass media requires that reporting passes the ''objectivity test.'' Journalists and editors are held responsible for the ''veracity'' of publications and postings, leading to a high degree of self-censorship. In Kazakhstan, there are several cases of oppositional and independent media Web sites being suspended for providing links to publications about corruption among senior state offices and the president.

Evoking national security concerns, especially at times of civic unrest, as the justification for blocking specific Internet content and services. Most recently, this justification was evoked in Armenia when the opposition demonstrations that followed the February 2008 presidential elections turned to violence leading to the death and injury of several dozen protesters. A 20-day state of emergency was declared by President Kocharian, which also led to the de-registration of popular Armenian political and news sites, including a site carrying the Armenian-language BBC service and the filtering of YouTube (ostensibly because of allegations that footage of the rioting had been posted to the popular video sharing site).[22] Similar filtering occurred during the Russian-Georgian crisis of 2008 when Georgia ordered ISPs to block access to Russian media. The blocks had the unintended consequence of creating panic in Tbilisi, as some Georgians perceived the blocks as a signal of impending Russian invasion of the capital.

The technical capabilities typical of second-generation controls are calibrated to effect ''just-in-time'' or event-based denial of selected content or services.[23] These techniques can be difficult to verify, as they can be made to look like technical errors. One of the more common techniques involves formal and informal requests to ISPs. Providers in the CIS are under constant pressure to comply with government requests or face any number of possible sanctions if they do not, from visits from the taxation police to revocations of their licenses. Such pressures make them vulnerable to requests from authorities, especially those that are conveyed informally. In Russia, top-level ISPs are in the hands of large telecommunication companies, such as TransTeleKom and Rostelecom, with strong ties to the government. These providers appear

responsive to informal requests to make certain content inaccessible, particularly when information could prove embarrassing to the government or its officials. In one such case, the popular Russian site—Kompromat.ru—known for publishing documents and photographs of corrupt or illegal practices (roughly analogous to the Web site wikileaks.com) was de-registered or filtered by several top-level ISPs (including TransTeleKom and Rostelekom). Service was later restored, and the blocking of the site was deemed ''accidental.'' Nonetheless, the Web site was inaccessible throughout the February 2008 Russian presidential poll.[24] Similar incidents have been documented in Azerbaijan, where Web sites critical of President Ilham Aliyev were filtered by ISPs, apparently at the request of the security department of the office of the president. [25] A similar dynamic is found in Kazakhstan, where a number of Web sites are inaccessible on a regular basis, with no official reason ever being given.[26]

Other, less subtle but nonetheless effective technical means include shutting down Internet access, as well as selected telecommunications services such as cell phone services and especially short message services (SMS). Temporary outages of the Internet and SMS services were employed by Belarus authorities during the February 2006 presidential elections as a means to limit the ability of the opposition to launch street demonstrations of the type that precipitated the color revolutions in Ukraine, Georgia, and Kyrgyzstan. At first, authorities denied that any interruptions had taken place, and later they attributed the failures to technical reasons.[27] Similar instances were reported (although not verified) to have occurred during the 2007 elections in Azerbaijan.

Second-generation techniques also make extensive use of computer network attacks, especially the use of distributed denial of service (DDoS) attacks, which can overwhelm ISPs and selected sites, and which make tracking down perpetrators difficult, since the attacks themselves are sold and engineered by ''black hat hackers'' and can be ordered by anyone. Such attacks were used extensively during the 2005 Kyrgyz presidential elections that precipitated the Tulip revolution.[28] They were also used during the 2006 Belarus elections against opposition political and news sites. In 2008, presidential and parliamentary elections in many parts of the region saw the significant use of DDoS attacks against the Web sites of major opposition leaders as well as prominent human rights groups. Recently, computer network attacks have been conducted by state-sanctioned ''patriotic hackers'' who act as vigilantes in cyberspace. A Russian hacker who admitted that officers from the FSB encouraged him brought down the pro-Chechen Web site ''Kavkaz center'' repeatedly.[29] There is strong suspicion that the May 2007 DDoS attacks that brought down most of Estonia's networks were the work of state-sanctioned ''patriotic hackers'' responding to unofficial calls from the FSB to ''punish'' Estonia over the removal of a monument to Soviet soldiers in Tallinn. Such attacks were also a prominent feature of the Russian-Georgian crisis of 2008. Several prominent investigations have been undertaken to determine attribution

in this case—including an ongoing one by the ONI's sister project, the *Information Warfare Monitor*—and to date no definitive evidence has been found linking the attacks to the Russian security forces.

Third-Generation Controls

Unlike the first two generations of content controls, third-generation controls take a highly sophisticated, multidimensional approach to enhancing state control over national cyberspace and building capabilities *for competing in informational space* with potential adversaries and competitors. The key characteristic of third-generation controls is that the focus is less on *denying* access than successfully *competing* with potential threats through effective counterinformation campaigns that overwhelm, discredit, or demoralize opponents. Third-generation controls also focus on the active use of surveillance and data mining as means to confuse and entrap opponents.

Third-generation controls include enhancing jurisdiction over national cyberspace and expanding the powers of state surveillance. These include warrantless monitoring of Internet users and usage. In 2008, Russia expanded the powers previously established by SORM-II, which obliged ISPs to purchase and install equipment that would also permit local FSB offices to monitor the Internet activity of specific users. The new legislation makes it possible to monitor all Internet traffic and personal usage without specific warrants. The legislation effectively brings into the open covert powers that were previously assigned to FAPSI, with the twist of transferring to the ISPs the entire costs associated with installing the necessary equipment. The SORM-II law was widely used as a model for similar legislation in other CIS counties, and it is expected that the new law will likewise become a standard in the CIS. Although it is difficult to verify the use of surveillance in specific incidences, inferences can be drawn from specific examples. In July 2008, a Moldovan court ordered the seizure of the personal computers of 12 individuals for allegedly posting critical comments against the governing party. The people were accused of illegally inciting people ''to overthrow the constitutional order'' and ''threaten the stability and territorial integrity of the Republic of Moldova.'' It is unknown how the authorities obtained the names of the people, but some suggest that an ISP provided them with the IP addresses of the users.[30]

Several CIS countries are also pursuing the creation of national cyberzones. Countries such as Kazakhstan, Tajikistan, and Russia are investing heavily into expanding Internet access to schools. These institutions are being tied to special Internet connections, which limit access only to resources found in the national Internet domain. These ''national zones'' are popular among some Tajik and Kazakh ISPs because they allow the ISPs to provide low-cost connectivity, as traffic is essentially limited to the national segment. In 2007, Russian authorities floated the idea of creating a separate Cyrillic cyberzone, with its own domain space and addressing scheme. National cyberzones

are appealing because they strengthen the degree of national control over Internet content. They also appeal to consumers, since access to them is less costly and the resources that can be found there are almost exclusively in the local language.

Other aspects of third-generation controls, such as state-sponsored information campaigns in cyberspace, are difficult to document, as they use surveillance, interaction, and direct physical action to achieve a disruption of target groups or networks. The intent of these campaigns is to effect cognitive change rather than to deny access to online information or services. The ultimate source of these campaigns is also difficult to attribute and can only be established through careful research or insider knowledge, since they are designed to render opaque the role of state actors. These techniques include employing ''Internet Brigades'' to engage, confuse, or discredit individuals or sources. Such action can include the posting of prepackaged propaganda, *kompromat*, and disinformation through mass blogging and participation in Internet polls, or harassment of individual users, including the posting of personal information.[31] This technique, along with the use of surveillance of Internet traffic to affect direct action, saw a marked increase in the run-up to parliamentary and presidential polls in Russia. Numerous accounts allege that progovernment forces monitored opposition Web sites and disrupted planned rallies and marches. In some cases, members of the opposition were warned by cell phone not to participate in rallies or risk being beaten. In other cases, false information was disseminated by progovernment forces, leading to confusion among opposition supporters and, in one documented case, leading them into an ambush by progovernment supporters where several were severely beaten.

Assessing the Evolution of Next-Generation Controls in the CIS

The three generations of controls are not mutually exclusive, and several can exist concurrently. Taken together, they form a pattern of control that is both unique to each country and generalizable to the region as a whole. However, the degree to which a country is more or less authoritarian does seem to influence the choice of ''generational mix'' applied. Countries with stronger authoritarian tendencies tend to apply more comprehensive information controls in cyberspace, often using all three generations of controls. Conversely, countries that are ''more democratic'' tend to favor second- and third-generation strategies. None of the six countries scoring as ''hybrid regimes'' or ''flawed democracies'' applied first-generation controls (see Figure 2.1).

Several factors can explain this pattern. The most obvious explanation of the general tendency is that authoritarian states will seek to dominate the public sphere. These states tend to be the most vulnerable to mass unrest, prompting additional efforts by security forces to ensure that all channels of potential mobilization are controlled. A second factor worth noting is that these six counties are also experiencing the fastest rates of Internet growth and, with the exception of Belarus, have among the lowest

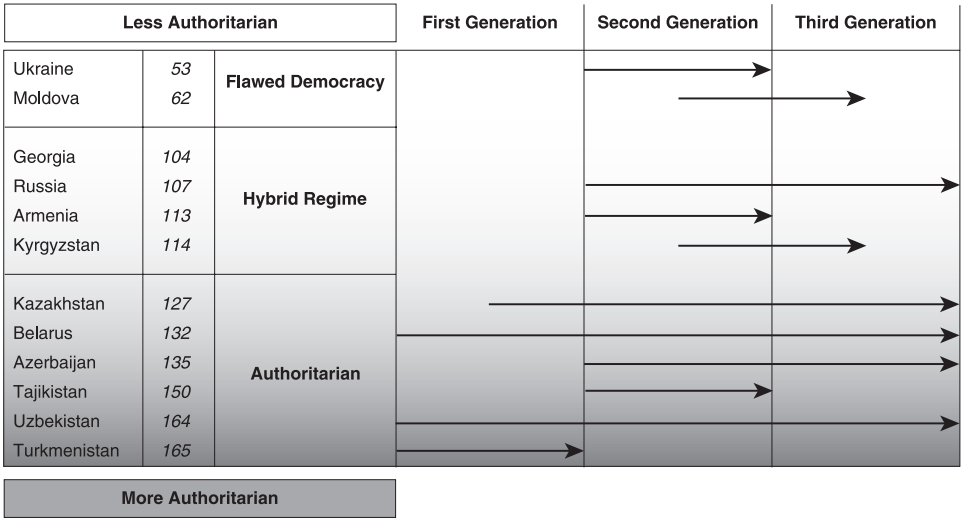


Figure 2.1
Spectrum of cyberspace content controls in the CIS (clustered by generation and EIU Index of Democracy)
Source: The Economist Intelligence Unit, "The Economist Intelligence Unit's Index of Democracy 2008," 2008, http://graphics.eiu.com/PDF/Democracy%20Index%202008.pdf.

levels of Internet penetration in the region. This latter explanation, which suggests that the RUNET in these counties has not become the locus for informal networks that it has in some of the less authoritarian countries, may make it more vulnerable and a target for filtering controls than what would be the case elsewhere in the CIS where the RUNET is more central to the political mainstream. In this respect, the maturity of the network itself seems to influence the degree to which filtering controls will be applied. This observation begs the obvious question—will the RUNET remain open even as countries in the CIS slide toward a new authoritarianism?

While the possibility of greater direct content controls being applied in the RUNET certainly exists, there is a far greater potential that information controls will continue to evolve along the evolutionary trajectory, toward strategies that seek to compete, engage, and dominate opponents in the informational battle space through persistent messaging, disinformation, intimidation, and other tactics designed to divide, confuse, and disable. In this respect, the patterns of information control in the CIS may in fact represent a model that will evolve elsewhere as governments are faced with the choice of imposing harsh controls and being labeled pariahs or doing nothing and risking that the technologies could become enablers of hyperdemocracy and undesired regime change.

Conclusion: Next-Generation Controls Beyond the CIS?

There are several obvious and not so obvious reasons to believe that second- and third-generation controls will become more common outside of the CIS and in fact may presage the future of cyberspace controls as a whole. First, the experience from other regions suggests that first-generation filtering is easy to circumvent. The ''Great Firewall of China'' is easily breached, as evidenced by the growing number of circumvention technology solutions, from Tor to Psiphon and others. As such techniques become more common, enabled and supported by large-scale and distributed efforts in the United States and Europe, the incentives to rely on less technologically static and temporally fixed methods characteristic of next-generation controls will likely grow.

It is also questionable whether first-generation controls in countries like Burma, North Korea, and China are really sustainable in the long run. In China's case, the floodgates may open sooner rather than later as the Chinese Internet itself becomes much more central to popular culture. First-generation filtering practices can produce economic and other social costs through collateral filtering and disincentives for foreign direct investment and tourism. As countries become more dependent on cyberspace for research, business, and other international communications, the friction introduced by filtering becomes increasingly unpopular, costly, and impractical.

More important than these factors, however, is the growing legitimization and frequent practice of policing the Internet through indirect and distributed means, and in particular through third parties, including the entities that actually support the cyberspace infrastructure, from connectivity to hosting to social networking platforms. Since much of cyberspace is operated by the private sector, there are practical and legal limits to the direct reach of government controls. Controls have thus evolved downward and in a distributed fashion, in a significant privatization of authority, in conformity with second- and third-generation controls outlined previously. Naturally, the scope for second- and third-generation controls differs among authoritarian and democratic countries, but examples of each can be found in both contexts.

In China, for example, while much of the attention focuses on the technologies of the Great Firewall of China filtering access to the Internet, at least as much, if not more, of the information controls exercised in that country happen in a more distributed fashion and by private actors. Web hosting and social networking services are now routinely obliged to sign self-discipline pacts and follow rigid hosting protocols that limit what can be communicated online; search engines—including those owned by American companies like Google, Microsoft, and Yahoo!—routinely filter their search results, often more aggressively than the government does itself; and in the most extreme example, volunteer citizen groups—sometimes known in China as *50 cent brigades* for the amount they are purportedly paid for each post—swarm the Internet's chat rooms, blogs, and other public forums making statements favorable to the government.[32] The latter was dramatically demonstrated, in a clear example of

third-generation controls, during the time of the Olympics, when thousands of Chinese bloggers posted aggressively to counter what they perceived as anti-Chinese propaganda.[33] Whether the volunteer posts were managed or encouraged by the state, or simply benefited the state coincidentally, or some combination, is a vexing question nearly impossible to untangle. Such attribution problems are, in fact, one of the key characteristics of second- and third-generation controls and one of their greatest challenges for research projects like the ONI.

Outside of authoritarian contexts and among democratic countries, it is now common to hear of legal and market pressures being invoked to remove content from Web hosting and social networking platforms, and there is also a very noticeable trend to offload policing activities to ISPs, particularly in the areas of content controls around pornography, hate speech, and copyright violations. In fact, most industrialized democratic countries have passed far-reaching surveillance measures that enable widespread eavesdropping on e-mail, cellular phone, and other communications activities by requiring ISPs to retain and, when required, turn over such information to legal authorities.

Perhaps the strongest impetus toward second- and third-generation controls has emerged from a growing emphasis on cyber security and the recognition of cyberspace as a domain of military action. Military actors have come to understand cyberspace as a domain equal in importance to land, air, sea, and space, requiring a full spectrum of capabilities. This has meant developing weapons and tactics designed to disrupt, destroy, and confuse potential adversaries. For the most part, these capabilities have been kept quiet and under classification, but they are similar in intent and execution to the network attacks characteristic of second-generation information controls. Russia, China, and the United States have all developed doctrines and capabilities for operations in cyberspace that include computer network attacks, as well as psychological operations designed to shape the domain through selective filtering, denial of access to information, and information engagement. The intent and effect of these emerging doctrines is the same as those we have documented in second- and third-generation controls in the CIS—to silence information that is strategically threatening and sow confusion and doubt among opponents dependent on cyberspace for information and organization.

Overall, the lexicon of cyber security is shifting norms around acceptable behavior for intervention into cyberspace and generating new incentives for technological development. Pervasive surveillance, including deep packet inspection, is now an acceptable part of compliance with good security practices, despite the impacts on privacy protections. Similarly, the political rush to secure cyberspace is generating economic opportunities not seen since the Internet boom of the 1990s. However, unlike the 1990s when the rush was led by companies seeking to open up cyberspace, the current momentum is in the other direction. The fact that defense contractors are now lining up to compete in this domain only raises the troubling concerns that some of the

valuable freedoms gained over the last 15 years in cyberspace will be sacrificed at the altar of security.

These are troubling tendencies, and ones with implications far outside of the democratic countries of the OSCE. The confluence of second- and third-generation controls, the militarization of cyberspace, and the legitimization of surveillance are contributing to a dangerous brew. The cyberspace enjoyed by the next generation of users may be a very different, more regulated, and less empowering domain than that which was taken for granted in the past.

Notes

1. The Economist Intelligence Unit, *The Economist Intelligence Unit's Index of Democracy 2008* (The Economist, 2008), http://graphics.eiu.com/PDF/Democracy%20Index%202008.pdf.

2. Radio Free Europe, ''Putin Quizzed by Internet Users,'' July 6, 2006, http://www.rferl.org/featuresarticle/2006/07/40C4C298-C619-4FCC-9D9D-D24787E10EAB.html.

3. ''When did you start to have sex?'' asked Kommersant reporter Andrei Kolesnikov on behalf of 5,640 Internet users. ''I don't remember when I started. But I can remember the last time,'' Putin replied. *St Petersburg Times*, ''Putin Weighs In on Robots, Sex Following Internet Conference,'' July 11, 2006, http://www.sptimes.ru/index.php?action_id=2&story_id=18178.

4. As of March 31, 2009, the top five countries with the highest number of Internet users, in order are China, United States, Japan, India, and Brazil. See Miniwatts Marketing Group, ''Top 20 Countries with the Highest Number of Internet Users,'' March 31, 2009, http://www.internetworldstats.com/top20.htm.

5. Between 2005 and 2007 Internet use in the Russian Federation jumped from 15% to over 28% of the population.

6. Popular sites include rutube, a Russian version of the popular U.S. site YouTube, as well as the social network sites odnoklassniki.ru and vkontakte.ru, which are modeled after Classmates.com and Facebook.com.

7. China has an estimated 298 million Internet users, while Korea possesses over 48 million users and a 76.1 percent Internet penetration. Miniwatts Marketing Group, ''Top 20 Countries with Highest Number of Internet Users,'' March 31, 2009, http://www.internetworldstats.com/top20.htm.

8. More significant yet than its impressive growth has been the emergence of RUNET at the center of Russian popular culture. Internet memes and jokes once marginalized to the community of computer specialists and aficionados are now in the mainstream of popular culture. For example, *Preved Medved*, an allusion to primitive cartoon of a bear surprising a couple having sex in a field and shouting *preved* (a deliberate misspelling of *privet*—a greeting, and *medved*—bear) has become a cultural icon, showing up on the cover of mainstream journals, in advertisements, and even, as a joke, in a question put to President Putin. Similarly, the *Olbanian* language, once an obscure Internet in-joke is now mainstream enough to have warranted a joke by President-elect Medvedev, who, when asked whether it should become a school subject, replied, ''One cannot ignore the

necessity of learning the Olbanian language.'' The slang it inspired has also led to political neologisms, such as *Putings*, which refers to political meetings in support of former President Putin, and whose usage on-air (as opposed to online) landed a Russian TV journalist a stiff fine.

9. LiveInternet, ''Report: From Search Engines, 2009,'' http://www.liveinternet.ru/stat/ru/searches.html?slice=ru.

10. Nick Wilsdon, ''Yandex Releases Autumn Report on Russian Blogosphere,'' Multilingual Search, November 12, 2007, http://www.multilingual-search.com/yandex-releases-autumn-report-on-russian-blogosphere/12/11/2007.

11. Yandex, ''*Sostoyaniye Blogosfery Rossiykogo Interneta*,'' [The State of the Russian Blogosphere], 2007, http://download.yandex.ru/company/yandex_on_blogosphere_autumn_2007.pdf.

12. Raymond Gordon, Jr., ed., *Ethnologue: Languages of the World*, 15th ed. (Dallas: SIL International, 2005).

13. For example, in mid-1995 the Federal Agency for Communication and Information (FAPSI) announced a joint venture with Relcom to create a secure business network. But FAPSI's interests in Relcom had less to do with security than with its growing business interests. FAPSI recognized that the Internet was becoming an important channel for business transactions. It wanted part of this market. Its intervention was part of a broader effort aimed at using its special position with respect to responsibility for state communications and security as means to seek rents from Russian and foreign businesses. By the time the Relcom deal had been announced, FAPSI had already secured legislation that required all use of cryptography in Russia to be licensed by FAPSI. Similarity, it had won the exclusive right to produce smart cards for the Russian market. Both moves had essentially given it a monopoly over the critical technologies required by the banking sector in Russia, as well as a future stake in all e-commerce. FAPSI was also present in the telecommunications market more broadly. The legislation creating Sviazinvest, the state-dominated holding company that owned shares in Russian telecommunications companies, required that senior officers from the service were present on the board of every major telecommunications player. Before long, the boards of most telecommunications companies were filling up with retired ex-FAPSI generals. In 1997 the FAPSI–Relcom deal collapsed, as the agency itself was disbanded over allegations of corruption by its leadership. Its assets, which included most of Russia's signal intelligence capacity, were reabsorbed into the FSB. But by that time, Relcom's dominance of the Internet market in Russia and the CIS was on the wane. The Internet was becoming a profitable business, and telecom operators where quickly entering the Internet market and putting former Relcom nodes out of business.

14. The term ''prihvatizatsia'' is slang, and a neologism of the Russian word for ''theft'' and the English term ''privatize.''

15. See the Russia country profile in this volume.

16. Security Council of the Russian Federation, *Information Security Doctrine of the Russian Federation*, 2000, http://www.scrf.gov.ru/documents/5.html.

17. Two years prior to the publication of this doctrine, Russia began actively working through the UN to establish an arms control regime in cyberspace. A. A. Streltsov, ''International Information Security: Description and Legal Aspects,'' *Disarmament Forum* 3 (2007), 5–13.

18. Ronald J. Deibert, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain, eds., *Access Denied: The Practice and Policy of Global Internet Filtering* (Cambridge, MA: MIT Press, 2008).

19. OpenNet Initiative, ''Special Report: Kyrgyzstan,'' April 15, 2005, http://opennet.net/special/kg/.

20. OpenNet Initiative, ''The Internet and Elections: The 2006 Presidential Election in Belarus (and Its Implications),'' April 2006, http://opennet.net/sites/opennet.net/files/ONI_Belarus_Country_Study.pdf.

21. Mikhail Doroshevich, ''Major Belarusian Internet Portal TUT.BY Introduces Restrictions for Internet Forums,'' E-Belarus.Org, June 28, 2005, http://www.e-belarus.org/news/200506281.html.

22. OpenNet Initiative Blog, ''Armenia Imposes Internet Censorship as Unrest Breaks Out Following Disputed Presidential Elections,'' March 11, 2008, http://opennet.net/blog/2008/03/armenia-imposes-internet-censorship-unrest-breaks-out-following-disputed-presidential-e.

23. Ronald J. Deibert and Rafal Rohozinski, ''Good for Liberty, Bad for Security? Global Civil Society and the Securitization of the Internet,'' in *Access Denied: The Practice and Policy of Global Internet Filtering*, ed. Ronald J. Deibert, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain (Cambridge, MA: MIT Press, 2008), 123–149.

24. As reported by ONI field researchers.

25. See the Azerbaijan country profile in this volume.

26. Reuters, ''Kazakh Bloggers Say Can't Access Popular Website,'' October 10, 2008, http://ca.reuters.com/article/technologyNews/idCATRE4995D020081010.

27. OpenNet Initiative, ''The Internet and Elections: The 2006 Presidential Election in Belarus (and Its Implications),'' April 2006, http://opennet.net/sites/opennet.net/files/ONI_Belarus_Country_Study.pdf.

28. OpenNet Initiative, ''Special Report: Kyrgyzstan,'' April 15, 2005, http://opennet.net/special/kg/.

29. John Varoli, ''In Bleak Russia, a Young Man's Thoughts Turn to Hacking,'' New York Times, June 29, 2000, http://www.nytimes.com/2000/06/29/technology/in-bleak-russia-a-young-man-s-thoughts-turn-to-hacking.html.''

30. Sami Ben Gharbia, ''Moldavia: Sequestration of Personal Computers of 12 Young People for Posting Critical Comments Online,'' Global Voices Advocacy, June 13, 2008, http://advocacy.globalvoicesonline.org/2008/06/13/moldavia-destruction-of-personal-computers/.

31. Anna Polyanskaya, Andrei Krivov, and Ivan Lomko, ''Commissars of the Internet: The FSB at the Computer,'' Vestnik Online, April 30, 2003, http://www.vestnik.com/issues/2003/0430/win/polyanskaya_krivov_lomko.htm.

32. David Bandurski, ''China's Guerrilla War for the Web,'' *Far Eastern Economic Review*, July 2008, http://www.feer.com/essays/2008/august/chinas-guerrilla-war-for-the-web.

33. See the China country profile in this volume.

J. Wright, T. de Souza and I. Brown (2011). “Fine-Grained Censorship Mapping: Information Sources, Legality and Ethics.” FOCI'11 (USENIX Security Symposium), San Francisco, 8 August.

Fine-Grained Censorship Mapping
Information Sources, Legality and Ethics

Joss Wright
Oxford Internet Institute
joss.wright@oii.ox.ac.uk

Tulio de Souza
Oxford University Computing Laboratory
tulio.de.souza@comlab.ox.ac.uk

Ian Brown
Oxford Internet Institute
ian.brown@oii.ox.ac.uk

Abstract

We examine the problem of mapping internet filtering, or censorship, at a finer-grained level than the national, in the belief that users in different areas of a country, or users accessing the internet through different providers or services, may experience differences in the filtering applied to their internet connectivity.

In investigating this possibility, we briefly consider services that may be used by researchers to experience a remote computer's view of the internet. More importantly, we seek to stimulate discussion concerning the potentially serious legal and ethical concerns that are intrinsic to this form of research.

1 Introduction

Many nations around the globe participate in some form of internet filtering[3]. Whilst filtering and censorship can, to an extent, be open and transparent, their nature tends towards secrecy. In order to understand the extent and nature of filtering around the world, we desire the ability to experience directly the limitations imposed on these internet connections.

National-level filtering, however, is simply the crudest form of such mapping. Whilst many states have national filtering policies, there is some evidence that the specific implementation of these may vary from region to region, from ISP to ISP and even from computer to computer. In order to fully understand filtering and its role in the globally networked world, it is extremely useful to explore connectivity at a more geographically and organisationally fine-grained level.

To this end, it is desirable to experience the Internet as viewed by a computer in a location of interest. There are a number of existing services specifically designed to allow this: VPN software and proxy services are well-known tools to allow a remote computer to route through a given remote network, and the well-known Tor anonymising network provides a similar service specifically aimed at bypassing national-level filtering.

For the purposes of wide-scale research, however, many of these services are relatively rare and require explicit access. Further, many of these services are employed directly to avoid filtering and thus to allow filtered users to access unfiltered connections. Clearly, such a service is less likely to exist on heavily filtered connections. In deliberately investigating filtered connections, it may be necessary also to explore other forms of information.

2 Motivation

There are many technical approaches to internet filtering employed around the world, applied to a greater or lesser extent. The most well-known filter is almost certainly China's "Golden Shield" (金盾工程, *jīndùn gōngchéng*), commonly known as the "Great Firewall of China", which represents arguably the largest and most technologically advanced filtering mechanism in use today.

Despite the technological sophistication of the Chinese national firewall, it is subject to a number of limitations. With a population of roughly 1.3 billion and an internet penetration rate estimated at almost 32%, the number of Chinese internet users is comparable to the combined populations of the US and Mexico. At such a scale economies must be made in the mechanisms of filtering to reduce the required resources to a manageable level. An excellent study of the technology underlying the Chinese national firewall was presented by Clayton et al[2].

Many other countries, however, perform internet filtering with significantly lower budgets and technical investment. Technologies range from crude blocking of large portions of the internet, to sophisticated and subtle blocking of specific content. A global view of internet filtering has been comprehensively presented in [3]. This work is

notable not just for its scope, but for its focus on the sociological as well as technical aspects of filtering, covering the nature of filtered topics and the levels of state transparency in the filtering process.

At a national level, however, filtering beyond crude mechanisms is often considered infeasible due not only to computational, but also to the organisational requirements of such systems; even if sufficient technological resources are available, the dynamic nature of the internet imposes a significant administrative burden in maintaining up-to-date filtering rules.

In solving this second problem states may choose to provide broader filtering guidelines to be implemented by local authorities or individual service providers, resulting in potential differences between the filtering experienced between users in different geographical locations or those using different providers. It is also possible, and has been observed in a number of cases, that a state may deliberately choose to restrict internet services to a greater or lesser extent in certain locations as a result of unrest or disaster.

To understand the technologies employed by states in filtering the internet, and the decisions behind this filtering, we therefore see great interest in studying the extent and nature of filtering at a regional and organisational, rather than national, level. We believe that this will provide a much more sophisticated picture of filtering around the globe, and provide a valuable source of information for internet researchers.

3 Filtering Technologies

The development of the internet was neither carefully planned, nor accurately predicted. It has expanded through the accretion of protocols, services and applications that have been extended and improved far beyond their original purpose. As such, many of the protocols provide opportunities both for filtering technologies, and for attempts to bypass or study those technologies.

There are a number of methods applied to filter internet connections at a national level. These have been usefully categorised by Murdoch and Anderson[7] as follows:

- **TCP/IP Header Filtering:** IP, the Internet Protocol, is the fundamental protocol by which traffic passes across the internet, encoded in IP *packets*. Filtering may occur via inspection of the *header* of an IP packet, which details the numerical address of the packet's destination. Packets may therefore be filtered according to lists of banned destination IP addresses. This method is simple and effective, but difficult to maintain due to the potential for services to change, or to have multiple, IP addresses. This approach may also incur significant "collateral damage" in the case of services that share IP addresses, causing multiple innocent services to be blocked along with the desired target.
- **TCP/IP Content Filtering:** Rather than inspecting the header, a filter may search the content of traffic for banned terms. This is a far more flexible approach to filtering, allowing packets to be blocked only if the include banned keywords or the traffic patterns of particular applications. This approach is also known as *deep packet inspection*, and is known to be employed to some extent by the Chinese national firewall. Deep packet inspection can be partially defeated by using encrypted connections, however filters may choose simply to block all encrypted connections in response, or to block traffic according to identifying traffic signatures that can occur even in encrypted protocols. The most significant limitation of this approach is that inspection of traffic content comes at a significant computational cost.
- **DNS Tampering:** The DNS protocol maps human-readable names to IP addresses on the internet, and is thus critical for most user-focused services such as the web. By altering DNS responses, returning either empty or false results, a filter can simply and cheaply block or redirect requests. This mechanism is simple to employ and maintain, but limits filters to entire websites, and can be relatively easy to bypass for technical users. This approach is employed by, among others, the Turkish state when blocking websites.
- **HTTP Proxy Filtering:** A more sophisticated approach is to pass all internet traffic through an intermediary "proxy" service that fetches and, typically, caches information for users. This is a common internet service that can be used to speed up internet connections and reduce traffic. A suitably enabled proxy can, however, employ sophisticated filtering on certain destinations, whilst leaving other connections alone. This approach can, by ignoring the majority of traffic, be efficient on a national scale while still allowing for detailed filtering similar to TCP/IP content filtering.
- **Other Approaches:** A variety of other means can be taken to regulate content on the internet. States can request that websites are removed from the internet, either by taking down their servers or by removing their names from the global DNS records. A state may also choose not to block a connection entirely, but to slow any connection to that site to unusable levels. At a less technical level, legal and

social constraints can be imposed to may accessing certain services illegal or socially unacceptable.

It has been noted, in [3] that many states begin by employing IP header filtering before moving on to more sophisticated methods as citizens protest the limiting of their connections. In the case of sophisticated national-level connections it is likely that a combination of these methods will be employed in order to meet the various constraints of large-scale filtering.

4 Mapping Filtering

A number of projects exist that provide insight into internet censorship around the world, both from the perspective of learning which sites are filtered and from the more practical approach of bypassing filtering. The most thorough study of global internet filtering is from Deibert et al[3], who present an in-depth global study of tools and techniques of filtering. The related Herdict project[11] allows users to report apparently blocked websites, via a browser plugin, to build up a global map of filtered sites. The Alkasir project[1] combines user-based reporting of blocked content with an anti-censorship tool that attempts to penetrate such filtering.

In bypassing internet filtering, the most well-known technology is the Tor project[4], which allows users to reroute their connections through a global network of volunteer-run anonymising proxy servers. This network, originally designed to preserve the connection-level privacy of users, was found to be an excellent tool for bypassing national filtering and now invests significant resources in supporting this use. Similar tools include Psiphon[8] as well as numerous Virtual Private Network (VPN) servers that allow users to evade national filters. All of these services work in a similar manner: by rerouting a connection through a server located in a different country, the user experiences the internet as if their connection originated in that country. Thus, a user from Saudi Arabia can route their connection through a US computer and bypass all filters run by their state, at the cost of some slowing of their connection and gaining those filters, if any, imposed by the US.

From these examples, we can observe two major possibilities for studying internet filtering. The first is to ask users in a given country to report their experience, as exemplified by the Herdict project; the second is to make use of an available service, such as a Tor node, in that country to experience the filtering directly. Both of these approaches have limitations that we explore in detail below.

Fundamentally, both of the aforementioned approaches suffer from a lack of availability that we see no easy way to avoid. In requesting users to directly report their experiences, Herdict relies on reaching interested and informed users. Tor relies on technically knowledgeable users to set up relays that require both significant resources and a willingness to face potentially serious legal issues[10]. In particular, at time of writing the Tor network does not report any publicly available servers in China[1].

The advantage of using a system such as Tor, Psiphon or VPN services is that they allow a researcher directly to control the flow of traffic. Sites of interest and even specific patterns of traffic can be directly sent and examined. This allows for a much more detailed examination of the technical measures employed on a given network. The approach taken by Herdict, however, cannot currently reproduce this level of sophistication. In the absence of a large network of experienced and technically capable users. user-level reporting only provides that a site appears to be unavailable, without reference to the conditions that cause the unavailability[2].

In order to achieve the fine-grained mapping of filtering that we desire, there are two major points of interest beyond those commonly considered by the most well-known current mapping projects. The first of these is the precise geographical location of a particular computer. The ability to determine the originating country of an IP address is relatively well known, and location to the level of an individual city can be achieved with some accuracy. Recent results[12] have proposed mechanisms that achieve a median accuracy of 690 metres, albeit within the US. This simple extension, we propose, would provide a valuable source of data on the applications of filtering. In many cases it is also possible to determine which organisation has been allocated any particular IP address, to the level of an ISP or major company. Both of these pieces of information can be used to build up a much more detailed view of filtering.

The second point of interest is to study, in detail, the technical nature of the filtering that is imposed on a given connection in a given location. While work has been conducted into specific methods, as in the work of Clayton et al. relating to the Chinese national filter, most large-scale projects appear to be focused more on the existence of filtering rather than the details of its implementation.

4.1 Extending Reporting Approaches

The approach taken by the Herdict project, which relies on volunteer participation to gather data, can be highly

[1] Specifically, there are no announced *exit nodes*, which would be the most feasible way to examine network filtering, reported as located on the Chinese mainland.

[2] The Herdict project does allow a user to express their opinion as to the cause of the blocking, but in the absence of direct experimentation this data has significant limitations.

effective if sufficient volunteers can be found. Herdict currently provides a webpage that attempts to direct a user's browser to load a random potentially-blocked site, and to report their experience. The project also makes available a web browser plugin that allows users to report sites that appear blocked. By focusing on the web browser environment, Herdict greatly reduce the effort required for user participation. The importance of this approach to usability, and the trust implicitly gained through the familiarity of the web browser, should not be overlooked.

This volunteer approach could naturally be extended to the use of more sophisticated tools to detect the presence of filtering automatically and, where possible, test the mechanisms employed. The detection of DNS filtering, IP blocking and even deep packet inspection is often simple enough in itself, particularly when the results of requests can be compared against reference requests made in other countries. It is, however, much more difficult to discover specifics of filtering mechanisms without direct, interactive access to the filtered network connection.

Our own experiments have resulted in a simple application that can detect a number of basic types of filtering, and has been tested on our own servers against deliberately filtered IP ranges and poisoned DNS responses. We make use of the freely-available MaxMind GeoIP database[6] to resolve IP addresses to the city level with a tolerable level of accuracy. At this point, however, our research has been limited, in part due to ethical concerns that we detail below, to proof of concept experiments for which we do not have useful results to present.

A dedicated application to detect and categorise filtering allows for a much higher level of accuracy with respect to the nature of reported filtering. Whether an effective number of users could be persuaded to run such an application is another matter. Therefore, while a standalone tool to map filtering would offer great flexibility, the barrier to entry for volunteers is potentially too high. Browser-based environments, such as JavaScript or Java applets, are likely to strike a useful balance between power and ease for end-users.

It is worth noting the Switzerland tool[9] developed by the Electronic Frontier Foundation, that aims to detect ISP-level filtering of peer-to-peer applications and violations of network neutrality principles. This tool detects many forms of network manipulation applied to an end user's connection, and offers the potential to be adapted for the purposes discussed here.

4.2 Direct Information Sources

As we have seen above, obtaining direct access to filtered connections is desirable for maximum flexibility. This can be achieved through Tor, Psiphon or open VPN services, all of which are specifically design to route traffic for third parties. Although some restrictions may exist on the access to these services, they provide an excellent platform for examining filtering when available. We note above that China does not appear to have any available Tor nodes; many other nations that reportedly engage in significant filtering, which are thus of greatest interest, show similarly low availability of such services. Despite the size and success of the Tor network in achieving its goals of anonymity and anti-censorship, this lack of availability limits its use for mapping global filtering. Where available, however, it is arguably the most powerful tool available to us. Similar services to Tor, including open VPNs, suffer from similar lack of scale to a far greater extent.

It is worth considering, therefore, if common services exist that allow for indirect exploration of filtering. The most obvious of these are DNS servers; these are widely available across the internet, often as an open service available to any users that choose to connect to them, and run a distinctive service that can be easily discovered. Their involvement in one of the major types of filtering, namely DNS poisoning, makes this particular type of filtering trivial to detect across much of the globe – one can simply connect to a DNS server in a locality where filtering is suspected and make DNS requests. If inconsistent results are found then these can be compared against reference requests from a trusted, non-filtered DNS server.

There are a small number of other well-known internet services that can be made to relay connections for a third party, although these are not typically common enough to allow for broad-scale research. Certain IRC servers, open shell access through telnet or SSH, open mail relays and various others offer the potential, however their scarcity and the difficulty of discovery make them a poor avenue of enquiry.

If we consider more legally and ethically questionable methods, there are a number of protocols that have the potential to be "repurposed" for the detection of filtering. Peer-to-peer filesharing networks result in large networks of home PCs running services that are accessible from any computer and that are themselves designed to connect to, and relay for, third parties. While these are unlikely to offer the flexibility of services such as Tor, there are several protocols, such as BitTorrent, that are amenable to this form of information gathering. It is worth highlighting at this point that such deliberate misuse of a service is likely to fall foul of the law in many jurisdictions, whilst simultaneously opening the operator of the service to potential repercussions if their connection is detected attempting to access banned content.

We find it impossible to resist mentioning a possibility open to those willing to throw law and ethics aside

entirely: many modern computer viruses exist solely to create networks of infected, or "zombie", PCs that can be entirely controlled from a central location. These captive systems are typically used, for the benefit of organised criminals, to send high volumes of spam emails or to blackmail organisations through denial-of-service attacks on their networks. Gaining access to such a botnet, some of which have been known to comprise tens of million PCs distributed across the globe[5], would provide an impossibly rich platform for these, and many other, network experiments.

5 Ethics and Legality

While many technical approaches, and challenges, exist for mapping global filtering, there are a number of serious legal and ethical issues to be faced with performing this research.

We have already mentioned that deliberate misuse of a network service may be illegal in many jurisdictions, and such misuse without a user's consent may well be considered unethical. Even when using openly available and general-purposed services, however, there are serious considerations when attempting to access blocked content via a third party.

In many situations, a user is unlikely to face repercussions for being seen to be attempting to access blocked content. The scale of internet use, even in smaller countries with low internet penetration rates, is simply too high for there to be serious policing of users who request filtered content. It is likely that, in the vast majority of cases, such attempts may not be logged at all. However, users in specific contexts may be put at risk.

The legality of attempting to access filtered content is also a concern. Many nations have somewhat loosely-defined computer crime laws, and often prefer to prosecute crimes involving computers under existing legislation rather than through creation of new laws. The legal status of attempting to access blocked content, however, and of attempting to bypass such blocks is not something a researcher can afford to ignore.

From the point of view of a researcher, these concerns are exacerbated by two factors: the concentrated attempts to access filtered content that is caused by a detection tool, and the wide variety of laws and social conventions that exist around the globe.

By their nature, the filtering detection mechanisms that we have discussed, and any that we can feasibly imagine, detect filtering by attempting to access filtered content: by requesting websites or IP addresses that are known, or are believed or likely, to be banned. As we have stated above, it would be largely impractical for a state to take note of every blocking action taken by their filter. It is possible, however, that sufficiently high-volume requests for banned content may be considered worthy of further action. A user innocently aiding a researcher in mapping their national filter, resulting in their computer suddenly attempting to connect to all forms of banned content, may find themselves under very unwelcome scrutiny.

It is also of great concern that a researcher not cause a user to unwittingly break the law with respect to the content that they direct a user to access. With the wide global variance in law, great care would have to be taken that a censorship tool not attempt to access content that was directly illegal. Pornography, particularly with respect to those under the local age of legal consent, *lèse majesté* and insults to religion are all sensitive issues that vary widely between cultures.

Volunteers that participate in research of this nature by running a filtering detection tool must do so having been fully informed as to the nature of the tool and the potential risks involved. From this perspective there is a significant added burden on the researcher to state to the participant, who may well not have any significant level of technical expertise, what the tool will do and what particular risks they run.

In the case of relay services, such as Tor or Psiphon, consideration must be given to the safety and security of the user operating the service. Due to their nature these services are frequently abused, and operators of such services must be prepared to defend their operation of the service. The Tor Project, in particular, invests significant efforts in education both for operators and for users. This does not, however, reduce the burden on a researcher taking advantage of such a service to ensure that they do not harm or endanger the operator through their actions.

6 Conclusions

We propose that it is in general false to consider internet filtering as an homogeneous phenomenon across a country, and that the practicalities of implementing a filtering regime are likely to result in geographical and organisational differentiation between the filtering experienced by users. We believe that the study of these differences are of great interest in understanding both the technologies and the motivations behind filtering, and propose a number of mechanisms that could be employed to gain this understanding.

However despite the existence of a number of technological and social avenues to aid in this research, we see a number of serious legal and ethical concerns that must be thoroughly considered in order to undertake broad-scale research of this nature. Beyond the more obvious pitfalls of misusing third-party services in an attempt to conduct this research, there are more subtle issues. The necessity

of attempting to access blocked content, and the legality and ethics of performing this via a third-party volunteer or service operator are all worthy of serious discussion by researchers in this field.

Despite these concerns, and the technical hurdles to gaining a detailed picture of global internet filtering, we consider that research into this subject presents a number of interesting problems, and can provide insight into the development of the internet and its ongoing social and political role both the national and international level.

References

[1] W. Al-Saqaf. Alkasir for Mapping and Circumventing Cyber-Censorship. `http://www.alkasir.com/`. Accessed May 8th, 2011.

[2] R. Clayton, S. J. Murdoch, and R. N. M. Watson. Ignoring the great firewall of china. In *In 6th Workshop on Privacy Enhancing Technologies*. Springer, 2006.

[3] R. J. Deibert, J. G. Palfrey, R. Rohozinski, and J. Zittrain. *Access Denied: The Practice and Policy of Global Internet Filtering (Information Revolution and Global Politics)*. MIT Press, 2008.

[4] R. Dingledine, N. Mathewson, and P. Syverson. Tor: The Second-Generation Onion Router. In *Proceedings of the 13th USENIX Security Symposium*, August 2004.

[5] Matt Thompson. Mariposa Botnet Analysis. Technical report, Defence Intelligence, 2009.

[6] MaxMind Inc. MaxMind GeoIP City Database. `http://www.maxmind.com/app/city`. Accessed May 8th, 2011.

[7] S. Murdoch and R. Anderson. Tools and Technology of Internet Filtering. In R. Deibert, editor, *Access Denied: The Practice and Policy of Global Internet Filtering (Information Revolution and Global Politics Series)*, chapter 3, pages 57–72. MIT Press, 2 edition, Dec. 2008.

[8] Psiphon Inc. The Psiphon Project. `http://www.psiphon.ca/`. Accessed May 8th, 2011.

[9] The Electronic Frontier Foundation. Switzerland Network Testing Tool. `https://www.eff.org/testyourisp/switzerland`. Accessed May 8th, 2011.

[10] The Electronic Frontier Foundation. Tor Project Legal FAQ. `https://torproject.org/eff/tor-legal-faq.html.en`. Accessed May 8th, 2011.

[11] The Herdict Project. The Herdict Project. `http://www.herdict.org/`. Accessed May 8th, 2011.

[12] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang. Towards Street-Level Client-Independent IP Geolocation. In *NSDI*. USENIX Association, 2011.

3. Website histories and historiographies

M. Dougherty, E.T. Meyer, C. Madsen, C. van den Heuvel, A. Thomas and S. Wyatt (2010). *Researcher Engagement with Web Archives: State of the Art.* London: JISC

JISC

Researcher Engagement with Web Archives
State of the Art

August 2010

Meghan Dougherty
Eric T. Meyer
Christine Madsen
Charles van den Heuvel
Arthur Thomas
Sally Wyatt

Acknowledgements

This report was funded by JISC, the Joint Information Systems Committee, from April to August 2010. The project was a partnership between the Oxford Internet Institute at the University of Oxford in the United Kingdom (http://www.oii.ox.ac.uk) and the Virtual Knowledge Studio at Maastricht University in the Netherlands (http://virtualknowledgestudio.nl/). Questions or queries about this report may be directed to:

Dr. Eric T. Meyer, Project Director
Oxford Internet Institute, University of Oxford
1 St Giles, Oxford, OX1 3JS, United Kingdom
Tel: +44 (0) 1865 287210
Email: eric.meyer@oii.ox.ac.uk

Neil Grindley, Programme Manager
JISC, Digital Preservation & Records Management
1st Floor Brettenham House (South), 5 Lancaster Place, London, WC2E 7EN, United Kingdom
Tel: +44 (0) 203 006 6059
Email: n.grindley@jisc.ac.uk

Please cite this report as:
Dougherty, M., Meyer, E.T., Madsen, C., van den Heuvel, C., Thomas, A., Wyatt, S. (2010). *Researcher Engagement with Web Archives: State of the Art.* London: JISC

Table of Contents

Boxed highlights in bold

Page is intentionally blank

4

Executive Summary

In this report, we summarize the state of the art of web archiving in relationship to researchers and research needs. This is a different focus than much of the earlier work in this area, including the JISC PoWR report which focused on institutional strategies for archiving web resources (JISC, 2008). It is important to note that this report focuses on the uses and needs of individual researchers. Research groups are also important, as some of the challenges that face individual researchers can quickly spiral into deeply complex tangles when dealing with collaboratories. For instance, national selection policies and national copyright rules can stand in the way of international projects, even if there are sound academic reasons to pursue international collaboration. While these issues are addressed here when appropriate, the bulk of the report focuses on individual researchers and institutions.

One of the main issues underlying this report is that there is still a gap between the *potential* community of researchers who have good reason to engage with creating, using, analysing and sharing web archives, and the *actual* (generally still small) community of researchers currently doing so. In this report, we identify some of the main reasons for archiving web pages, web sites, web domains, and the web in general. Beyond the fact that the web is allowing for the constant creation and distribution of huge volumes of information, it is also a valuable resource for understanding human behaviour and communication in the late 20th and early 21st centuries. To really reach the potential of web archives as objects of research, however, it is necessary to begin to take web archiving much more seriously as an important element of any research programme involving web resources.

A number of approaches are possible within this realm, and in the report we identify the differences in scope and scale of web archives, and present examples of how web archives can be used to address a number of research questions. Another key theme throughout are the challenges that still face researchers who wish to engage seriously with web archives as an object of research.

This report also makes a number of recommendations regarding developing additional capacity for web archiving and for research into web archives. These recommendations are grouped into three themes: building community, building tools & resources, and building practices.

Building Community

- Encourage the creation of communities that increase the accessibility and usability of web archiving tools
- Sharing tools and sharing web archives should become the norm
- New multidisciplinary approaches should be encouraged
- Privacy and property issues should be made more understandable
- Local instances of collections should feed into meta-collections to maximize the value of consortia

Building Tools & Resources

- There are two related and connected streams of support required to build infrastructure and to support the needs of individuals to archive

- Tools should be sharable and easy for researchers and librarians to implement
- Efforts should be made to diversify tool and interface development beyond preservation and into use
- Workflow tools should be used to orchestrate collections of standardised building blocks
- Tools should be developed that are able to execute query searches over multiple web archives
- Shared typologies, or vocabularies, of metadata need to be developed
- Standards, protocols and methods of quality control are need for interoperability, but not at the cost of flexibility
- Multiple access points into archives are needed to support administrative, descriptive, and conceptual access to web archives
- Shared archives of web archives need to be developed

Building Practices

- Web archiving needs to be integrated into the practices of institutions
- Additional training to understand the structure of web content will help researchers understand how to make use of archival web content in their research
- The possibilities of web archives should be communicated to a much broader research community
- Researchers need help to better match available tools to their needs
- Funding postgraduate students in areas that require web archives and providing them with the necessary skills will yield growth in this area in the long term
- Support for experimentation with web archives is vital for innovation
- Mentorship of new researchers is important for instilling the importance of archiving the web materials that researchers are increasing using as objects of study
- Measuring the impact of shared web archives is good practice

These recommendations are described more fully in the body of the report. We hope that these recommendations will be taken seriously, and that they will inspire researchers to see the advantages of working with web archives for research purposes.

Why archive the web?

The World Wide Web provides unprecedented access to information on virtually every known topic, and is a constantly growing and evolving information source that continues to develop as users and consumers of information and technology become increasingly knowledgeable. Information distributed on the web encompasses a vast array of the activities and artefacts of humanity. The *New York Times* reported in 2006 that the extent of human knowledge is summarized in "32 million books[1], 750 million articles and essays, 25 million songs, 500 million images, 500,000 movies, 3 million videos, TV shows and short films, and 100 billion public web pages" (Kelly, 2006). At the time, it was estimated that the sum of knowledge generated throughout human history could be contained in 50 petabytes (10^{15}) of storage space. The Internet, however, is increasing the rate at which textual, visual, and audio information is being produced and shared. By 2008, Google reported that their systems had found 1 trillion (10^{12}) unique URLs on the web at once (Alpert & Hajaj, 2008). The Internet Archive, which is a collection of historical copies of web pages representing the most complete source of the history of the Internet to date, currently contains 3 petabytes (10^{15}) of data, and is growing at a rate of 100 terabytes (10^{12}) of archived data *each month*. The sheer quantity of data appearing on the web represents a rapid expansion in human knowledge, including a comprehensive record of information production and social interaction over time. As Dr. Kirsten Foot put it when interviewed for this report:

> *At this point in our social material history, the extent of intertwining between online and offline phenomena is so thorough…that if we don't capture the online phenomena in at least the same rigor that we archive newspapers and other kinds of artefacts of cultural significance, we will have nothing to study retrospectively. There is a significant collective consciousness that is heading to a dark ages where we aren't writing anything down, in fact we are writing lots down on the web, but then we are writing over what we just wrote. It will be very hard for future scholars even in five years, ten years to understand what kinds of political and social and cultural moments or phenomena retrospectively without the key aspects of the web.* (Foot, personal communication)

What is a web archive?

In interviews for this research, stakeholders suggested the following answers to the question *"What is a web archive?"*

- A set of web objects that have been collected and verified with a particular purpose or goal in mind (where the goal could be to collect everything). What makes it an archive is the intentionality, collecting process, and then some level of verification
- Artefacts that are born digital, created on the web for the web, and are interesting for curatorial or analytical reasons
- A web archive is any offline storage of web content, created either manually or with an automation tool by an individual or group of people
- An accessible archive is one that has an interface that allows users to see objects in the archive
- A national collection representing website materials of interest to a nation
- A domain collection (e.g. ac.uk)
- A specialist collection based on one or more related specialist subjects
- A records management solution for business and legal purposes (one that treats a website as an organisational record)
- A collection designed to provide content of value to researchers (once one knows who the user community is)
- A collection of data that could be text-mined, or analysed statistically, or in other ways, to give interesting results
- A history of website design and application usage

[1] Although more recently, the Google Book project estimated the total number of books at a much higher count of approximately 130 million (Taycher, 2010). Estimates of this sort from any source are bound to be inaccurate in one way or another, if one wishes to take into account all languages at all times, but they can give one a sense of the scale at which one is operating when dealing with this much information.

This constant change is one of the web's greatest advantages to its end users: consumers of information are able to find the most up-to-date news and information at the touch of their fingertips. Yet this changing nature is also one of its chief frustrations as a data source: pages disappear, content is re-edited, comments are deleted, and wikis are vandalized. Without printed volumes, the history contained within the content of web pages is often lost. Researchers, archivists, librarians, students, citizens and corporations seeking knowledge or records previously but no longer available on the Internet are often at a loss, and those needing to know the history of content on the web are likely to struggle to get any significant information. Over the past fifteen years, most of the content of the web has disappeared as it is replaced by new pages and new content. There is, in fact, rapid turnover: several studies found that within a given week 35-40% of web pages changed their content[2] (Cho & Garcia-Molina, 2000; Fetterly, Manasse, Najork, & Wiener, 2004), and that this change is even more rapid when looking at subjects visiting dynamic pages such as news sites. For instance, in one study, 69% of web sites changed when revisited after a day or more (Weinreich, Obendorf, Herder, & Mayer, 2008), and another found that certain dynamic information is likely to change more frequently than once an hour (Adar, Teevan, Dumais, & Elsas, 2009). Pages are updated and refreshed continuously, but older versions are rarely archived by content producers. Web pages decay over time, and on average have a half-life of little more than two years, depending on the type of content (Koehler, 2004). This evolution and decay of content further results in a phenomenon referred to as 'link rot' as relationships and connections between data are lost over time (Taylor & Hudson, 2000).

In addition to this ever-changing content, the Internet and the web continue to show a dizzying pace of technological evolution – new multimedia types, new ways of displaying content (e.g. on mobile, rather than PC-based, platforms), increasing use of executable content such as JavaScript -- all pose new challenges for the web archive community. Worse still, much of the web's content (up to 90% by some estimates) is increasingly hidden behind forms-based query interfaces, and the actual content is held in databases which are inaccessible to crawlers; the development of methods to allow these "deep Web" contents to be collected poses another major challenge. Other, even more fundamental changes, such as the growing pervasiveness of social media such as Facebook and Twitter, among many others, point to a potential sharp decline in the relative importance of the "traditional" web, as is pointed out by in an article (Anderson & Wolff, 2010) which is engendering considerable controversy as this report goes to press. In this new world, there is a risk that open content, protocols and interface behaviours will be replaced by closed systems, content and interactions which are absolutely invisible to traditional archiving practices.

[2] Although the rate of change varied considerably by domain: .com pages changed much more quickly than .edu pages, for instance.

Web archives as research objects

Starting in the mid-1990s, researchers began partnering with librarians, as well as working on their own, to create archives of web objects that could be queried to draw generalizations about a variety of topics in the humanities, social, and physical sciences. Research using these methods range from studies about politics on the web (Foot & Schneider, 2006), to explorations of the web presence of different cultures (Franklin, 2005), to linguistic studies (McEnery & Wilson, 2001). These types of inquiry have contributed to shaping the descriptive, methodological, and theoretical bases of scholarship centred on web archives.

Web archives case: Election Web Spheres

Foot & Schneider's work (2006) was one of the earliest innovative research projects to use purpose built web archives as a means of answering a research question. In building their archive of web campaigning in the 2000, 2002, and 2004 elections in the United States, they conceptualized their objects of study as a *web sphere*. They define web sphere as *"a set of dynamically defined, digital resources spanning multiple web sites deemed relevant or related to a central event, concept, or theme…enabling analysis of communicative actions and relations between web producers and users developmentally over time"* (p. 27). By building an archived collection of websites produced by a variety of political actors during election campaigns, Foot & Schneider were able to better understand campaign strategies, tensions within campaigns, and more generally how technology is influencing the practice of political campaigning.

In the early 2000s, as web archives became more accessible and more widely known, a number of researchers and librarians worldwide began to investigate the potential and the limits of such a resource as a complement to exploration of the live, or currently active, web. Advocates of web archiving draw on methods in the relatively new area of digital cultural heritage to harness the quantity and variety of data available, in the hopes of advancing the potential for studying new genres such as blogs, web forums, and collections of emails. It is also possible using these methods to observe change in the content of the web as it takes place (Foot & Schneider, 2006; Kilgarriff & Grefenstette, 2003). Some sceptics, however, have questioned the trustworthiness of archives collected by researchers, arguing that control over sources and long-term replicability and stability in the building of such collections should be better defined (Brügger, 2005).

While many debates about the potential uses of web archives still remain at both a theoretical and practical level, web archiving is increasingly accepted by most cultural heritage institutions as an important complement to more traditional forms of collection development. Many researchers, too, have moved forward to explore the building and the resulting value of such archived web collections empirically. The development of social actions have been explored with the use of web archives (Foot & Schneider, 2006), object-oriented approaches in web historiography have been compared to topic and event oriented approaches (Dougherty, Schneider, & Jones, 2010, Forthcoming; Schneider & Foot, 2010), the ethical and legal impacts of saving artefacts from a highly volatile semi-public cultural space have been addressed (Dougherty, Foot, & Schneider, 2010). Within this body of work, technical and methodological approaches vary substantially: from the use of Google queries to derive artefacts from a web sphere to capture and archive (Schneider & Foot, 2004), and expert derived sets of artefacts to archive from the entirety of the web, to more targeted approaches delineating very specific sets of carefully defined web objects such as pages or sites (Brügger, 2005), and downloading quick-and-dirty specialized corpora for evaluating the language of the web (see, e.g., the papers in Baroni & Bernardini, 2006). While this work has provided interesting tools and new insights, none so far has succeeded in coalescing and making available to the larger research and heritage community an infrastructure that combines the advantages of the web in terms of inclusion and access with the advantages of traditional methods in archive research in terms of stability and control.

This report presents an overview of the current state of web archiving, including the diversity of practices as they are evident in a variety of inquiry modes, attempts at standardization, and the

loose web archiving infrastructure that has emerged to support e-research and e-heritage. The focus of this project, though, is on the current state of researcher engagement with web archives – how are researchers currently making use of web archives and what sort of technical and policy infrastructures will they need in the future in order to facilitate their work?

State of the art

Stewardship of cultural heritage is a story of loss and reconstruction. Artefacts deteriorate, or become otherwise corrupted, and stewards of the cultural heritage those artefacts represent - whether they be scholars, curators, archivists, or interested amateurs - feel a responsibility to reconstruct not only the artefacts, but often the meaning the artefact holds for interpreting our past. This holds true for stewardship of digital cultural heritage as well, not only in the construction of narratives about our past on the web, but also for the way practices are developed for handling the web artefacts that help researchers to construct those narratives.

The World Wide Web is now largely recognized as an essential access point for cultural, historical, and scientific information. Nonetheless, it is still a highly fragmented environment that is often changing, always evolving, and often disappearing. In recognition of this problem, several groups are now successfully archiving large portions or selected segments of the web. Through these activities, they aim to create an archival record of web culture or of contemporary culture as manifest on the web. This record is intended "to resemble a digital library" from which historians, curators and scholars can draw data to support their research (Lyman & Varian, 2003).

Library and information science have been developing practices for collection and archive development for decades that have come to dominate web archiving. In some ways, the practices and standards of this discipline are a good fit because they are extensively developed and ready to handle the content management and delivery systems required by web archives. Further, they offer an existing policy framework for the collection of contemporary cultural materials. However, there are consequences to relying heavily on libraries and archives to deal with web archives. As European Archive director Julien Masanès points out:

> *It is a utopia to hope that a small number of librarians will replace the publisher's filter at the scale of the global Web. Even if they have a long tradition in selecting content, they have done this in a much more structured environment that was also several orders of magnitude smaller in size. Although this is still possible and useful for well-defined communities and limited goals…, applying this as a global mechanism for Web archiving is not realistic. But the fact that manual selection of content does not scale to the Web size is not a reason for rejecting Web archiving in general. It is just a good reason to reconsider the issue of selection and quality in this environment.* (Masanès, 2006, p. 4)

While library and information science norms have been the basis for many of the developments in web archiving policy and infrastructure, the resulting focus on collection development and preservation of artefacts has often been done with little regard to the question of how the web archives will eventually be used. Viewing the web archive as a collection of documents and bibliographic records is an efficient approach to storing and preserving the web. Whether it is flexible enough to accommodate the uses that researchers will want to put web archives to is another question. This has set up a point of contention between librarians and information scientists who would like to build widely valuable and accessible collections, and humanities and social science researchers who would like to develop web archiving as a method for understanding digital cultural heritage or web historiography. The two perspectives are not diametrically opposed, but there are certainly points of contention that are derived from differently held philosophical undercurrents that motivate each (Dougherty, 2007). Librarians and archivists are inclined (and trained) to build

Web archives case: Iranian Elections

In June 2009, Iran participated in its tenth democratic presidential election. As the results were tallied, allegations of electoral fraud were voiced and protests mounted. Most of the anti-Ahmadinejad actions known as the Green Movement were coordinated online. According to one researcher, "*Immediately after the election there were lots of digital materials online – campaign materials, online activism, video clips, citizen journalism, and a lot of really good stuff in Facebook. Essentially there was a huge amount of Iranian cultural artefacts online. Nothing like this had ever happened before.*"

A group of researchers distributed around the world attempted to archive these materials. They had two motivations: *"The first is selfish, really. That these would make a great research archive at some point. Something to go back to. The second is political. Through this archive it would be easier to reproduce the narrative of the Green Movement."*

Unfortunately, the project ran into technical problems due to a lack of easy to use tools and server space. Without an immediate source of funding to pay for commercial services, the researchers were not able to save most of this material. This underscores the need to have better and more accessible methods to archive and save materials related to unfolding events that are now being lost.

collections that will last for centuries, even 'forever' as is the mandate for some institutions. Researchers are interested in first building or collecting something that can help them answer their current research questions or even design new ones. The longevity of the data beyond their own career or even beyond a project, for researches, is generally of secondary importance.

Collaboration and partnership is a complex issue. During an interview for this project, Kirsten Foot reflected on the issues that arise in institutional collaborations. She identified the various partners who are interested in partnering around web archiving: national libraries in the US, Europe, Australia and Asia; and museums and archives that are recognizing the value of born digital objects for their collections. She mentioned universities as institutions that are taking an interest, but quickly explained that they do not seem to have yet developed any discernible strategy for collecting born digital materials. In describing her experience as an academic entering into a multi-institution partnership, she explained that even as an individual researcher, there needs to be some university-level commitment to support inter-institutional web archiving activity. She mentions that there are legal considerations as well as technical considerations that serve as a foundation. The more complicated issues are the detailed protocols about what curation consists of, and the basis for collection development. These issues are approached very differently by social researchers versus librarians versus archivists. Foot said, "It is important to really thrash through those [differences] and work out a protocol." Technical questions of storage, quality assurance, and capture are also issues to be negotiated. When Foot was asked to elaborate on "thrashing through differences" to determine protocol, she said that she learned the hard way that these are necessary conversations. People from different types of disciplines have different concepts in mind even when using the same terms and it is important to bring those differences to the forefront when collaborating. She was particular about the definition of what it means to be "systematic" and the level of rationale or criteria needed to complete a given project successfully. There are different practices from professional communities and domain expertise. Thoughtful agreement around these issues are increasingly important as proprietary technology can obscure how we access and capture web materials - different search engine algorithms will lead to different results much the same as different search strategies will surface different results. These differences have deep epistemological and disciplinary roots.

As a result, large libraries and archives continue with their efforts to build large multi-purpose web archives that further institutional missions, while researchers - either on their own, or partnering with archivists - develop their own archives for use in their research. Archives cannot justify allocating resources to project-specific archives, but researchers cannot always find useful materials for their work in the large multi-purpose archives being built by archivists. The core tools for creating basic web archives are now widely in use, but there is no underlying infrastructure in place to support the research into these archives.

Consequently, web archiving is currently in a state of flux where boundaries around traditional roles of researchers and stewards are blurring. Stewards are seeking out researchers to learn their needs. Researchers are building their own collections and seeking the expertise of archivists to sustain those collections. These types of collaboration are resulting in the need to experiment with different approaches that are guided by multiple motivating principles. Web archives created by a social scientist will inevitably differ from those created by a librarian, or by a linguist. The tools needed to make the archives usable to each group will vary as well. Each practitioner is motivated by a different mission, be it institutional, methodological, or epistemological. Diverse approaches to web archiving are resulting from this experimentation and are increasingly leading to conversation and collaboration across fields to develop inclusive practices.

In addition to this older community, who have been principally interested in the content of the web, we now see the appearance of a relatively new community – the Web Scientists – who are interested in the web itself as a technological artefact and object of study (Hendler, Shadbolt, Berners-Lee, & Weitzner, 2008). There are many fascinating issues about the network structure of the web, and the ways in which that structure evolves over time, which have intrinsic interest, as well as telling us a good deal about how human beings use communication technologies in innovative ways to interact and collaborate in the creation of new cultural artefacts. The interests of these new students of the web are not necessarily best served by the library and information science approach, and future developments in web archiving will need to take these new requirements into account.

A diversity of approaches distinguished by purpose

Each of these diverse approaches to archiving web objects develops from certain modes or styles of inquiry. Researchers in the social sciences and humanities are guided in their practices by methodological concerns and specific research questions when approaching the web and attempting to stabilize objects of analysis there. Cultural heritage professionals are guided by institutional mission statements and clientele.

The greatest contention among these professionals is based on fundamental differences in how we understand the world, and how we determine what things are. These epistemological and ontological beliefs provide a driving force for activities of collection, documentation, classification, and are eventually filtered through to defining points of access. Divergences in the beliefs that underscore the development of these activities can entrench practices later, so much so that change becomes quite difficult. Support for experimentation in practices is vital at these early stages as the field is still being defined.

The following categorization of web archiving projects is not comprehensive, but shows the evolution of multiplying practices and tools. Each step problematizes the previous one and creates its own new path while respecting the value of the previous. Each new path proposes its own set of practices as an addition to add value to previous collection practices.

Broad collections: diverse future uses

Both scholars and cultural heritage institutions recognize the need and value of preserving content on the web (e.g., Arms, Adkins, Ammen, & Hayes, 2001; Burner, 1997; Day, 2003; Foot & Schneider, 2002, 2006; Hodge, 2000; Kahle, 1997; Kahle, Prelinger, & Jackson, 2001; Lyman & Kahle, 1998; Masanès, 2002, 2005, 2006; Schneider & Foot, 2002, 2004, 2005), and have launched efforts to archive web content.

In 1997, Brewster Kahle published a short article in *Scientific American* entitled, “Preserving the Internet” in which he described his Internet Archive project that would attempt to do just that. This

was not the first or only mention of web archiving at the time – the Finnish EVA project was launched in the same year and Australia's PANDORA archive was launched in 1996 – but it marked the beginning of the most ambitious effort to preserve artefacts from the web to date. The Internet Archive (IA)[3] takes a whole-domain approach, with the goal to preserve the entire content of the global web. This approach builds a comprehensive collection of websites and online resources using harvesters to automatically retrieve artefacts in broad sweeps of the web. Other broad sweepers of the web include the European Archive[4], while projects such as the Swedish Kulturarw3[5] and the UK Web Archive[6] limit their domain to national web spaces. The Preserving Access to Digital Information (PADI) page[7] maintains a list of national web archiving programmes.

Broad scale collecting strategies result in very large collections of archived sites, but generally with little documentation or metadata about the objects. Due to the sheer scale of IA's crawls, for example, only machine readable data is collected. This results in archives that are difficult to navigate as archived sites can only be retrieved via URL, as in the IA's case via their *Wayback Machine* interface[8]. This interface problem is exacerbated by the fact that the quality and reliability of these archives often do not meet the standards of completeness and replicability required of researchers in the humanities and social sciences. However, new tools from IA such as *Archive-It*[9] are being developed to allow for more focused collections with advanced features such as search, a feature which is not yet technically feasible across the entire *Wayback* collection (see box). As will be discussed further below, access, interfaces, and selection policies are all creating challenges for those wishing to broaden the use and re-use of web archives.

In addition to building collections, large-scale projects such as the Internet Archive and the European Archive have parallel missions to make their collections usable and accessible to the public and to researchers. For the former this is focused on universal accessibility--that is, to the widest audience possible. To date, their efforts have been primarily focused on providing "native replay" of individual archived sites and pages. With this capability now well established they are turning their attentions to providing new ways for researchers to use their archive (primarily through the development of new APIs). The European Archive, too, is focused on building tools that allow researchers to engage with their archives, for example to run analytics or to perform linguistic analysis. Through their *Living Knowledge* project[10] their goal is "goal is to bring a

Web archives challenge: Search

In 2009, the Internet Archive ran a pilot in providing full-text searchability, making the first five years of their archive (1996-2000) available for searching. The search ranking mechanisms available at that time were not adequate, however, and the search results were full of spam. To date, there is still no reliable full text search tool for web archives and although several groups are currently working on the problem, it remains one of the greatest obstacles to providing archives usable for a wide variety of researchers.

Search in general is still not able to adequately work with items in digital archives to the standard many researchers desire. For instance, with regard to the *New York Times* digital archive of news content dating back to 1851: *"We can say, 'show me all the articles about Barack Obama,' but we don't have a database that can tell us when he was born, or how many books he wrote… Such a resource will not only help the research community move the needle for our company but for any company with a large-scale data-management problem."* (Evan Sandhaus, New York Times Research and Development Labs, quoted in Simonite, 2010)

[3] http://www.archive.org/
[4] http://www.europarchive.org/
[5] http://www.kb.se/english/find/internet/websites/
[6] http://www.webarchive.org.uk/ukwa/
[7] http://www.nla.gov.au/padi/topics/92.html
[8] http://www.archive.org/web/web.php
[9] http://www.archive-it.org/
[10] http://livingknowledge.europarchive.org/index.php

new quality into search and knowledge management technology for more concise, complete and contextualised search results."

A number of studies have established the Internet Archive as a valid tool for research in the social sciences. In particular, scholars have used the Internet Archive's *Wayback Machine* as a tool for estimating the age of a website, the frequency of updates, and for evaluating and coding the content within sites (Brock, 2005; Thelwall & Vaughan, 2004; Veronin, 2002). Further, Murphy, Hashim & O'Connor (2008) validated measures of age and frequency of updating against third-party data, illustrating the overall strength and reliability of these measures as research tools. Thus, there is support for the use of data from the Internet Archive as attributes and characteristics in research studies. The *Wayback Machine* can additionally be used as an evolutionary research tool to track the development of technology over time, for instance, to track changes in content over time. Chu, Leung, Van Hui & Cheung (2007) conducted a longitudinal study of e-commerce websites, using the Wayback machine to track the development of site content. Similarly, Hackett & Parmanto (2005) used the *Wayback Machine* to analyze changes in website design in response to technological advances over time. Efforts along these lines include the *Memento* project[11], which adds a time dimension to the HTTP protocol to better integrate the current and past web, and the *Yahoo Time Explorer*[12] which is being developed to build timelines from searches in news archives. A number of scholars have conducted historical research using data from the Internet Archive. This previous work has clearly established the utility of data from the Internet Archive as a source of research data. Yet large-scale studies using this source are hampered by the size of the database, the structure of the data itself and the complexity of linkages between sites (Murphy, et al., 2008). To date, they have used tools that have been time-intensive to develop, that are custom-made for particular topics and therefore not widely usable, and that have encountered many other difficulties and limitations.

Web archives case: The Twitter archive

In 2010, the U.S. Library of Congress announced that Twitter had given its entire archive of public tweets to the Library for preservation and to make it available for research use. According to the FAQ for the collection, *"Twitter is part of the historical record of communication, news reporting, and social trends – all of which complement the Library's existing cultural heritage collections. It is a direct record of important events such as the 2008 U.S. presidential election or the "Green Revolution" in Iran. It also serves as a news feed with minute-by-minute headlines from major news sources such as Reuters, The Wall Street Journal and The New York Times. At the same time, it is a platform for citizen journalism with many significant events being first reported by eyewitnesses. The Library of Congress collections include items such as the very first telegram ever sent, by telegraph inventor Samuel F.B. Morse, oral histories from veterans and ordinary citizens, and many other firsthand accounts of history. These collections and others have left behind glimpses of the lives of ordinary people, thereby enriching knowledge of the context of public events recorded in government documents and newspapers. Individually tweets might seem insignificant, but viewed in the aggregate, they can be a resource for future generations to understand life in the 21st century."* (Raymond, 2010)

Directed collections: flexible, immediate uses by individuals and institutions

Other web archiving approaches are selective, thematic, deposit-based or a combination of these approaches. Selective approaches identify web artefacts to collect by specifying certain inclusion criteria such as a theme, by quality or significance, or through identifying specific intervals at which to take impressions or snapshots of web artefacts. This type of selection at the harvesting level is employed by Australia's PANDORA[13] project, which collects selected Australian online publications deemed to be of national significance and long-term research value. The U.S. Library of Congress employs a thematic approach with its Library of Congress Web Archives[14] (originally called the

[11] http://www.mementoweb.org/
[12] http://fbmya01.barcelonamedia.org:8080/future/
[13] http://pandora.nla.gov.au/
[14] http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html

Web archives methods: Collecting

Ed Pinsent of the University of London Computer Centre provided the following general steps he uses in creating a web archive.

1. Discover that the target site exists - for example by checking jisc.ac.uk and other sources to see what new projects have started up, whether they have websites, and determine if they fit the scope of the collection.
2. Seek permission from the website owner to make a copy. I use a form and mail merge to do this. The project manager is usually regarded as the owner. If and when consent is given, enter the details of their Institution into Web Curator Tool, thus creating a permissions record.
3. Create a target entity in Web Curator Tool and link it to the permissions record.
4. Set harvest in motion.
5. QA the results. If necessary, change parameters of the harvest for future gathers (e.g. add or remove filters), or "prune" the gather to remove material we don't need
6. Submit the harvest to the archive.

MINERVA project) by selecting artefacts that fit a specific theme. Its *United States Election 2000 Web Archive*[15] (also done in 2002[16], 2004[17] and 2006[18]), *September 11, 2001 Web Archive*[19], *Iraq War, 2003 Web Archive*[20], and *Papal Transition 2005 Web Archive*[21], for instance, used these themes to guide selection. Deposit-based projects, such as projects at the National Library of the Netherlands (Koninklijke Bibliotheek)[22], rely on voluntary deposits of web artefacts. The National Library of the Netherlands is also working with experts on collection strategies within specific identified humanities-related topic areas.

Several projects aimed at preserving national digital cultural heritage employ a combination of these approaches. France and Denmark combine comprehensive sweeps with targeted selective and thematic collection strategies in an effort to guarantee good coverage of certain highly valuable portions of web artefacts within a larger broader sweep of content. The Digital Archives for Chinese Studies[23] (DACHS) with branches at the University of Heidelberg and Leiden University, and Virtual Remote Control[24] (VRC) at Cornell University represent a 'by discipline' approach to web archiving that is popular among research institutes and universities. The British Library takes a similar hybrid approach, focusing on building discrete collections of "websites with research value that are representative of British social history and cultural heritage".[25] Several of Harvard University's libraries[26] are working on very narrow but deep collections, known to fall within the existing collection scope of the library, such as *Blogs: Capturing Women's Voices*[27] and the *Constitutional Revision in Japan Research Project*[28]. At both the British Library and Harvard University Library, archiving of web content is being integrated with standard collection development practices. These approaches provide varying degrees of nuance in all the processes of web archiving. Libraries, archives and large cultural heritage institutes can have broader objectives and thus employ broader practices in their approaches.

[15] http://lcweb2.loc.gov/diglib/lcwa/html/elec2000/elec2000-overview.html
[16] http://lcweb2.loc.gov/diglib/lcwa/html/elec2002/elec2002-overview.html
[17] http://lcweb2.loc.gov/diglib/lcwa/html/elec2004/elec2004-overview.html
[18] http://lcweb2.loc.gov/diglib/lcwa/html/elec2006/elec2006-overview.html
[19] http://lcweb2.loc.gov/diglib/lcwa/html/sept11/sept11-overview.html
[20] http://lcweb2.loc.gov/diglib/lcwa/html/iraq/iraq-overview.html
[21] http://lcweb2.loc.gov/diglib/lcwa/html/papal/papal-overview.html
[22] http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/index-en.html
[23] http://www.sino.uni-heidelberg.de/dachs/
[24] http://handle.library.cornell.edu/VRC/
[25] http://www.bl.uk/aboutus/stratpolprog/digi/webarch/index.html
[26] http://wax.lib.harvard.edu/collections/home.do
[27] http://wax.lib.harvard.edu/collections/collection.do?coll=61&lang=eng
[28] http://wax.lib.harvard.edu/collections/collection.do?coll=101&lang=eng

Narrow collections: known, immediate uses by researchers

Oftentimes, a researcher's systematic approach and sometimes-narrow topical scope guides the creation of narrow collections in web archiving. In these researcher-led cases, the selection of artefacts is driven by the boundaries of the research project for which a sampling scheme has been developed. Categorization follows coding strategies informed by prior inquiry into the field and developed to address certain concepts tested in the project. These collections are limited in size and scope. They typically focus on an initial list of seed URLs, or the contents of one website, and contain frequent (sometimes hourly or more) captures of artefacts resulting in very full, but limited collections. There are sound methodological reasons for creating a web archive; as project interviewee Dr. Steve Schneider, of SUNY Institute of Technology in the United States, puts it:

> *I think it is not possible to study social phenomenon on the web, especially in an ongoing/ developmental sense, at any medium-to-large scale and with any hope of replicability, without archiving material. So the benefit is that archives make it possible to do the quality social science research that is, in a sense, competitive methodologically with large-scale survey research. My thoughts are that the way we approach web sphere analysis has the opportunity to bring the methodological sophistication (including the ability for others to replicate our research) of public opinion research to the study of online social phenomena.* (Schneider, personal communication)

Some of the more technical aspects of web archiving such as indexing and curation are similar in both widely sweeping archiving schemes and narrowly bound scholarly web archiving. Scholarly web archiving is a focused development of a collection following narrowly defined collection strategies, while individually produced web archives are designed to be a source of data generally for one particular project. Researchers develop these collections on their own, and in conjunction with larger institutions with better resources, however the extent to which these collections can be described as archives varies. Individual collections with no public access and no claims to longevity can hardly be called archives, but this does not reduce their potential value to the research community. They merely lack infrastructural support.

Traditional collection development can follow similar individual procedures, but without a specific research project in mind. Collection development is an ongoing task that follows set policies, but is a different act than the sampling procedures in a research project that tend to guide scholarly web archiving collection development. Fundamentally, the archivist aims to develop a collection that may be widely used for any number of known and

Web archives case: Immigration web storm

Interviewee Dr. Kirsten Foot of the University of Washington in the USA, recently compiled a web archive of what she calls a web storm, which she defines as *"a flurry of productive activity that happens on the web in unpredictably predictable ways. You don't know when it will happen, but there will be bursts of generative activity on the web in which many actors are producing material about the phenomenon."* As a social science researcher studying social phenomena, she often has an eye out for unanticipated web storms that fit into other arguments that she is interested in theoretically.

This particular case involved the Yahoo News site and the recent immigration debate taking place through links to Photoshopped images of a particular cartoon character. Foot noticed that Yahoo News was aggregating reports from other news sources reporting the photo manipulation as political commentary, but were presenting the content on their site in a guarded way. She noticed that Yahoo News was providing access to the politically and emotionally charged images through a link to an outside server and providing their own disclaimer in text surrounding the link on their page. Foot saw this as an example of strategic coproduction, and began capturing snapshots of the Yahoo pages, and its target links. Once she noticed what was happening and identified the event as an example of a concept she works with, she explains that she knew there were certain aspects of the phenomenon that she needed to capture on the pages that were linked together in this event. She needed to capture evidence of the particular dimensions she saw as relevant to the concept she was observing: who was hosting the images, who was pointing to them, what the various pages the portal provided were, the various levels that it took to navigate to the page with the image, etc.

unknown users for different purposes. The individual researcher's web archive is a set of data collected to support a specific inquiry that may be re-purposed for another project later.

One particular form of narrow archive is the “idiosyncratic archive" (of the type described in the accompanying box on the Immigration web storm, and also discussed in Dougherty, et al., 2010, Forthcoming). The web is often the site of "unpredictably predictable" activity, a type of activity that is not necessarily tied to the definition of a web storm presented in the box, but is an undercurrent concept that drives all activity and retrospective analysis on the web. It is degrees of this "unpredictable predictability" that illustrates the difference between different styles of narrow archives. So, for instance, there is a difference between the unanticipated web storm such as the Yahoo News example, and an event such as the recent case of Steven Slater, the JetBlue employee who dramatically quit his job by exiting the plane by the emergency chute. Though the time scale is still short, there is a moment - no matter how short - between the event of Slater's dramatic exit from his flight attendant job and the coming web storm for which you can predict what online actors will produce a short-lived burst of related content. In contract, Foot's Yahoo News web storm brews more slowly from events originating on the web.

Individual or research-led web archiving usually includes rich metadata, interpretation, and representation. These are technical and analytical steps that actively engage the user or reader. These steps go beyond other methods of web archiving by invoking research methodology designed to answer specific questions, rather than simply to catalogue and preserve information. This added data makes the resulting web archive particularly useful to the researcher or archivist who created it. The risk, of course, is that without an ontological understanding of those methods and collection development policies, these collections may be difficult for other researchers to use.

Page is intentionally blank

18

Web archiving: A developing field

No matter what the approach, web archiving is a complicated process involving many steps to selecting and acquiring objects to archive, and also determining solutions for storage, documentation and access (Arms, et al., 2001; Foot & Schneider, 2006; Hodge, 2000; Masanès, 2002, 2005, 2006). While there are many planning strategies and policy formats, collection development policy making takes knowledge, experience, and intuition, but it also aims to reflect the needs and interests of the collection's community of users (Johnson, 2009). Like most others defining the scope of web archiving, Julien Masanès, director of the European Archive, does so by developing practices already established in librarianship and archiving. The practices he describes cover collection policy development and collection building, but fall short of delving deeply into the other areas of access, categorization, interpretation, and representation.

Masanès (2002, 2006) points out that applying traditional strategies for collection management to web collections is difficult, a point also noted in interviews done for this report. At this point in time, there are few comparable collections against which to evaluate the completeness of web collection strategies. The inconsistent publishing procedures and formats on the web, and the connectedness of the medium both create a need for a different and more open approach to discovery and dynamic selection.

A cultural environment exists with this technical media environment that is also fluid. It is this type of context and environment that allow us to recognize artefacts and their uses. We use these environments to create genres into which we can categorize artefacts to account for their meaning and usefulness (Innis, 1951; Levinson, 1997). The preservation of a digital document is tied to its production. Every time you read a digital artefact, it must be reproduced and reconstructed entirely – it must be rendered in a human-readable format. With born-digital documents, preservation is no longer an artefact-centric problem. The integrity of the media environment surrounding and supporting the artefact must be preserved in addition to the integrity of the artefact.

Some web archivists discuss preservation, but their discussion of what they call preservation also addresses issues of selection and capture (Day, 2003). The rate of resource decay on the Internet, the rate of change in web tools and standards, and the continuing development of the Semantic Web, where information is given well-defined meaning so machines can recognize, understand and process it accordingly, are all issues to address when developing a collection policy, and they will influence choices of how to collect, when to collect and what to collect. None of these considerations address how to preserve the artefacts once they are collected, nor do they address how to preserve the varied uses and interpretations the artefacts took on during their active time in the cultural world (that active time may overlap with the time spent in the archive once collected). As one of the librarians interviewed for this report said, “innovation in web technologies is both a challenge and a threat. We are always catching up.” The social life of the artefact, defined by the uses to which it was put to produce new knowledge and the interpretation it was assigned by different users at different times are additional avenues in which to collect metadata to preserve not only the object itself, but some meaning about the object so its cultural value can be revisited and evaluated as it changes over time.

Tools for building and using web archives

Each individual tool for personal desktop archiving has a different set of goals and so different design elements. Simply archiving sites you’ve visited during a particular research setting does not always meet the needs of the researcher. Often, the researcher does not know what metadata elements are missing, or what indexing elements are missing from a certain archiving tool until it is too late. Social science researchers find themselves with archives that are full of redundancies that need to

be cleaned out, missing seeming redundancies that actually show significant change, or contain a mess of archived sites with no logic of how the individual objects can be related to one another. Personal desktop archiving tools are designed from a "basic needs" perspective. The designer's assumption is that the user wants to save websites to view later. The next level of design complication is that the user may want to know the exact click stream followed when navigating a site. Neither of these assumptions touches upon the complexity of what a social science researcher thinks it means to save a website for retrospective study, or to archive a click stream for analysis. In an interview for this project discussed above, Kirsten Foot described problems in inter-institutional collaborations in web archiving; she explained that people from different disciplines have different concepts in mind even when we use the same terms. These differences can surface in the design of personal desktop archiving tools. It is important to surface those differences early. It is important for researchers to be very clear about their research goals, and thorough about what metrics they will need to reach their goals. It is also important to develop some tools that are not multi-purpose. Not all tools need to be accessible to the casual user, and special research tools can be designed to meet the higher level needs of the researcher.

Web archives case: Personal Facebook archives

Interviewee Frank McCown recently led a research project that produced a Facebook archiving add-on for Firefox (ArchiveFacebook, available at https://addons.mozilla.org/en-US/firefox/addon/13993/). The add-on is a tool that users can install and run by themselves to produce an offline, fully navigable archive of their Facebook account. This kind of individual-use tool reinforces the current trend of creating fail-safes and living-wills for online identity profiles. This is a specific perspective on archiving the web, which has potential to find a large popular following of users for this type of tool, but does not necessarily help researchers create, access, or analyze web material retrospectively. More often than not, this is the type of archiving tool that is leading the current state of web archiving tool development.

The overarching challenge is not recognizing the importance of archiving web content in general, or more specifically a particular metric, concept, or method until it is far too late. Certain questions cannot be answered, certain concepts cannot be illustrated, certain methods cannot be used if measures are not set up to be indexed as an archive is built, and studies cannot be replicated if the ephemeral digital primary materials aren't archived. Even if the researcher was clever or lucky enough to capture all the different data required, there are two additional challenges. The first is finding software that suits the researcher's needs, and as a corollary finding a researcher who is capable of evaluating the available tools to match their needs. It is hard to find and figure out which archiving software is going to be useful and user friendly for the kind of use in practice that that individual has. The second challenge in use is organization. The structure of the objects you collect matters. Foot described seeing eager researcher-archivists collect strategically, only to find that their collection was inaccessible due to tremendous redundancies, and structural chaos in the archive: "Many of the tools available are simply not robust enough" (Foot, personal communication).

In 2003, twelve institutions including the Internet Archive and eleven national libraries (Australia, Canada, Denmark, Finland, France, Iceland, Italy, Norway, Sweden, British Library, US Library of Congress) formed an international collaboration focused on Internet Archiving. The International Internet Preservation Consortium[29] (IIPC) focuses on creating tools and standards for web archiving as well as providing support and advocacy for its members. The IIPC open source tools now comprise the standard package used by most cultural institutions engaged in web archiving. These include the Heritrix crawler, the Web Curator Tool (WCT) for collecting, NutchWAX for indexing, and the Wayback interface for access.[30]

[29] http://www.netpreserve.org
[30] http://www.netpreserve.org/software/downloads.php

As described by one of the web archiving project managers from a national library, *"we have just now gotten good at what we do – downloading copies of static text from the web."* Nearly twenty years after the introduction of the first web browser, we have finally made progress in capturing and preserving some of the earliest web documents. This pace, when contrasted with the speed of innovation on the web, will shortly become a significant challenge facing web archiving communities. New protocols such as iPhone apps[31] are being introduced and popularized across the web. New mobile devices are providing new ways of looking at the same data, increasing the difficulty of providing "native replay" of archived materials. This begs the question: if some web based content appears differently in a web browser and on an iPhone, does a web archive need to capture that? If so, where does it stop – do archives need to capture all competing versions as well? Even the introduction of embedded metadata in to existing protocols (RDF, for example), which could help with indexing and access to pages within archives, provide new challenges. For example, if the public content of a web page does not change, but its tags do, does this represent a new version? As Wendy Gogel of the Harvard University Library commented to us for this project, *"In the future the sheer number of formats is going to be overwhelming and the problem is not the capture of these, but in being able to preserve display of them."*

Future challenges and opportunities for using web archives

Differing inquiry modes for web archives

As outlined above, there are several current approaches to building web archives - some arising from frustration with existing resources, some developing from institutional mission statements, and all developing from limited understanding of the end-users' needs.

Temporary *ad-hoc* practices that are developed to circumvent obstacles were discussed in several pilot interviews conducted in autumn 2008 by project partner VKS (Dougherty, 2008), and in new interviews for this project conducted in summer 2010 with a range of researchers and librarians engaged in some variation of web archiving. All respondents are facing similar sets of obstacles despite their approach. The ways in which these obstacles are handled determines, among many things, the character of the resulting archive, the limitations of use as set by access points to the resulting archive, and ultimately the perceived value the resulting web archive offers to different communities of researchers.

Common obstacles

The common thread through conversations among researchers and archivists using and building web archives is that researcher-users all want different aspects of the same things. Firstly, they want ***stabilized web objects*** that can be reliably studied and cited. They want to be able to ***clearly define*** what that stabilized archived object represents in reference to the live web. They want to have access to archived ***representations of the most fine-grained features*** of web objects in order to suit their research needs. Most of all, they want to ***work with those objects, enriching and annotating them*** on whatever level is appropriate for their analysis.

In terms of the archive itself, three things are clear: an archive must be trustworthy, long-lasting, and reliable. These are fundamental elements of any archive; and these elements need to be extended to bolster web archiving processes as they develop. Researchers and small-scale libraries are increasingly seeking the help of large established archives to meet these standards. Resources for downloading, archiving, and serving archived objects are often too costly to implement for individual researchers and small libraries. Even with the availability of software tools such as those provided by

[31] Short for applications. Generally small, inexpensive, single purpose programs designed to pull data for instant display.

IIPC[32], the limited access to human and technical resources and expertise is often cited as the main obstacle for small libraries and researchers wanting to participate more actively in the web archiving community. Even with free technical resources available, small operations have limited human resources to run and maintain it. These parties recognize that the criteria for legitimately calling their collection a valuable archive that serves a research purpose in the future are often beyond their reach. They are seeking to collaborate with larger archive institutions to share resources and expertise.

As small collections seek collaborative opportunities, they move forward, doing their best to meet standards of a legitimate archive, and face the next set of obstacles: **access**. Often, access obstacles are also fundamentally a problem with lack of resources. In the case of access, not only is there a lack of labour resources, there is also a lack of technical infrastructure to support that work.

For these archives to find value in the world or research, they need to have multiple access points: **administrative**, **descriptive**, and **contextual**. These types of access points are experimented with and employed in myriad ways in different archives. Again, there are few shared practices, and no standards across archives. Shared practices exist only as a coincidence if two archives use the same harvesting software, or object-rendering software. Further, these three access points are even described differently using disciplinary language that is not shared between researchers and archivists, or even between researchers in different fields. According to an archiving engineer we interviewed at one of the national archives, quality assurance still requires extensive manual work as few automated tools exist, exacerbating the problem since manual steps are more difficult to duplicate unless they are meticulously documented. Each description of how an archivist or researcher would like to have access to an archive contains elements of these three strata, but none share a common language.

Administrative access enables a user (or archivist) to examine an artefact and determine exactly what it is (when it was archived, with what software, from what organization, including what file types, etc). This type of access is imperative for the structure of the archive itself. Administrative data enables an archivist to rebuild an archive after a data crash, for example. Administrative access is also valuable for content comparison across archives or across archived objects.

Descriptive access is basic catalogue access to artefacts in an archive. Basic cataloguing information makes artefacts findable (Morville, 2005). This descriptive metadata is equivalent to the information in a library that would help a user find one book among many on a shelf. The metadata answers the question, "What is it?" for every object in the archive.

Contextual access places artefacts in a thickly described and purposeful context. Contextual access does not place an artefact in its original context; rather it makes an artefact findable via its relationship to other objects in a research project. Contextual access has been experimented with in several collections; two of the most notable are DACHS[33] and the former Politicalweb.info, which somewhat ironically is no longer available online.[34] Users enter an archive and view archived artefacts via the research of another. Archived artefacts, in this sense, can be seen as a collection of objects to which a research project refers. This metadata answers the question, "What is it *about*?" for any object in the archive, and this question can be answered differently many times over depending on the perspective and purpose of the researcher-user.

[32] http://www.netpreserve.org/software/downloads.php
[33] http://leiden.dachs-archive.org/
[34] See the Wayback Machine to view archived versions of the site: http://web.archive.org/web/*/Politicalweb.info. The domain name currently points to an advertising site.

A related issue is contextualization in the form of **annotation**. Hanzo Archive, for instance, has created tools that allow individual annotations to artefacts within a web archive. The need for collective annotation of web archives, however, is only recently being acknowledged (Dougherty, 2007; van den Heuvel, 2009). By allowing collective annotation of web archive objects, researchers can build up additional levels of data to enhance our collective memory (van den Heuvel, 2009, pp. 282-283).

Making collections valuable and re-usable for researchers does not have to involve a large-scale effort to build platforms and maintenance-heavy metadata structures for search. Researchers are eager to become involved. They are eager to use the collections that exist and to create their own. One of the primary obstacles to the involvement of researchers in early phases of web archiving projects, though, is a lack of user-friendly **interfaces**. While the tools for capturing and documenting websites are now in place, there are still not sound, intuitive interfaces for interacting with web archives, particularly at the scale of the larger archives. Currently, in order to access an archived version of a website in most collections, users must know the URL of that site. Searchability of web archives is still minimal. If a site no longer exists it is therefore buried, unless the user remembers the site's URL or finds it via an archived hyperlink from another site. The scale of web archives alone presents challenges for providing usable and intuitive interfaces and the temporal and versioning aspects of web archives compound these challenges further.

Ultimately, and fundamentally, there is an epistemological conundrum about ***what constitutes a document*** in a web archive. This conundrum is at the heart of the disconnect in understanding access points across collections. This is a fundamental and persistent discussion in web archiving. Web archiving is a creative process. For each "archived object" we have an impression that it is an approximate representation of what was on the live web. We cannot verify its veracity with the live web. As web technology advances, the notion of the "live web" becomes less and less static - web objects are served differently to different people. Our archived impressions are often incomplete. At times they are loose representations of the objects we wish to capture. At worst, they are snapshots of one instantiation of a dynamic object that may look, in detail on the live web, very different to the many individual users viewing simultaneously. As web historiography develops as a field, it will no doubt develop different methodological approaches to dealing with this epistemological problem.

The problems described above are only a sample. The challenge to web archivists and those building tools to support their use is to build sustainable systems that can weather the coming epistemological rifts in methodology that will arise as the field grows. This epistemological conundrum begins at the earliest stages in the web archiving planning process and continues through to research, and takes hold in the subsequent re-use of previously collected archives. It is this epistemological conundrum that makes many current web archives difficult to re-use.

There are so many different valuable research-oriented approaches to an archive. These approaches, or methods of search and retrieval, are often reduced to tools that represent the few most basic methods (e.g., full-text search without lexical indexing, or specific item search and retrieval based on strict metadata points). Other richer and more powerful search strategies focus less on searching, but rather more on temporary sorting. These methods are experimented with largely in research settings where researchers are working alongside archivists and librarians to build robust collections. As collections are being built, and as researchers are using them, they can add value themselves. Their additions, in turn, make the collections valuable and re-usable for future users. Each new slice through the collection by each new researcher adds to the robustness and re-usability of the collection. Each new way of searching through the data may not be valuable in and of itself to the next researcher who uses the collection, but it may spark interest and creativity.

While the IIPC toolset mentioned above is becoming more heavily embedded into the web archiving practices of institutions, the creation of web archives by individual researchers and end-users is still an elusive and often *ad hoc* practice. The goal among all involved in web archiving should be to turn existing, institution-level technology and resources into accessible and stable services that any user in any discipline can share, adapt, and repurpose. This statement is made with the current culture of personalization, Web 2.0 and 3.0 technologies, and the self-directed and democratic characteristic of web culture itself in mind. The focus on the ability to re-use, repurpose, and personalize research resources mirrors this trend in web culture, and shifts focus to the users' role in developing, not only making use of, humanities and social science research resources for the web. Perhaps users are a valuable resource for web archiving documentation that is yet to be tapped.

Web archives challenge: Knowing the users

One of the big challenges for the organizations who host web archive collections is that it is difficult for them to know how, or even if, their collections are being used. According to Ed Pinsent of the University of London Computer Centre, *"Not much is known about the users of the JISC web archive. The public do not feed back to the JISC or to ULCC as to what use they make of the collections. The only evidence we have is statistical evidence, generated from the log files by the British Library. But this simply records visits to the UK Web Archive and doesn't tell us anything about who these people are, why they are visiting, what they expect to find when they get there, what they take away with them, or whether they have experienced any degree of satisfaction."* One possible approach is to apply impact tools, such as the JISC-funded *Toolkit for the Impact of Digitised Scholarly Resources* (TIDSR) to web archives, just as they have been to other types of digital collections.

Copyright also remains an obstacle at several levels. In terms of access to archives, in some cases researchers have to go to the library building where the web archive is housed to consult the resource and a result of copy right issues. This obviously makes accessing the electronic resource inconvenient for researchers not located near the archive. In addition, copyright issues regarding harvesting potentially copyrighted content into a new web archive can be difficult to navigate, and the legal issues are not at all clear in this area (Knutson, 2009; Patel, 2007). Also international differences in copyright can stand in the way of international research collaborations and projects. These issues are important to clarify so that researchers and institutions will have greater confidence that their collection building and research can be carried out without infringing the rights of others.

The role of the user

Too little is known about users' behaviours in relation to web archives. Most archiving institutions therefore rely on semi-hypothetical use cases to refine and expand their usability and interfaces. One particularly detailed study was conducted at the National Library of the Netherlands (Ras & van Bussel, 2007). This structured experiment, run similarly to a task-oriented usability study, evaluated user comfort level with search and access tools and attempted to determine user satisfaction with archive contents. Several use-scenarios were posited. Few native users have been studied to date, and reports of these studies remain unpublished works-in-progress. We do not have much to draw on when speculating about users in web archives. However, those who are developing their own web archives for directed and narrow research purposes can provide some insight about how they use their archives to produce knowledge in their field.

Web archives challenge: Chickens and eggs

"We tell them what's possible and we want them to tell us what's useful" – Helen Hockx-Yu, The British Library

From the perspective of libraries and large archiving efforts, working with users presents a "chicken and egg" scenario. Usable web archives are just emerging, such as the one released by the British Library in February 2010, and institutions are just now beginning to understand what is possible. Researchers are being asked how they might like to use web archives, but until recently have not known what is possible. Several user-focused initiatives are being led by institutions such as the British Library and the European Archive, and the results of these studies will be pivotal in understanding what will come next.

Despite the reciprocal relationship between the development of research in the humanities and social sciences, there is tension between archiving practices used by researchers and their subsequent access requirements and archiving practices and perceived access requirements in heritage institutions. Each recognizes value in archiving artefacts from the web, but each has followed different paths to develop web archiving practices with a special focus on characteristics most relevant to their immediate environment. More and more, each community is beginning to understand the particular sets of expertise each community can offer to the cause; and members from each community are beginning to understand the value in partnering to achieve the shared goal of stabilizing and preserving artefacts from the web. Ultimately one aim in these efforts is to develop or identify key elements to support the emergence of an infrastructure for web archiving activities for research in the humanities and social sciences.

Researchers, technology developers, and cultural heritage institutions need to work together in order to build this infrastructure with an acute awareness of preservation, accessibility, and interpretation in all their different permutations in the diverse sets of practice. Keeping a diverse set of users in mind, preservation, accessibility, and interpretation can come to be more inclusive and representative of expert and lay-expert views together. To date, most institutions actively archiving web objects focus on some limiting definition to bound, or stabilize, web objects as documents, and place emphasis on an efficient system for generating metadata to enable smooth transitions between archived web objects and other documents. This is highly influenced by traditional library practices. However, the ephemeral and dynamic nature of web objects questions traditional notions of the document. The unclear definitions of web objects lends itself to experimentation with practices in documentation, notably the inclusion of broad annotating activity by diverse users to describe web artefacts and add value to archives for researchers in the humanities, sciences, and social sciences.

Page is intentionally blank

Recommendations

To draw together existing web archiving technologies into an infrastructure that will support e-research and e-heritage, we must foster community and create an abundance of tools and resources that are usable by a variety of users.

Building Community

- Web archiving resources remain largely inaccessible; **the creation of communities that increase the accessibility and usability of web archiving tools should be encouraged** so researchers and librarians can have a common space to share best practices and develop standards.

- Researchers and librarians are often re-building, re-stabilizing, and re-conceptualizing web archiving for each new project undertaken; **sharing tools and sharing resulting web archives for research should become the norm** for both researchers and librarians. These shared resources should enable participants to share archives in a flexible way that meets both institutional missions and individual research needs. The idea of **virtual collections made up by on-demand integration of information from multiple physical collections** would allow users to create thematic collections with much less effort than at present.

- Contributions are being made on a practice-level, a structural level, and a theoretical-conceptual level, but are disconnected in the scholarly literature and professional practice communities; **new approaches should enable connections across disciplines and professions** to encourage web archiving to grow as a flexible field.

- **Privacy and property issues should be made more understandable** in the web archiving space. Many people working in e-research and e-heritage are limited in their use of tools, sharing of practices, and sharing of results due to international law, institutional missions, publication restrictions, and often individual personal preferences in protecting data and methods. Much more powerful tools (based, e.g. on Digital Rights Management technologies) are needed to allow archivists to collect, and users to navigate ethically and legally through these minefields, and to publish with some confidence that they will not run into future liabilities.

- International collaboration remains an important, albeit costly, element to the continued development of tools, resources and standards. In parallel with these continuing international approaches, **local instances off these collaborative outputs need to be created that can feed back into community meta-collections in order to maximize consortial efforts.** The development of such tools, as exemplified by the *Archives Hub*[35], will help avoid duplication of collection efforts and serve to give users a much richer overview of what content may be available, and where.

Building Tools & Resources

- In balancing between the top-down needs of institutions and the bottom-up needs of researchers, there need to be **two related streams of support: one for infrastructure and one for individual archiving**. Crucial to this two-pronged approach, however, is **building a way to connect the two**.

- Technical obstacles are keeping many researchers and librarians out of the emerging web archiving community. **Tools should be both sharable and easy for researchers and librarians to**

[35] http://archiveshub.ac.uk/

implement. Tested solutions to struggles in technology should be easy to find and execute. Usability in installation and use should be a primary concern in future tool development in order to attract more researchers to working with web archives.

- Current web archiving efforts rely heavily on the same set of existing tools, but few of these are specifically focused on extracting data from archives in a manner that enables serious research. **Efforts should be made to diversify the development of tools and interfaces beyond preservation and into use.** These tools, as mentioned above, should be shared widely as a normal practice. Ideally, such tools, should aim to blur the distinction between live and archived content, and also allow much more powerful visualisations of the structure of complex collections, and their changes over time.

- An approach based on modern software engineering practices (e.g. the establishment of collections of Web services or other programmatic interfaces) would allow the current, rather monolithic tools to be replaced by **collections of standardised building blocks whose activities could be orchestrated by workflow tools**.

- Researchers and librarians struggle to use the archived web in research and heritage because there are currently so few ways to parse the information gathered in a crawl; it should become commonplace for researchers in varied fields to have **tools to execute query searches over multiple web archives** to find themes in content that go beyond the results provided in a full-text and 'presence or absence' search.

- Standards for metadata vary by researcher, field, and tools; it should become commonplace to call up a **typology, or vocabulary, of metadata particular to the line of inquiry** that inspired the original query. Metadata should be relational and movable for the needs of the audience at hand. The development of new metadata standards outside the library community, such as the Resource Description Framework (RDF)[36] and Linked Data[37] conventions, point out new ways in which rich and flexible metadata can be used not only for retrieval but also for linking together documents and data sets from different sources, in different formats.

- **Development of standards, protocols, and methods of quality control will help to make web archives more interoperable.** However, the diverse needs of researchers need to be taken into account, so standards must be built that have the flexibility to accommodate innovative uses.

- **For these archives to find value in the world or research, they need to have multiple access points: administrative, descriptive, and contextual.** Administrative access allows for structural integrity, descriptive access allows one to understand the catalogue of contents in an archive, and contextual access places the artefacts within the archive in a thickly described and purposeful context.

- While considerable effort has been put into developing data archives, there has been considerably less commitment to building places to store and share web archives. **Resources need to be developed that allow researchers to deposit and publish their web archives** that are searchable, with organized metadata, and with transparency in the collection criteria, period of capture, and other technical details so that researchers will know what they are dealing with when accessing and re-using the web archives. The adoption of cloud storage technologies may allow the stretched resources of the Web archive community greater economies of scale, leading to an eventual change from "collecting the needles" (assuming that archivists know

[36] See, e.g. http://en.wikipedia.org/wiki/Resource_Description_Framework
[37] http://www.linkeddata.org

ahead of time what needles their users may interested in) to "collecting the haystack," thereby giving much more freedom to the users to ask unanticipated questions and navigate in unanticipated ways.

Building Practices

- **Web archiving practices need to be integrated into the daily practices of cultural institutions.** Libraries have existing policies and practices for collection development that can and should be expanded to encompass web-based materials.

- To understand the possibilities for research uses of web archives, researchers need to have some understanding of how websites are built and how they behave. **Basic training in the area of web content design can lead to a better understanding of how to capture, archive, use, and interpret content from websites.** If they need to make important decisions based on what is stored in a web archive (for example, a lawyer trying to prove a web page contained a certain image on a certain date), they are certainly going to need to be trained about the basics of HTML, web browsers, CSS, JavaScript, web crawling, and possibly other factors. Or they are going to need an intermediary who can explain to them what they need to know in layman's terms.

- **The possibilities of web archives should be communicated to a much broader research community.** A number of examples of potential uses are given below.

- There need to be **better resources for researchers to be able to match available tools to their research needs**. It is currently too difficult to find and understand which archiving software is going to be useful and user friendly for any given practical use.

- Postgraduate training is an excellent way to engage new researchers with new methods and objects of research. **Funding students to look at questions which require the use of web archives, and providing them with the skills to help create the next generation of tools**, has the potential for enabling considerable growth in web archiving for research and for encouraging creative uses of web archives.

- **Support for experimentation in practices is vital at these early stages as the field is still being defined.** Creative new uses may emerge from unexpected quarters, and providing support for these unexpected innovations is crucial.

- **Mentorship of new researchers is necessary to instil the importance of archiving the materials one studies as one studies them.** We need to encourage our undergraduate, post-graduate, and post-doctoral researchers to follow best practices in archiving the web materials they are studying, to build these practices, and also develop the resources that will be available to researchers for further study.

- Funding bodies such as JISC are increasingly recommending that holders of digital collections measure the impact those collections have on various audiences. **Using methods such as those in the JISC-funded *Toolkit for the Impact of Digitised Scholarly Resources* (TIDSR)[38] to measure and enhance the impact of collections of web archives is good practice.**

[38] http://microsites.oii.ox.ac.uk/tidsr/. Members of the project team creating this report were also responsible for developing the TIDSR resource. Other approaches to understanding audiences and enhancing impact would also be appropriate.

Sample potential uses of web archives

There are many potential uses of web archives. To get researchers thinking about the possibilities of web archives, the following ideas represent examples of the types of questions that either could be answered with current tools and methods, or that could be answered with the development of new tools and methods. Of course, this list is suggestive, not exhaustive; many other areas are possible.

Web archives challenge: Imagining the uses

Niels Brügger, an Associate Professor in the Department of Information and Media Studies at Aarhus University, Denmark, was interviewed and discussed how researchers need to imagine the potential uses of web archives:

I guess I would like to see as many people as possible doing history using archiving stuff. To use this kind of material. Have people using it, asking questions of the archive, developing a little what they can do. In one of my texts, I distinguish between five strata that you can focus on in the web. There's the web as such, then the web sphere (clusters of web sites), you can have a web site, a web page, and the web element. And I would like to see studies in all these strata in a way. I am not advocating that we should only do web site research - they are all important and they are all context of each other. I would like every historical study as possible on all these 5 strata. For example, can one imagine, as Kirsten [Foot] and Steve [Schneider] do the history of a web sphere – that's what they do with their presidential elections. Web sphere analysis. Web site history/analysis is what I try to do web pages, that could be, for instance, we heard Megan Sapnar talk about. The design. Web elements - there was a person at a recent meeting who did not give a presentation, but she is working with ads on the web. Banner ads - that history. That would be the history of the element. I hope that people start doing all these things.

If you want to study the web sphere, the links are crucial. And the web site, the outgoing links might be important. Maybe the targets aren't important, but you want to know that it was a link. Studying pages, there you probably find it necessary to have all the elements on the page. I think that would be important for an archive. And the elements, and again, if you study streaming media, the use of video throughout the history of the web, it is important that the archive have those elements. So each of the strata might pose different demands.

Humanities Scholars

There are many sites on the web covering historical topics. Take the two World Wars, for instance: many sites contain personal testimony and copies of original sources such as photographs, letters and official documentation (Meyer, Carpenter, & Middleton, 2009). It may be that members of the public who might not think to approach an archive or library with their own story or personal mementos would be more likely to mount details or copies of their mementos online, which has happened with the *Great War Archive* project at Oxford[39]. Often people have responded to sites which invite those who lived through these events to contribute their memories, and people may be more willing to do so in the privacy of their homes via the internet or through a local event. One of the attractions of these sites to historians, therefore, might be that they offer previously unavailable or untapped primary sources. Other humanities scholars such as those interested in the web as corpus for linguistics are natural potential users of web archives (Hundt, Nesselhauf, & Biewer, 2007; Kilgarriff & Grefenstette, 2003).

Sample questions include:

- How many photographic sources are available on the web for a particular historical event or time period? If places are tagged in these photos, is it possible to reconstruct a virtual panorama of the place or time in question?
- How many personal reminiscences are available across different websites? Do the same people, events, and places in these reminiscences occur in different accounts? When were the reminiscences written, by whom, and for whom? Tools to find, analyse, and view

[39] http://www.oucs.ox.ac.uk/ww1lit/gwa

related documents that refer the same people, places and events would greatly expand possibilities here, such as those being developed in the *Cultures of Knowledge*[40] project.

- Do alternative sources and accounts that are on the web challenge the current historiography? To what extent have these sources been overlooked by traditional historians? Have the kinds of historical sources and documents available on the web changed over time? Does this tell us anything about either history, or about the practices of historians and those members of the public interested in history?

- Are there topics that are of broad interest to the amateur historians and the public that appear frequently on the web, but are largely absent from the traditional historical discourse? Have amateurs developed interesting areas, or found novel ways to present historical information? Are the documents on the web any more or less reliable that other sources?

- Using the huge amount of language available on the web, what can we understand about language change? How is written language changing to reflect new technology? What languages are rising or falling in dominance on the web?

Internet Researchers and Social Scientists

Scholars who are interested in the Internet and its impact on society are clear candidates to become users of web archives. Most research in the area of Internet studies has been cross-sectional, based on data collected at a particular point in time. Now that the web has been around for the better part of 20 years, there is a need to start understanding changes over time on the Internet. Some examples of the kinds of questions one might ask using web archives:

- How has the growth of online news varied country by country over time? Given the claims made by some newspapers that the Internet is killing newspapers, is there historical evidence for any relationship between the depth of online content and a newspaper's financial solvency? How does the contribution of online news affect democratic debate?

- Where has discussion of climate change been most active? How has this changed over time? Is it possible to map the geographical spread and the topics covered in the debate to the geography of climate change effects and attempts at mitigation?

- What kind of predictive indicators about future potential financial crises can be uncovered through the retrospective and real-time data mining of the web?

- Using hyperlink analysis of the structure of the web to understand the social processes around topics and events. While some hyperlinking behaviour is formal or institutional (e.g. government agencies linking to one another as authoritative sources of information or providers of services), a lot of hyperlinking activity is more informal, reflecting the grassroots networking of bloggers, NGOs, special interest and advocacy groups. How can changes in linking over time help us to understand the role of informal communication as part of the feedback loops influencing developing issues? What hyperlinking behaviour is exhibited by these actors, and how can this be related to social science models of collective action?

[40] http://www.history.ox.ac.uk/cofk/. While this project is not focused on web archives *per se*, it is developing methods for linking between similar references in letters that would be applicable to a researcher looking for these sorts of links in web documents.

- How has the visibility of topics changed over time? Do websites that fall within certain network clusters during one point in time ever move to different clusters, or do they remain stable? For those that move, what separates them from the more stable parts of the network?

- Can we develop better tools to analyse web archives statistically? How many sites exist on certain topics? How has this changed over time? What languages are the pages in? Are there clusters around which pages are created, or have they grown steadily over time? Are certain topics more interlinked than others? Can websites be divided into categories that we can uncover using cluster analysis? Can we compare sites by statistics such as the average size of the website in different categories, average number of links, amount of non-textual data (photographs, images, etc), age of content between updates, frequency of updates, type of interface (static versus dynamic, for instance).

- Can we visualize web archive data using methods such as tag clouds of the website titles or keywords or of all the textual content on the website? Can we do linguistic analysis of the terms and words used, and sub-divide the sites into different clusters linguistically?

- With regard to user creation of content, much of the hype around the web, particularly web 2.0, is that users are creating more and more content. This shift from the passive consumption of media content about the world to active participation in the generation of content is clearly happening in areas such as the creation of YouTube videos. Can we measure anything about this non-professional content creation? For instance, what proportions of the collections reside in different domains (.edu, .ac, commercial domain, yahoo website, etc.)? Can we determine which kinds are more likely professional versus amateur creators? If so, can we distinguish between them using the measures in the sections above (links, types of data, age of content, age of site, size of site, etc.)?

Many other questions are possible, as these are just a few to get people thinking about the possibilities for web archives as research objects.

Conclusion

Building community and tools with the features listed above will result in a shift in perspective in e-research and e-heritage that:

- Recognizes and enables the reciprocal relationship between e-research and e-heritage on and about the web;

- Fosters historical and heritage work as well as contemporary research on and about the web in the humanities, sciences, and social sciences; and

- Establishes a domain of distributed repositories, services, and expertise.

Participants in web archiving have expressed the need for multiple and varied access points to the same archived web resources. Therefore, focusing on the creation of access points that are suitable for different disciplines who are using the same primary resources - can build interdisciplinary communities that cut across fields with shared resources and common methods.

A participatory, inclusive and representative knowledge ecology can achieve what current knowledge management practices have failed to do – create an inclusive knowledge ecology where access means readability, retrievability, connecting disparate and closely related information, and enabling connections between users in order to make meaning that can be used to create new

knowledge but also support preservation. Social, community-built tools provide viable alternatives to authoritative systems that derive their management from strict process, workflow, security and control and can make user-driven meaning-making part of the process of accessibility. Hierarchical and ontological information management cannot include the deep contextual and cultural usage meanings that might easily place one object in multiple categories. The restrictions that arise from authoritative management of knowledge can be avoided with the participatory, inclusive and representative knowledge ecology that is fostered by social, community tools, although an approach that is too decentralized runs the risk of having a chaotic approach to standards, or no standards at all. Or, as Julien Masanès of the European Archive suggested when interviewed, *"what we need is a CERN for web archives."*

Page is intentionally blank

34

Appendix A: Interviews

For this project, we supplemented desk research with 17 interviews with a number of stakeholders in the web archiving community. We are grateful to the following individuals for generously helping us to better understand how archivists and researchers are engaging with web archives.

Niels Brügger
Associate Professor, Department of Information and Media Studies
Aarhus University, Denmark

Richard Davis
Repository Service Manager
University of London Computer Centre, United Kingdom

Katrien Depuydt
Head of the Language Database Department
Institute for Dutch Lexicology, The Netherlands

Kirsten Foot
Associate Professor of Communication
University of Washington, United States of America

Wendy Gogel
WAX Project Manger
Harvard University Library, United States of America

Alison Hill
Curator, Web Archiving, Modern British Collections
The British Library, United Kingdom

Helen Hockx-Yu
Web Archiving Programme Manager
The British Library, United Kingdom

Hanno Lecher
Librarian, China Studies
Leiden University, The Netherlands

Julien Masanès
Director
European Archive, France

Frank McCown
Assistant Professor of Computer Science
Harding University, United Kingdom

Mark Middleton
CEO, Hanzo Archives, United Kingdom

Martin Moyle
Digital Curation Manager
University College London (UCL) Library Services, United Kingdom

Kris Carpenter Negulescu
Director of the Web Archive
Internet Archive, United States of America

Ed Pinsent
Digital Archivist/Project Manager
University of London Computer Centre, United Kingdom

Steve Schneider
Professor & Interim Dean, School of Arts & Sciences
SUNY Institute of Technology, United States of America

René Voorburg
Crawl-engineer & Coordinator of web archiving
Acquisition and Processing Division – E-depot
Koninklijke Bibliotheek, The National Library of the Netherlands, The Netherlands

Max Wilkinson
Datasets Programme Technical Lead
British Library, United Kingdom

References Cited

Adar, E., Teevan, J., Dumais, S. T., & Elsas, J. L. (2009). *The web changes everything: understanding the dynamics of web content.* Paper presented at the Second ACM International Conference on Web Search and Data Mining, Barcelona, Spain.

Alpert, J., & Hajaj, N. (2008, 25 July). We knew the web was big... Retrieved from http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

Anderson, C., & Wolff, M. (2010, September). The Web is Dead. Long Live the Internet. *Wired*.

Arms, W. Y., Adkins, R., Ammen, C., & Hayes, A. (2001). Collecting and preserving the Web: The Minerva Prototype. *RLG Diginews, 5*(2).

Baroni, M., & Bernardini, S. (Eds.). (2006). *WaCky! Working Papers on the Web as Corpus*. Bologna: GEDIT.

Brock, A. (2005). "A belief in humanity is a belief in colored men": Using culture to span the digital divide. *Journal of Computer-Mediated Communication, 11*(1), article 17.

Brügger, N. (2005). *Archiving websites: general considerations and strategies*. Århus: Center for Internet-forskning.

Burner, M. (1997). Crawling towards eternity: Building an archive of the world wide web. *Web Techniques Magazine, 2*(5), 37-40.

Cho, J., & Garcia-Molina, H. (2000, 10-14 September). *The evolution of the web and implications for an incremental crawler.* Paper presented at the 26th International Conference on Very Large Databases, Cairo, Egypt.

Chu, S.-C., Leung, L. C., Van Hui, Y., & Cheung, W. (2007). Evolution of e-commerce Web sites: A conceptual framework and a longitudinal study. *Information & Management, 44*(2), 154-164.

Day, M. (2003). *Preserving the fabric of our lives: A survey of web preservation initiatives*. Paper presented at the European Conference on Research and Advanced Technology for Digital Libraries, Trondheim, Norway.

Dougherty, M. (2007). *Archiving the Web: Collection, Documentation, Display and Shifting Knowledge Production Paradigms (Ph.D. thesis).* University of Washington, Seattle.

Dougherty, M. (2008). *Making web archives valuable for researchers: Exploring the state of the art, annotation practices, and possibilities for progress*. VKS Working Paper. Virtual Knowledge Studio. Maastricht, The Netherlands.

Dougherty, M., Foot, K. A., & Schneider, S. M. (2010). *Ethics in/of Web Archiving*. Paper presented at the Computer Supported Cooperative Work Pre-conference on Revisting Research Ethics in the Facebook Era: Challenges in Emerging CSCW Research, Savannah, GA.

Dougherty, M., Schneider, S. M., & Jones, J. (2010, Forthcoming). Web Historiography and the Emergence of New Archival Forms. In D. W. Park, S. Jones & N. W. Jankowski (Eds.), *The Long History of New Media: Technology, Historiography, and Newness in Context.* New York: Peter Lang Publishing.

Fetterly, D., Manasse, M., Najork, M., & Wiener, J. (2004). A large-scale study of the evolution of web pages. *Software-Practice and Experience, 34*, 213-237.

Foot, K. A., & Schneider, S. M. (2002). Online action in campaign 2000: An exploratory analysis of the US political Web sphere. *Journal of Broadcasting & Electronic Media, 46*(2), 222-244.

Foot, K. A., & Schneider, S. M. (2006). *Web campaigning*. Cambridge, MA: The MIT Press.

Franklin, M. (2005). *Postcolonial Politics, the Internet, and Everday Life: Pacific Traversals Online*. London: Routledge.

Hackett, S., & Parmanto, B. (2005). A longitudinal evaluation of accessibility: Higher education web sites. *Internet Research, 15*(3), 281-294.

Hendler, J., Shadbolt, N., Berners-Lee, T., & Weitzner, D. (2008). Web Science: An Interdisciplinary Approach to Understanding the Web. *Communications of the ACM, 51*(7), 60-69.

Hodge, G. M. (2000). Best Practices for Digital Archiving: An Information Life Cycle Approach. *The Journal of Electronic Publishing, 5*(4).

Hundt, M., Nesselhauf, N., & Biewer, C. (Eds.). (2007). *Corpus linguistics and the web*. Amsterdam: Editions Rodopi B.V.

Innis, H. (1951). *The Bias of Communication*. Toronto: University of Toronto Press.

JISC. (2008). PoWR: The Preservation of Web Resources Handbook. London: JISC.

Johnson, P. (2009). *Fundamentals of Collection Development and Management (Second Edition)*. Chicago, IL: American Library Association.

Kahle, B. (1997). Preserving the Internet. *Scientific American, 276*(3), 82-83.

Kahle, B., Prelinger, R., & Jackson, M. (2001). Public access to digital material. *D-Lib Magazine, 7*(4).
Kelly, K. (2006, 14 May). Scan this book! *New York Times Magazine*.
Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics, 29*(3), 333-347.
Knutson, A. N. (2009). Proceed with Caution: How Digital Archives Have Been Left in the Dark. *Berkeley Technology Law Journal, 24*(1), 437-473.
Koehler, W. (2004). A longitudinal study of Web pages continued: a consideration of document persistence. *Information Research, 9*(2).
Levinson, P. (1997). *The Soft Edge: Natural History and Future of the Information Revolution*. New York: Routledge.
Lyman, P., & Kahle, B. (1998). Archiving Digital Cultural Artifacts: Organizing an agenda for action. *D-Lib Magazine, July/August*.
Lyman, P., & Varian, H. R. (2003). How Much Information 2003? Retrieved 18 August, 2010, from http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/
Masanès, J. (2002). Towards Continuous Web Archiving: First results and an agenda for the future. *D-Lib Magazine, 8*(12).
Masanès, J. (2005). Web Archiving Methods and Approaches: A comparative study. *Library Trends, 54*(1), 72-90.
Masanès, J. (2006). *Web archiving*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.
McEnery, T., & Wilson, A. (2001). *Corpus linguistics*. Edinburgh: Edinburgh University Press.
Meyer, E. T., Carpenter, K., & Middleton, M. (2009). World Wide Web of Humanities: Final Report to JISC. Online: http://www.jisc.ac.uk/media/documents/programmes/digitisation/humanitiesfinalreport.pdf. London: JISC.
Morville, P. (2005). *Ambient Findability: What We Find Changes Who We Become*. Sebastopol, CA: O'Reilly Media, Inc.
Murphy, J., Hashim, N. H., & O'Connor, P. (2008). Take Me Back: Validating the Wayback Machine. *Journal of Computer-Mediated Communication, 13*(1), 60-75.
Patel, K. (2007). Authors v. Internet Archives: The Copyright Infringement Battle over WEB Pages. *Journal of the Patent and Trademark Office Society, 89*, 410-428.

Ras, M., & van Bussel, S. (2007). Web Archiving User Survey. Retrieved from http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/documenten/KB_UserSurvey_Webarchive_EN.pdf
Raymond, M. (2010, 28 April). The Library and Twitter: An FAQ. Retrieved from http://blogs.loc.gov/loc/2010/04/the-library-and-twitter-an-faq/
Schneider, S. M., & Foot, K. A. (2002). Online Structure for Political Action: Exploring Presidential Campaign Web Sites from the 2000 American Election. *Javnost-The Public, 9*(2).
Schneider, S. M., & Foot, K. A. (2004). The web as an object of study. *new media & society, 6*(1), 114-122.
Schneider, S. M., & Foot, K. A. (2005). Web Sphere Analysis: An Approach to Studying Online Action. In C. Hine (Ed.), *Virtual Methods: Issues in Social Research on the Internet* (pp. 157-170). Oxford: Berg.
Schneider, S. M., & Foot, K. A. (2010). Object Oriented Web Historiography. In N. Brügger (Ed.), *Web History*. New York: Peter Lang Publishing.
Simonite, T. (2010). A Search Service that Can Peer into the Future: A Yahoo Research tool mines news archives for meaning--illuminating past, present, and even future events. *Technology Review*. Retrieved from http://www.technologyreview.com/computing/26113/
Taycher, L. (2010, 05 August). Books of the world, stand up and be counted! All 129,864,880 of you. Retrieved from http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html
Taylor, M. K., & Hudson, D. (2000). "Linkrot" and the usefulness of Web site bibliographies. *Reference & User Services Quarterly, 39*(3), 273-276.
Thelwall, M., & Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research, 26*(2), 162-176.
van den Heuvel, C. (2009). Web Archiving in Research and Historical Global Collaboratories. In N. Brügger (Ed.), *Web History* (pp. 279-303). New York: Peter Lang Publishing.
Veronin, M. A. (2002). Where Are They Now? A Case Study of Health-related Web Site Attrition. *Journal of Medical Internet Research, 4*(2).
Weinreich, H., Obendorf, H., Herder, E., & Mayer, M. (2008). Not quite the average: An empirical study of Web use. *ACM Transactions on the Web, 2*(1), 1-31. doi: http://doi.acm.org/10.1145/1326561.1326566

B.A. Howell (2006). "Proving web history: How to use the Internet archive." *Journal of Internet Law*. 9(8): 3-9.

Proving Web History: How to Use the Internet Archive

By Beryl A. Howell

Showing what content a Web site *previously* contained (as opposed to what is currently on the site) may help answer questions that attorneys confront in a myriad of cases, ranging from copyright and trademark infringement to business torts and defamation. Showing that a particular Web site is currently using copyrighted text or images or protected marks may be all that is needed in a case, but documenting prior versions of the Web site can be critical to establish the scope or extent of the illegal or tortious conduct, the amount of the damages, or the requisite *mens rea*.[1] When the Web site at issue or the offending content on it has been removed or modified, the most effective, if not the only, way to document the content is to review prior versions stored in online archives of Internet sites.

Specifically, in trade secret and misappropriation cases, showing that the same information claimed to be secret or confidential has previously been made publicly available by the claimant, such as on the claimant's Web site, can be probative if not case dispositive. Diligently searching archived versions of the claimant's Web site for such evidence can be worthwhile.

Similarly, capturing evidence from archived Web sites is helpful in intellectual property infringement cases as well. For example, in cybersquatting and typo-squatting cases, where a trademarked name or a slightly misspelled trademarked name (*e.g.*, *mcrosoft.com*) has been registered as a domain name and used as an online address for a Web site, evidence from archived versions of the Web site may establish the period of time the offending Web site has been operational and the types of goods or services offered on the site over time. This evidence can help establish intent and harm. In cyberstuffing cases, where popular trademarked names are repeatedly embedded in hidden metatags and transparent text on a Web site, search engines will pick up on the trademarked names and push the infringer's Web page to the top of search engine results, diverting business from the trademark owner's site. Even if the Web site is modified after the infringer is notified of the claim, documenting the cyberstuffing activity on archived versions of the Web site can establish the nature of the offending activity, its scope, and duration.

No matter the legal context, gathering evidence of prior versions of Web sites should be performed in a careful forensic manner with cognizance of the underlying technology used in the archiving process. This article will review strategies and methods for capturing prior versions of Web sites from the most popular of the archives and considerations that counsel should be prepared to address in authenticating this evidence.

Beryl A. Howell is a Partner and heads the Washington, DC, office of Stroz Friedberg, LLC, a computer forensics and electronic discovery consulting and technical services firm with offices also in New York City, Minneapolis, and Los Angeles. She formerly served as a New York federal prosecutor and General Counsel of the US Senate Committee on the Judiciary. *Research assistance for this article was provided by Donald Allison and Jessica Reust, computer forensic examiners, and George McLean, Evidence Technician at Stroz Friedberg, LLC.*

"MAP" OF ARCHIVES

At the outset, archived versions of Web sites are available for free at multiple sites. The federal government, in particular, archives government Web sites and makes those archives accessible online. For example, the US Government Printing Office, in partnership with the University of North Texas, provides online access to federal Web sites that have ceased operation on a site called the CyberCemetery.[2] The archived deceased Web sites include Access America, Advisory Commission on Intergovernmental Relations, Office of Technology Assessment, and others. Similarly, the Electronic Research Collection (ERC),[3] which is a partnership between the United States Department of State and the Federal Depository Library at the Richard J. Daley Library, University of Illinois at Chicago (UIC), makes available the US Department of State Web pages archived from 1998 through January 2001. In addition, the National Archives and Records Administration harvested all of the federal agency public Web sites as they existed at the end of the presidential term on January 20, 2005, and makes these archives available at the 2004 Presidential Term Web Harvest site.[4]

Archives of Web sites that are not associated with the federal government are available at several sites. The Library of Congress sponsors a project called Minerva (Mapping the INternet Electronic Resources Virtual Archive), which harvests Web sites based on subject matter and then provides the collections as an archive, rather than try to harvest every Web site. The collections currently available include: Election 2002 Web Archive (July

1, 2002-Nov 30, 2002);[5] September 11, 2001 (September 11, 2001-December 1, 2001);[6] and Election 2000 (August 1, 2000-January 21, 2001).[7]

Certain Internet search engines, such as Google and Yahoo, also make archived Web sites available. The Google archive provides access to the last cached version of a Web site, but not to prior versions. These cached Web sites are a backup in case the original page is unavailable and are useful since they show the date and time stamps for when each page on the site was retrieved by Google. Google and other search engines often index a Web site about once a month, but Google explains that the "cache is the snapshot that we took of the page as we crawled the web" and cautions that "[t]he page may have changed since that time" or "[t]his cached page may reference images which are no longer available." Google states that many factors affect how often it indexes a site, but a 2003 survey showed that Google revisited most sites within one month.[8] Therefore, unless a page is defunct, a Google cached site often will be 30 days old or less. To look farther back in time, the Internet Archive is probably a better bet. Sites may not be cached if they have not been indexed or if the owners have requested that the content not be cached. The date-time stamps on the Google archive may be helpful in establishing, for example, when a site stopped operating within the last six months. If a site is no longer available online, a visit to the Google cache may indicate the date when the site was last indexed.

Yahoo has recently added the ability to view cached pages by clicking on a link entitled "cached." As with Google, clicking on "cached" brings up a copy of the Web page as it appeared when it was last crawled by the search engine. By contrast to the Google cached sites, however, the Yahoo archive does not date-stamp the version of the cached site but simply notes the following: "It's a snapshot of the page taken as our search engine crawled the Web. The Web site itself may have changed." To check the previous versions of the Web site, Yahoo directs users to the Internet Archive. As discussed in more detail below, the Internet Archive contains the most extensive archive of Web sites in terms of period covered, number of Web sites and pages archived, and the number of prior versions of Web sites archived.

Other search engines that provide cached Web sites include *search.msn.com* (MSN), *ask.com* and *teoma.com* (both from Ask Jeeves), *clusty.com* (from meta-search engine Vivisimo), and *Gigablast.com*. Of these, Gigablast may be the most helpful in researching historic Web sites because its search engine results include the date that the Web page was last modified, as well as the date that the page was last indexed by Gigablast. Gigablast also provides links to the cached site, a stripped version of the site without graphics, and a link to "older copies" found on *archive.org*.

THE INTERNET ARCHIVE AND THE WAYBACK MACHINE

The Internet Archive[9] is a free online resource that was created in 1996 to build a digital library of Web pages and other cultural artifacts in digital form with the purpose of offering permanent and free access to researchers, historians, scholars, and the general public.[10] Internet Archive provides not only an archive of websites but also of open source movies, feature films, cartoons, historic newsreels, and news video and music.

Five years after its creation, in October 2001, the Internet Archive launched the Wayback Machine, which provides the public with a free online service to search for and access archived Web sites. The name of the search service is derived from the Rocky and Bullwinkle cartoon in which the characters of a bow-tied dog, Mr. Peabody, and his boy assistant, Sherman, used a time machine called the WABAC Machine to travel back in time to famous events in history.

The Web pages are collected for the Internet Archive using a search engine technology called Alexa Crawl that traverses the Internet taking snapshots of Web sites. The Alexa Crawl currently captures about 1.6 terabytes (1600 gigabytes) of Web content per day and takes about two months to complete a snapshot of the more than 16 million Web sites accessible online.[11] This search-and-copy engine is owned and operated by Alexa Internet, a for-profit company that offers a free toolbar and a number of statistical services to subscribers based upon the Web content and usage information collected. The company donates a copy of each crawl of the Web to the Internet Archive, which may make the crawl results available after six months. Thus, there is a six- to 12-month lag between the date that a site is crawled and when it appears for free use in the archives of the Wayback Machine.[12] Alexa Internet is now offering a fee-based service to access its crawl results data before it goes to the Internet Archive.[13]

The Alexa Crawl does not purport to capture all Web sites accessible on the Internet, but instead prioritizes the Web sites and pages to copy based on the number of times that a Web site is requested through the Alexa search engine. Thus, not every Web site has an equal chance of being copied at all or copied in full. Alexa Internet uses a rating system for content at all that will be captured. Content that is not popular may be deliberately omitted if not visited often. This is related to Alexa's business model for selling databases of frequently visited sites to customers. The result is that the Wayback Machine does not hold

archived versions of all Web sites of copies of every page for the Web sites that are archived.

In addition, sites may not be archived if they are password-protected, the site owners have requested exclusion from the Wayback Machine, or the crawler is blocked by use of a technical flag installed by the site owner called robots.txt, or the site is otherwise inaccessible. When the site is blocked by request or use of a robots.txt flag, the Wayback Machine search engine will indicate this with an error message, such as "blocked site error" or "robots.txt query exclusion error."

At the inception of the Wayback Machine, the Internet Archive contained 100 terabytes of data that was growing at a rate of 10 terabytes per month. By 2005, the amount of data stored in it is more than a petabyte, with a growth rate of 20 terabytes per month, making the Internet Archive the largest data archive in the world. All of this data is stored in huge server farms in the Presidio of San Francisco.

The archived Web sites are stored across multiple servers. A version of a particular Web site that is shown as indexed on the Wayback Machine may not be available at the time when a user wants to access it. A replica of the Internet Archive is stored at the Bibliotheca Alexandrina in Egypt.[14] If a version of a particular Web site cannot be accessed on the Internet Archives' primary site, the replica site can be checked.

The replica on the Bibliotheca Alexandrina Web site is not updated frequently, however, and it does not contain as much content. Test searches conducted on *archive.org* reveals many Web sites that do not appear on Alexandrina's Web site. For example, a search for *cnn.com* yields results for pages from July 2000-September 18, 2001, on the Alexandrina's Web site, while the *archive.org* site has version from November 26, 2004.

To use the Wayback Machine, users simply go to the *archive.org* Web site, and type in the Internet address[15] in the provided search box. Any versions of the Web site corresponding to the Internet address that are archived on the servers of the Internet Archive will pop up in a chronological list. A user can review this list and select the version or date for review by clicking on the selected date. The archived version of the Web site for the date selected will then appear and can be reviewed.

The nature of the legal dispute may require analysis of multiple archived versions of a particular Web site in order to establish whether and how content changed. For example, in a contract dispute, the question of whether a party offered services or items in violation of terms in the license at issue may require documenting changes in a party's advertised offerings on its Web site during and after expiration of the license term. Critical text may simply be eyeballed as part of this analysis to document changes over time. In addition, the Wayback Machine notes changes in an archived Web site with an asterisk. This asterisk system alerts only to changes in text or graphics and not to modifications in internal or external links and or in the source code for the Web site. This may become critical if, for example, the archived Web site is cited as evidence that it was used to link to an offending site. The link to the offending site in the archive version may not, in fact, have existed or existed in the same form at the time that version of the Web site was copied for the archive.

The Wayback Machine also offers a free service of comparing any two versions of an archived site using a technology called DocuComp, which is a patented algorithm licensed by Advanced Software for use in the Wayback Machine. The comparison can show how the contents, including text, images, and links, have changed over time and between any two versions being compared.

"MISSING" ARCHIVED WEB SITES

When a search for an archived Web site has negative results, this does not mean that the Web site does not exist, is not archived, or is only of current vintage. The Web site may have been excluded from the archiving process or in fact, the Web site may be archived but review of the archived versions is blocked. The Internet Archive takes steps to avoid archiving web sites for which the owner has indicated a preference to be excluded. A universal technical standard that indicates an exclusion preference is called the standard for robot exclusion (SRE). A file called robots.txt can be added to the header information on a Web site or specific Web page by an owner, and a denial or disallow command within that file can serve as a flag that the owner does not want the entire Web site or particular Web pages copied or scanned by a Web crawler. In other words, the directions in the robots.txt file can be set to allow full or partial copying or copying exclusion. The Alexa crawler respects this preference and will not copy those sites or pages with a robots.txt file embedded.[16] Alexa Internet and Internet Archive take this respectful technology a step further: When robots.txt is added to a Web site, Alexa will exclude the site from being copied by its crawler, and the Internet Archive will go back into archived sites to remove content already captured.[17]

In addition, intellectual property owners who believe that infringing activity is occurring on a Web site may contact the Internet Archive and request exclusion of the offending Web site. The Internet Archive provides specific directions to copyright and trademark owners seeking to have third-party Web sites containing infringing works removed from the archive. These owners must specifically

identify the work allegedly being infringed and where it is located within the Internet Archive collections, contact information, and a statement made under penalty of perjury that use of the work is unauthorized by the copyright owner, along with an electronic or physical signature.[18]

The Internet Archives' respect for the exclusion preference of Web site owners and compliance with its own stated policy to remove Web sites with robots.txt flags is the subject of a recent suit in the Eastern District of Pennsylvania brought by Healthcare Advocates against the Internet Archive for, *inter alia*, breach of contract and misrepresentation due to a failure to block access to the plaintiff's archived Web sites.[19] The plaintiff operates a Web site that describes the services of the company, including helping the public get reimbursements for health care expenses, reporting on medical research, providing doctor referrals and information on discount prescriptions and healthcare plans. The company claims copyright in all of the Web site content. In mid-2003, the plaintiff installed the denial text string in the robots.txt file on the computer server hosting its Web site with the expectation that the Internet Archive would prevent users of the Wayback Machine from gaining access to the archived versions of its Web site.

Nevertheless, in another case brought by Healthcare Advocates against a competitor for misuse of proprietary and trade secret information, the defendant's counsel was able to access the archived versions of the plaintiff's Web site on the Wayback Machine by successfully circumventing the security offered by the denial text string in the robots.txt file. This circumvention was apparently facilitated by the fact that "the mechanism preventing www.archive.org from searching a particular web site's host computer server for a denial text string in the robots.txt file more than once per day was 'broken.'" In other words, when the Wayback Machine receives a query for an archived version of a Web site, the Web site is pinged for the presence of a robots.txt file denial string. If the string is found, the query is blocked, but apparently persistent queries will overcome the block. The defendant's counsel in the underlying lawsuit conceded that the plaintiff's archived Web sites on the Wayback Machine had been searched and accessed in connection with that underlying case. That counsel is now co-defendants with the Internet Archive in Healthcare Advocates' suit for copyright infringement and computer hacking.

This lawsuit will test the scope and merits not only of the claims at issue but also the indemnification provision of the Internet Archive's terms of use. Specifically, the terms governing the use of the collection of archived Web pages is predicated on the user's agreement "to indemnify and hold harmless the Internet Archive and its parents, subsidiaries, affiliates, agents, officers, directors, and employees from and against any and all liability, loss, claims, damages, costs, and/or actions (including attorneys' fees) arising from your use of the Archive's services, the site, or the Collections."[20]

CAPTURING ARCHIVED WEB SITES

Once an archived Web site has been located, the methods of capturing the virtual pages in a concrete form for use in court can vary. One method is to print each page that appears on the computer screen. The person performing or supervising the search and printing can attest to the date, time, and process used to obtain the printout. This method shows static pages of the Web site without any of the links that may remain active, other than any advertisements pushed to the site, even in the archived state. Similarly, screen-shots of each page viewed can be saved electronically for incorporation into expert reports or affidavits.

Importantly, Internet browsers and specialized tools used by computer forensic experts for downloading Web sites with metadata intact can be used to capture not only the graphical display of a Web page but also the underlying html code that is driving the display. Simply using the file save function on a browser can preserve code that may reveal who authored a contentious Web page. Saving underlying code in the same way may reveal a trademarked name written over and over again in white-on-white text, indicating that it was meant to be revealed to crawling search engines but hidden from a consumer's (or competitor's) naked eye. If two or more archived pages are linked to each other, download tools can provide a fuller layout of a Web site with its underlying code. At trial, this fuller layout can be presented to the judge or jury, and links and related pages can be navigated, much as an historic user might have surfed them.

In addition, specialized software tools are available that allow dynamic presentations, including demonstrations of any link that remains active on the Web site. One such software tool, called Camtasia, can be installed on the computer used to access the archived site to record every keystroke and screen shot appearing during review of the cached Web site. The recording of the review session is documented real time in video-like form that may be stored on a CDR or DVD for submission to court. For example, in a business diversion case, a recording of the cached version of the defendant's prior Web site may be able to show links that remain active and purportedly direct potential customers to the plaintiff's products, but the links instead actually channel users to the defendant's sites.

Beware when capturing an archived Web site that

different browsers display Web sites with differing degrees of accuracy and completeness, and this holds true for archived Web sites and Web pages as well. There are a number of different reasons why some Web pages look different depending on which browser is used to view the page, including browser adherence to Web page standards, browser support of different technologies, and Web sites that do not use Web page standard code. The World Wide Web Consortium (WC3) develops the standard elements for Web site programming, which some browsers adhere to and some do not. For example Firefox and Mozilla adhere to the WC3 standards, while Internet Explorer supports additional non-standard Web-programming technologies. The resulting difference in the way that Web pages are displayed may be as minimal as the color of the scrollbar to as inconvenient as the navigation menus not working or the site content not being displayed at all.

A Web site that uses or requires a certain technology to be viewed will not be displayed correctly or completely by a browser that does not support that technology. For example, Firefox does not support ActiveX, which are software components from Microsoft that enable sound, Java applets, and animations to be integrated in a Web page.[21] For example, using a browser that supports ActiveX is necessary in order to access the Windows Update Web site, which otherwise will simply not be displayed but with an alert to the viewer that content is hidden from view. The fact that content is not being displayed or displayed in a different way from the original site is not always apparent.

The key to capturing an archived Web site as accurately and completely as possible is to examine the underlying code used to create and support the Web site to determine whether a browser is incompatible. This can be done by an examination of the source code for the initial page of the Web site. The entry point for the Web site usually includes language that will query and collect information from the browser and its computer system settings to determine the best method of providing the information from the site. For Web sites that use only standard html coding, the content and features of the site usually have the least variance across browsers. Where non-standard html coding is revealed, forensic experts capturing Web sites for litigation purposes may display the Web site with multiple browsers as a test to ensure that the display does not vary by browser and if variances are noted, capture the Web site with the browser that displays the most content.

ADMITTING INTERNET ARCHIVE DATA

Information obtained from reputable or government-sponsored online sources has generally been held admissible. For example, in *U.S. Equal Employment Opportunity Commission v. E.I. DuPont De Nemours & Co.*,[22] the defendant moved to exclude as an exhibit the printout of a table from the Web site of the US Census Bureau as inadmissible hearsay and lack of trustworthiness. The court denied the motion, stating that the hearsay exception for a public record applied. In addition, the court concluded that the printout was sufficiently authenticated under Federal Rules of Evidence 901(a) since it contained the "internet domain address from which the table was printed, and the date on which it was printed."[23] The court performed its own verification as well, noting that "[t]he Court has accessed the website using the domain address and has verified that the webpage printed exists at that location."[24] Similarly, printouts of data from other government-sponsored Web sites have been admitted over objection to the reliability of the information.[25]

Reported cases involving Web site captures from the Internet Archive are rare, even though *archive.org* is an important resource for litigators trying to establish prior representations or actions on Web sites. Significantly, in the few cases where challenges have been interposed to Internet Archive versions of Web pages, the evidence has been admitted over hearsay and authentication challenges.

The leading case for admission of archived Web sites from the Internet Archive is *Telewizja Polska USA, Inc. v. Echostar Satellite Corporation*.[26] The plaintiff in this case claimed that Echostar improperly had used the plaintiff's trademarks in "TV Polonia," a Polish-language television station, to sell subscriptions to the Dish Network satellite TV service after the contract allowing such marketing rights had expired in early 2001. Echostar argued that plaintiff had itself advertised that the Dish Network carried TV Polonia on its Web site after the marketing rights had expired and offered an exhibit of the plaintiff's Web site at various times in 2001 confirming this past Web site content. The plaintiff filed a motion *in limine* to bar Echostar from offering the exhibit on the grounds of double hearsay and lack of authentication. The court rejected these grounds and denied the motion, stating that "the contents of [plaintiff]'s website may be considered an admission of a party-opponent and are not barred by the hearsay rule."[27] In addition, the court relied on the affidavit of "Ms. Molly Davis, verifying that the Internet Archive Company retrieved copies of the websites as it appeared on the dates in question from its electronic archives."[28] The plaintiff "presented no evidence that the Internet Archive is unreliable or biased" or "denied that the exhibit represents the contents of its website on the dates in question" or otherwise "challenged the veracity of the exhibit."

AUTHENTICATION CONSIDERATIONS FOR ARCHIVED WEB SITES

The versions of Web sites and pages archived on Internet Archive can provide valuable and significant probative evidence in a variety of cases. To authenticate copies of prior versions of Web sites obtained from the Wayback Machine, a party proffering the evidence must show, under Federal Rules of Evidence 901(a) that the "matter in question is what its proponent claims." This can be done by producing the testimony, either orally or in written form, of the person who copied or supervised the copying of the archived Web site and the process followed to accomplish this task. In addition, the proponent must establish the general reliability of the copy.

The capture and use as evidence of archived Web site material must be approached with a full appreciation of three primary technical features and limitations that may affect the archived copy in order to respond to any challenges that may be raised to the completeness, reliability, and authenticity of the copy. For this reason, expertise in digital forensics, including the methods of forensic capture and documentation of the archived Web site proffered, may be recommended depending on the issue for which the archived Web site is being offered.

First, archived Web sites on the Internet Archive are compilations made over time. While the archived versions of Web sites are date- and time-stamped, the pages for each version of the Web site may not have been copied simultaneously. The Alexa crawler may take multiple passes at a Web site over the course of up to two days to try to capture the entire Web site. In short, due to bandwidth and storage constraints, all of the data on a Web site may not be captured at the same time. The Internet Archive explains that "Sites are usually crawled within 24 hours and no more than 48."[29]

Second, the archived versions of Web pages available through the Wayback Machine may not contain all of the content on each Web page that is captured. What you see is not always the complete story.

For example, when a Web site contains elements that require interaction with the originating host, copying that page for archiving breaks the necessary link with the original site, thereby reducing the functionality or eliminating entirely that particular element. The result is that the archived Web page or site has missing material, which may not be apparent or flagged for the viewer. Similarly, links originally enabled with a java script, which the Alexa crawl technology disables during the capture of the Web site or Web page, would no longer work.[30] The Internet Archive acknowledges: "Not all images are archived nor are retrievable from the original site. If they no longer exist on the original site then the images will not be available and not displayed within the archived pages."[31] Other types of coded content that the crawler technology does not capture include Flash enabled content, some photographic images, and some html coded content.

Moreover, content may not ever be captured if problem technology, such a password protected pages, or respectful technology, such as a robots.txt flag, is encountered. Additionally, even after the capture is completed, archived copies of Web sites may have content deleted if a robots.txt flag is added to the site or if a request for deletion is sent to the Internet Archive. Thus, the archived copy may show what was captured but not what was skipped or subsequently omitted.

Finally, depending on the technical sophistication of the Web site and its use of internal and outside linked material, the copy of the archived version of the Web site may not show links that existed on the Web site at the time of the original capture. Links that may have worked at the date of capture may be inactive because they simply no longer exist or are not in the archive library.

The links on archived Web sites may remain active but link to different material from that associated with the Web page at the time that it was archived. The linked material may be to current sites or to other stored link sites from a different time. Indeed, links may connect to current active sites and show *current* banner advertisements available at the site, rather than linking to sites as they existed at the date of capture. When the active links on archived Web sites pull information from the current site, the owner of the current Web site can track how many times the Wayback Machine is being queried for archived versions of the Web site. Logs of incoming IP addresses maintained by the server hosting the current Web site can reveal whether the incoming IP address originated with the Internet Archive.[32]

Alternatively, the working link may connect to sites or pages archived on the Wayback Machine around the time of the original Web site to which the link connected. The Internet Archive explains: "When you are surfing an incomplete archived site the Wayback Machine will grab the closest available date to the one you are in for the links that are missing. In the event that we do not have the link archived at all, the Wayback Machine will look for the link on the live web and grab it if available."[33] In short, the process of copying a Web site for archiving may result in changes to the extent that the archived Web site may not show accurately the links that existed at the time shown for the Web site storage date. The Alexa Internet crawler technology rewrites the original link code in html to re-direct links to current or stored links.

Determining whether the content on a linked site is contemporaneous with the archived version of the site or dates from another time may be critical. For example, establishing that a linked promotion to a site containing infringing material persisted after notification from the copyright owner may be important to establish knowledge and intent in a copyright infringement suit. Each link must be checked for the date code embedded in the archived URL, or location within the Wayback Machine database, to verify whether the linked content is contemporaneous, current, earlier or later than the version of the archived Web site or page. The Internet Archive provides the following example: "in this url *http://web.archive.org/web/20000229123340/http://www.yahoo.com/* the date the site was crawled was Feb 29, 2000, at 12:33 and 40 seconds."[34]

Increasingly, documentation of offending activity that occurred on Web sites of opposing parties is relevant and, in some cases, dispositive of certain types of claims. Searching for, reviewing, and capturing archived copies of Web sites can be easily accomplished from the Internet Archive, but litigators should consider carefully the methods of capture and the issues surrounding the completeness, reliability, and authenticity of the Web site copies.

NOTES

1. *See, e.g.*, Van Wetrienen v. Americontinental Collection Corp., 94 F. Supp. 2d 1087, 1109 (D. Or. 2000) (contents of defendant's Web site relevant to determination of whether defendant's conduct was so egregious as to merit an award of punitive damages).
2. The CyberCemetery is located at *http://govinfo.library.unt.edu*.
3. ERC is located at *http://dosfan.lib.uic.edu/ERC/*.
4. The 2004 Presidential Term Web Harvest is located at *http://www.webharvest.gov/collections/peth04/*.
5. *http://www.loc.gov/minerva/collect/elec2002/index.html*.
6. *http://www.loc.gov/minerva/collect/sept11/index.html*.
7. *http://www.loc.gov/minerva/collect/elec2000/index.html*.
8. *See http://searchengineshowdown.com/stats/freshness.shtml*.
9. The Internet Archive is located at *www.archive.org*.
10. Kahle v. Ashcroft, 2004 U.S. Dist. LEXIS 24090, *5 (N.D. Cal. Nov. 19, 2004).
11. *http://pages.alexa.com/company/technology.html*.
12. *http://www.archive.org/about/faqs.php#The _Wayback_Machine*.
13. *http://websearch.alexa.com/welcome.html*.
14. *http://www.bibalex.org/english/initiatives/internetarchive/web.htm*; *see also http://en.wikipedia.org/wiki/Bibliotheca_Alexandrina*.
15. The technical term for an Internet address is Universal Resource Locator or URL.
16. Directions for removal of a Web site from the archive are found at *http://www.archive.org/about/exclude.php*.
17. *http://www.archive.org/about/faqs.php#2* ("By placing a simple robots.txt file on your Web server, you can exclude your site from being crawled as well as exclude any historical pages from the Wayback Machine.").
18. *Id.*
19. Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey, Civil Action (E.D. Pa., filed July 8, 2005), copy at *http://www.geocities.com/bledrydudenet/Healthcare_Advocates_v._Harding_Complaint__FINAL.pdf*. Healthcare Advocates, Inc. unsuccessfully moved to have the counts against the law firm for, *inter alia*, violations of the DMCA and the Computer Fraud and Abuse statute added to the underlying complaint, but that motion was denied. Flynn v. Health Advocate, Inc., 2004 U.S. Dist LEXIS 12536, *12 (E.D. Pa. July 8, 2004).
20. *http://www.archive.org/about/terms.php/*.
21. *See http://webmaster.lycos.co.uk/glossary*.
22. U.S. Equal Employment Opportunity Commission v. E.I. DuPont De Nemours & Co., 2004 U.S. Dist. LEXIS 20753 (E.D. La. Oct. 18, 2004).
23. *Id.* at *5.
24. *Id.*
25. *See* Chapman v. San Francisco Newspaper Agency, 2002 U.S. Dist. LEXIS 18012 at*2 (N.D. Cal. Sept. 20, 2002) (computer printout of page from US Postal Service Web site was sufficiently reliable to be admissible public record). *But see* St. Clair v. Johnny's Oyster & Shrimp, Inc., 76 F. Supp. 2d 773, 774 (S.D. Tex. 1999) (court deemed plaintiff's proffered data from the US Coast Guard's online vessel database insufficient since "any evidence procured off the Internet is adequate for almost nothing").
26. Telewizja Polska USA, Inc. v. Echostar Satellite Corp., 2004 U.S. Dist. LEXIS 20845 (N.D. Ill). *See also* Attig v. DRG, Inc., 2005 U.S. Dist. LEXIS 5183, at *5, n.1 (E.D. Pa. Mar. 30, 2005) (in copyright infringement suit, parties agreed that copies of websites at issue obtained from *archive.org* are admissible evidence); Louis Vuitton Malletier v. Burlington Coat Factory Warehouse Corp., 42 F.3d 532, 535 (2d Cir. 2005) (in trademark infringement suit, evidence of defendant's Web site advertisements presented through *archive.org* capture of the site content at particular time).
27. *Id.* at *16-17.
28. *Id.*
29. *http://www.archive.org/about/faqs.php#The _Wayback_Machine*.
30. *Id.* ("javascript enabled links and actions are disabled in the comparison results to prevent errant scripts from being run").
31. *Id.*
32. This feature of the Wayback Machine is what alerted Healthcare Advocates in the pending lawsuit discussed *supra*, at n.20 that prior versions of its Web site had not been blocked as requested but instead were being accessed by the defendants.
33. *http://www.archive.org/about/faqs.php#The _Wayback_Machine*.
34. *Id.*

J. Murphy, N.H. Hashim & P. O'Connor (2007). "Take me back: Validating the Wayback Machine." *Journal of Computer-Mediated Communication*. 13(1).

Murphy, J., Hashim, N. H., & O'Connor, P. (2007). Take me back: Validating the Wayback Machine. *Journal of Computer-Mediated Communication*, *13*(1), article 4. http://jcmc.indiana.edu/vol13/issue1/murphy.html

Take Me Back: Validating the Wayback Machine

Jamie Murphy
Noor Hazarina Hashim
The University of Western Australia Business School

Peter O'Connor
IMHI, Essec Business School, France

Go to a section in the article:

Abstract

Although fields such as e-commerce, information systems, and computer-mediated communication (CMC) acknowledge the importance of validity, validating research tools or measures in these domains seems the exception rather than the rule. This article extends the concept of validation to one of an emerging genre of web-based tools that provide new measures, the Wayback Machine (WM). Drawing in part on social science tests of validity, the study progresses from testing for and demonstrating the weakest form of validity, face validity, to the more demanding tests for content, predictive, and convergent validity. Finally, the study tests and shows nomological validity, using the diffusion of innovations theory. In line with prior diffusion research, the results of tests for predictive and nomological validity showed significant relationships with organizational characteristics and two WM measures: website age and number of updates. The results help validate these measures and demonstrate the utility of the WM for studying evolving website use.

Introduction

A growing trend is researchers drawing on output from online archival databases such as Google PageRank (Garofalakis, Kappos, & Makris, 2002; Murphy & Scharl, 2007), Google Scholar (Bakkalbasi, Bauer, Glover, & Wang, 2006; Hall, 2006; Kousha & Thelwall, 2007; Pauly & Stergiou, 2005), and several products from Alexa (Palmer, 2002; Ryan, Field, & Olfman, 2003; Thelwall & Wilkinson, 2003; Vaughan & Thelwall, 2003; Veronin, 2002). Most of these studies imply validity, ignore the subject, or note it as a limitation. Yet failure to validate raises issues of trust in research findings (Straub, 1989; Straub, Boudreau, & Gefen, 2004). Computer-mediated communication (CMC) studies validate research instruments such as questionnaires (Koh & Kim, 2003-4; Wade & Nevo, 2005-6), but few address the validity of measures obtained from online tools.

Despite a 1989 call for rigorous instrument validation in management information system research (Straub, 1989), the field has yet "'to reach the point where validation is the rule rather than the exception"' (Boudreau, Gefen, & Straub, 2001, p. 11). Validation is inadequate, in part due to the difficulty in tracking rapid technological changes (Straub, 1989), yet establishing validity is particularly important for new instruments (Bagozzi, 1981; Hinkin, 1995). In addition to validating research instruments such as survey questions, the computer science field acknowledges validating software or expert systems as an important

step in the development of new tools (Kitchenham, Pfleeger, & Fenton, 1995; O'Leary, Goul, Moffitt, & Radwan, 1990). Similarly, social science often validates research instruments, such as the psychometric properties of questionnaire items (Babbie, 1997; Straub et al., 2004), rather than output from online tools. Validating the output from archival databases is an important new challenge.

In the expert systems domain, a review of validation literature found no standard definition of validity and different terms used interchangeably to describe validity (O'Leary et al., 1990). A business research methods text defines validity as the degree to which a research instrument provides adequate coverage of the topic under study (Sekaran, 2003). In computer science and expert systems, validation is the ability of software or a system to comply with defined standards or adequately represent an expert's knowledge (Kitchenham et al., 1995; Mosqueira-Ray & Moret-Bonillo, 2000; O'Leary et al., 1990).

Common to most definitions across disciplines is determining suitability and accuracy. With regard to types of validity, Straub el al. (2004) argued that *predictive* validity was optional, highly recommended *content* and *nomological* validity, and mandated *convergent* validity. New measures, however, require substantiation of predictive, content, and nomological validity (Bagozzi, 1981; Straub et al., 2004).

Apart from a single study that included convergent and nomological validity for three website measures from Alexa—content, download time, and navigation (Palmer, 2002)—to the authors' knowledge, no CMC studies have validated the output from third-party online tools. Comparing Alexa results with jury ratings and a web-based agent, this sole study found significant correlations and suggested that the three measures had convergent validity. The study further suggested nomological validity for website content and website navigation (Palmer, 2002).

The Wayback Machine (WM) from Alexa provides two complementary measures of website evolution: website age and website updates. Scholars have used the WM to investigate archived website content (Brock, 2005; Hackett & Parmanto, 2005; Ryan et al., 2003; Thelwall & Wilkinson, 2003; Veronin, 2002), infer website age (Vaughan & Thelwall, 2003), and study website evolution (Chu, Leung, Van Hui, & Cheung, 2007). While these studies have drawn upon WM measures, to the authors' knowledge only one published article attempts to validate website content, and no studies have attempted to validate website age. Moreover, no known studies have used a third measure provided by the WM: the number of website updates. Thus, the present study:

1. tests the content validity of three measures provided by the Wayback Machine: archived web pages, website age, and website updates;
2. tests the predictive, nomological, and convergent validity of two measures provided by the Wayback Machine: website age and website updates; and
3. adds to the small number of studies validating measures from third party online tools.

The following sections introduce tests of validity, followed by discussion of the Wayback Machine and diffusion of innovations theory. The article then describes the study population. After testing for face and content validity of three WM measures—website content, website age, and website updates—the article uses the study population to test for predictive, nomological, and convergent validity of the latter two measures. The article closes with suggestions for future research directions for academics studying website evolution or using third-party online tools for research.

Literature Review

Validity Tests

As its name implies, *face validity* relates to face value and relies upon experts' personal opinions and judgment. Because of the vagueness and subjectivity that can result, face validity is a weak test of validity, and some researchers question its use (Sekaran, 2003). Given the lack of validation of third-party online tools, checking face validity seems a reasonable first step prior to moving on to more demanding tests. Closely related to face validity is *content validity*, which ensures that a measure includes an adequate and representative set of items to cover a concept. Content validity also relates to sample-population representativeness, for example, the ability of a questionnaire to represent the larger population. When experts agree that a measure provides adequate coverage of a concept, the measure has content validity (Sekaran, 2003).

Predictive validity, also known as practical or concurrent validity, measures how well an independent variable or set of independent variables relates to the characteristics of research interest (Sekaran, 2003). Scholars debate whether predictive validity falls in the general category of *construct validity* (see below) or the extent that the operationalization of a concept actually measures that concept (Straub, 1989). Predictive validity can also show the applied value of research (Straub et al., 2004). For example, a business could predict its online sales based on the number of website visits and email enquiries. To demonstrate validity, the firm could periodically correlate website visits in a particular month with sales in that or subsequent months. Repeatedly high correlations would suggest predictive validity, thus allowing the firm to use website visits to forecast future sales. Depending on the objective, researchers typically use correlation or regression analyses to test such hypothesized relationships (Hinkin, 1995).

Combined with predictive validity, *nomological* and *convergent* validity help achieve *construct* validity—the empirical and theoretical support for a particular interpretation (Straub, 1989). Nomological, or lawful, validity links a theoretical concept with observable results (Cronbach & Meers, 1955). "'If theoretically-derived constructs have been measured with validated instruments and tested against a variety of persons, settings, times, and, in the case of IS research, technologies, then the argument that the constructs themselves are valid becomes more compelling"' (Straub et al., 2004, p. 395). *Convergent validity* results when two variables measuring the same construct correlate highly (Straub et al., 2004). Triangulation of multiple research results, rather than relying on a single line of evidence, helps achieve convergent validity.

The Wayback Machine

The Wayback Machine is part of the Internet Archive (www.archive.org), which amasses websites, moving images, texts, audio, and recently, educational resources (FAQs, 2007). Drawing upon results from the Alexa webcrawler, this U.S.-based non-profit organization permanently stores publicly accessible websites in an enormous digital archive. By preserving human knowledge and artifacts and making its collection available to all, the Internet Archive envisions resembling ancient Egypt's legendary Library of Alexandria (FAQs, 2007). The archive contains snapshots of over 55 billion web pages—more information than in any library including the U.S. Library of Congress—even though archiving began only in 1996. The archive adds about 20 terabytes (10^{12} bytes) of digital content monthly (FAQs, 2007), with each sweep of the estimated 16 million archived websites taking over two months (Howell, 2006).

Via the WM, users can view the original version of each site, as well as the dates and content of subsequent updates. To call up archived websites, users type the URL of the desired site into the address box on the WM homepage. The WM then returns the date of original site creation, number and date of site updates, and links to archived sites. Figure 1 shows the WM homepage, and Figure 2 shows the results for a Malaysian hotel, the Timotel in Mersing. The WM also provides information on site updates. An asterisk beside the dates in Figure 2 indicates more than 50% changes to the website since the last visit.

Figure 1. Homepage of the Internet Archive

Figure 2. Wayback Machine results for www.timotel.com.my

Tracking the evolution of a site can be useful. For example, a researcher could investigate the evolution of *Hyatt.com*'s online customer relationship programs by analyzing consecutive archived versions of the company's site. As noted earlier, researchers have used the WM to track and measure web content development (Chu et al., 2007; Hackett & Parmanto, 2005). The WM is also gaining legal acceptance with

trademark and intellectual property issues (Howell, 2006). In a landmark 2004 U.S. case, the court ruled that pages culled from the WM were admissible as evidence (Gelman, 2004).

Although massive, the WM has limitations. It archives publicly accessible sites written in simple HTML, but has problems archiving password-protected or dynamic sites (Veronin, 2002). Furthermore, sites can decline inclusion by emailing the Internet Archive or using the Standard for Robot Exclusion (see www.robotstxt.org) to specify files or directories not to crawl (FAQs, 2007). Intellectual property owners concerned about infringements on third party sites can also request removal of such content (FAQs, 2007). Any of these actions stops future indexing, removes site content from the archive, and limits the archives' comprehensiveness. Finally, a condition of use of the Alexa webcrawler is that the Internet Archive must wait at least six months after surveying before including site updates in the archive. Coupled with the requisite time to survey the 55 billion archived pages, this results in a time lag of six to 12 months for an archived snapshot to appear (FAQs, 2007; Howell, 2006).

Diffusion of Innovations

As noted above, testing for nomological validity relies upon an established research stream such as diffusion of innovations (Rogers, 2003). Diffusion studies show that technology evolves from simple to complex use at both the individual (Karahanna, Straub, & Chervany, 1999; Pavlou & Fygenson, 2006) and organizational levels (Cooper & Zmud, 1990; Raho, Belohlav, & Fiedler, 1987; Zmud & Apple, 1992).

Since 1994, academics have investigated evolutionary aspects of business Internet use (Teo & Pian, 2004) and noted the importance of understanding the evolution of websites (Hoffman, Novak, & Chatterjee, 1996; Park & Thelwall, 2003). For example, net-based customer service may evolve over five phases—experimentation, value creation, focus, differentiation, and relationships (Piccoli, Brohman, Watson, & Parasuraman, 2004). Alternatively, a longitudinal study proposed that e-commerce websites evolved over four eras, from the pre-web to the integrative web era (Chu et al., 2007).

Examining websites' evolutionary aspects helps researchers investigate what factors lead to successful website implementation, including which features organizations add, and leave, on their websites. Evolution itself, however, is a research limitation; a single evaluation at a single point in time cannot capture such evolution. While longitudinal studies would let researchers track changing relationships, performing multiple evaluations is difficult and cumbersome (Chatterjee, Grewal, & Sambamurthy, 2002). Furthermore, some websites may no longer exist and some changes are ephemeral. For instance, a study of over 1,000 websites across six categories found only two-thirds of the sites still functioning at the same URL five years later (McMillan, 2002).

Another research limitation of diffusion studies is relying upon stated behavior rather than measuring actual behavior (Damanpour, 1991; Rogers, 2003). For example, to measure website age, researchers could email webmasters to ask when their websites first went online. However, a webmaster might not reply, might not know, or might give incorrect information. Domain name age, based on when an organization originally registered its domain name—such as Hyatt hotels registering *Hyatt.com*—provides an actual measure of Internet adoption (Adamic & Huberman, 2000; Murphy, Olaru, & Schegg, 2006). Yet domain name age as a measure of website evolution has limitations. With names registered in the most common domain, *.com*, changes in domain name registrars render the recorded age invalid (Murphy et al., 2006). Similarly, organizations may buy a domain name but wait months or years before hosting a website at that name, thus making the registration date an unreliable measure of when a website went live. Using data from the WM, which archives actual website pages, helps overcome such limitations and establish the real date of site creation.

The study context is Malaysian hotels, for three reasons. Of those industries going online, travel leads other service industries in its share of e-commerce (Dinlersoz & Hernández-Murillo, 2005). Second, hospitality e-commerce studies often draw upon the diffusion of innovations (Matzler, Pechlaner, Abfalter, & Wolf, 2005; Murphy et al., 2006; Murphy, Olaru, Schegg, & Frey, 2003; Wang & Fesenmaier, 2005), and using this research stream facilitates testing for nomological validity. Finally, most tourism Internet research focuses on developed countries; there is a lack of research in developing countries (Frew, 2000; Hashim, Murphy, & Hashim, 2007). This may be partly due to Internet use being at an early and formative phase in developing countries such as Malaysia (Le & Koh, 2002). Studies of Malaysia's hospitality industry are limited, particularly regarding Internet use (Hashim & Murphy, 2007). Using Malaysian hotels contributes to the body of knowledge in these domains, while at the same time achieving the study objectives.

Data Preparation and Preliminary Nomological Results

Testing the content, predictive, convergent, and nomological validity of the WM measures necessitated a database. With no comprehensive list of Malaysian hotel websites available, the study began with 540 hotels registered with Malaysia's Ministry of Tourism, and the Malaysian Accommodation Directory's (MAD) 2003/2004 list of hotel website addresses. In May 2006, keying the 540 hotels' names into Google and Yahoo! helped find more hotel websites and verify the MAD website addresses, yielding 310 websites. The WM failed to give results for 19 sites (about 6%), due to trouble locating the site or the site declining indexing by the Internet Archive. Of the remaining 291 websites, some chain hotels shared the same domain name, such as *hyatt.com* and *hilton.com* for all Hyatt and Hilton hotels in Malaysia. To avoid duplication, excluding 116 hotels with the same domain name left 175 websites. Of these 175 hotels, 96 hotels hosted their website in the global *.com* domain, and 79 hosted their website in Malaysia's country domain, *.my*.

Diffusion of Innovations Findings

Table 1 shows the final sample and suggests that in line with diffusion of innovations research, high rated, chain-affiliated, and large hotels tended to lead in website adoption (Murphy et al., 2003; Siguaw, Enz, & Namiasivayam, 2000; Wei, Ruys, van Hoof, & Combrink, 2001). The first five-star hotel went online almost three years earlier than the first one-star hotel, early 1997 versus late 2000. The first chain hotel went online nearly a year earlier than the first non-affiliated hotel, late 1996 versus mid 1997. Finally, the first online hotel with over 300 rooms was about two years older than the first online hotel with under 200 rooms.

	Websites accessible via the WM	Sample without same domain name	Sample with *.my* domain	First website	Most updates from 1996-2005
Rating					
1-star	4	3	0	10.11.2000	35
2-star	50	31	12	3.11.1999	35
3-star	95	62	28	27.8.1997	63
4-star	71	48	24	22.12.1996	72
5-star	71	30	15	25.1.1997	60
Affiliation					
Chain	205	89	40	22.12.1996	72
Non-chain	86	86	39	27.8.1997	63
No. of					

Rooms					
1-99	70	53	16	1.12.1998	63
100-199	84	50	21	25.1.1997	33
200-299	55	35	22	25.1.1998	56
>299	82	37	20	22.12.1996	72
Total	291	175	79		

Table 1. Sample characteristics

Similarly, high rated, chain-affiliated, and large hotels led in updating their websites. The five-star hotel with the most updates from 1996-2006 changed its site 60 times, compared to 35 times for the leading one-star hotel. Likewise, the leading chain-affiliated large hotel, which was also a large hotel, made 72 updates on its website versus 63 updates for the leading non-affiliated hotel that was also a small hotel. This discussion of website age and number of updates suggests *nomological validity* in line with the diffusion of innovations, but the results are just for one hotel—the leading hotel in each category—and not the entire sample of hotels.

Thus, the next section tests the validity of the Wayback Machine's website age and website updates using the entire sample. Three transformations were necessary prior to testing. A new variable, update frequency, was the website age divided by the total number of website updates. Using this new variable, the most frequently updated website was a three-star independent hotel in Terengganu, which averaged an update every 35 days. At the other extreme, a two-star independent hotel in Melaka updated its website on average once every five years. As update frequency and the number of rooms had an abnormal distribution based on a one-sample Kolmogorov-Smirnoff test, a logarithmic function transformed these two variables into a normal distribution.

Validating the Wayback Machine

The following discussion draws on instrument validation and research of individual measures to validate three measures provided by the WM—website content, website age, and website updates. Starting with the weaker and more subjective tests, this study assessed *face validity* based on published research, feedback from three website managers, and a comparison with Malaysia's domain name database. The courtroom acceptance of the WM (Gelman, 2004; Howell, 2006), mentioned earlier in this article, demonstrates face validity by legal experts. Next, an email invited two Malaysian hoteliers to test their website in the WM. The WM provided archived versions of their sites, and they agreed that the WM provided accurate ages and archived versions. Similarly, an author of this study verified that the WM provided accurate dates and versions of one U.S. and four Australian websites that he managed. The study also examined the face validity of the WM by investigating four hotel homepages shown in a 1996 study (Murphy, Forrest, Wotring, & Brymer). The WM results showed the same homepages as those in the article.

A final test of face validity compared the website age provided by the Wayback Machine with the domain name age provided by Mynic, Malaysia's domain name registrar (whois.mynic.net.my). In principle, a hotel would register a domain name to house the website prior to launching the website. Comparing the WM website age with the domain name age for the 79 hotels using a .my domain name showed that 68 hotels had a domain name age older than the WM website age. Three hotels changed domain names, evidenced by the links and content on archived web pages. For example, the Hotel Flamingo began at www.twosteps.com/flamingo on August 23, 2000 and then changed to www.flamingo.com.my on June 3, 2002. The other eight hotels changed their Mynic information, resetting the registration date on file with

Mynic. These two issues highlight shortcomings of using domain name age as a measure of Internet adoption and provide face validity for the Malaysian hotels' website age.

Content validity was assessed based on the representativeness of websites and adequacy of the website age information provided by WM. As noted above, the WM provided universal coverage for the four sites in the published study and the seven sites managed by three webmasters. Furthermore, as noted in the data preparation section, the WM returned archived versions for 291 of 310 hotel websites, which suggests representativeness. In summary, confirmation by website managers, comparison with a published study, and representation of 291 Malaysian hotels in the WM suggest face and content validity of the WM's website age, website updates, and archived web pages.

Predictive validity stemmed from the number of website updates recorded. Literature on the evolutionary nature of websites (Chatterjee et al., 2002; Chu et al., 2007; Piccoli et al., 2004; Teo & Pian, 2004) led to the prediction that older websites would have a higher average frequency of updates. The result of a one-tailed Pearson correlation test—a significant positive relationship between website age and the logarithmic value of update frequency ($r=.274$, $n=175$, $p<.001$)—shows older websites were updated more frequently and suggests predictive validity.

The diffusion of innovations served as the theoretical base for testing *nomological validity*. This theory argues that certain organizational characteristics relate positively to organizational technology use (Matzler et al., 2005; Wang & Fesenmaier, 2005). U.S. and Swiss studies showed that high rated, large, and affiliated hotels led in technology adoption (Murphy et al., 2003; Siguaw et al., 2000). Compared to lower rated, smaller, or non-affiliated hotels, such hotels had more resources and expertise to facilitate IT implementation. Similarly, emerging Malaysian research and early global research found that large, high rated, and affiliated hotels led in the use of advanced website features (Hashim & Murphy, 2007; Wei et al., 2001). Based on the similarity in these studies, star rating, hotel size, and brand affiliation were the independent variables for testing nomological validity.

Table 2 shows the results of one-way Pearson correlation tests for the logarithmic number of rooms, Spearman correlation tests for star rating, and independent t-tests for chain-affiliation against the dependent variables of website age and number of updates. As mentioned earlier, the analysis used logarithmic values for the update frequency and number of rooms. Given the possible correlation among the three independent variables—size, number of stars, and affiliation—two multiple regression tests examined the predictive importance of the independent variables on website age and number of updates. No independent variables were significant predictors for number of updates, and star rating was a significant predictor of website age ($\beta=.203$, $p=.031$).

Correlation coefficient/ t-value, significance level	Size	Rating	Affiliation
Website age in days	0.161, p=0.017	0.239, p=0.001	2.737, p=0.004
Average update frequency	0.112, p=0.070	0.193, p=0.005	1.775, p=0.039

Table 2. Correlation and T-test results for website age and number of updates (N=175)

Although the low correlation coefficients in Table 2 indicated significant relationships, and the multiple regressions showed low predictive importance, the results were in line with diffusion of innovations research. Larger, higher-rated, and affiliated hotels launched their websites earlier and updated their websites more often than smaller, lower-rated, and non-affiliated hotels did, helping support nomological

validity.

Convergent validity was evaluated by measuring the relationship between domain name age and the creation date of a website at that address. Despite the limitation of a temporal gap between owning a domain name and having a live website, studies use an organization's domain name age as a proxy for Internet adoption (Adamic & Huberman, 2000; Murphy et al., 2006). Although as explained earlier, a domain name age is an imperfect proxy, a high positive correlation between a website's domain name age and that same website's age as provided by the WM would suggest convergent validity.

Establishing the age of names in global domains such as *.com* or *.org*, however, is problematic. On November 30, 1999, the Internet Corporation for Assigned Names and Numbers shifted from a sole domain registrar to a Shared Registration System (SRS) with multiple registrars in the *.com*, *.net*, and *.org* domains (see www.icann.org for a history of domain names). The SRS makes gathering valid global domain name ages unreliable, as companies may change domain registrars, resetting their domain name's creation date and rendering the data invalid (Murphy et al., 2006).

At the country level, however, such as *.at* and *.my* for Austria and Malaysia respectively, gathering the domain name age is less problematic. There is usually just one domain name database for each country, such as in Malaysia. Due to the difficulty validating ages in the *.com* domain, the study used the 79 websites with a *.my* domain to test convergent validity. Eliminating the 11 hotels that changed domains or Mynic information, the result of one-way Pearson correlation for the 68 hotels hosted in *.my* showed a significant positive correlation between website age and domain name age ($r=.933$, $p<.001$). This strong correlation supports convergent validity for the website age provided by the Wayback Machine.

Conclusions and Future Research

Researchers frequently adopt instruments from other studies, which can contribute to flawed measures for at least two reasons. Researchers fail to validate the adopted instrument or make major alterations to a validated instrument without re-testing it (Straub, 1989). This study reinforces the importance of the first reason, failure to validate, for metrics from the growing field of third-party tools such as those provided by Google and Alexa. As researchers continue to use these tools, it is important to address the validity of both the tools and their measures.

This article augments research on the evolutionary and dynamic nature of CMC (Chatterjee et al., 2002; Chu et al., 2007; Hoffman et al., 1996; Murphy et al., 2006; Park & Thelwall, 2003; Piccoli et al., 2004; Teo & Pian, 2004) by suggesting and validating two hitherto underutilized website adoption measures, website age and number of website updates. Website age helps overcomes limitations associated with domain name age and gives researchers a valid temporal measure of website adoption. Furthermore, the archived websites and number of website updates allow researchers and practitioners to study websites over time.

Although the Wayback Machine has limitations such as not indexing some websites, the results of this study showed content validity for three WM measures—website content, website age, and number of updates—as well as predictive, nomological, and convergent validity for website age and number of website updates. This article thus adds to the minimal research on validating online third party tools. Validation studies often deal with a survey instrument or a software process, but results from third party tools such as Alexa seem a new and fruitful area for validation studies.

Future Research

As this study investigated just four types of validation, future research should address other validation tests, as well as the reliability of third party tools (Boudreau et al., 2001; Straub, 1989; Straub et al., 2004). While this article suggests that the WM provides valid website ages and website updates, future research should revisit these two WM metrics in other industries and extend the concept of validation to measures from other web-based third party tools. For example, Alexa provides measures of website popularity and incoming links to a website. Google provides a toolbar that ranks websites on Google's proprietary PageRank, and a beta tool, Google Scholar (scholar.google.com), provides popularity measures for scholarly articles (Bakkalbasi et al., 2006; Hall, 2006; Jacsó, 2005, 2006; Kousha & Thelwall, 2007; Pauly & Stergiou, 2005). While widely used, to the authors' knowledge these tools remain unvalidated.

In addition, this article relied on social science methods to validate measures provided by the WM, rather than validating the WM itself. CMC researchers should draw upon and collaborate with colleagues in computer science and expert systems to apply methods such as evaluation, validation, and verification to the WM (Kitchenham et al., 1995; Mosqueira-Ray & Moret-Bonillo, 2000; O'Leary et al., 1990).

Finally, now that the Wayback Machine seems validated as a viable research tool, an interesting range of research possibilities arise. Researchers can now have greater confidence in the data generated by the tool and can incorporate such data into their research on website development and e-commerce. As suggested elsewhere in this article, the WM facilitates studies of website development over time. Taking a historical perspective and exploiting this opportunity should lead to a better understanding of website evolutions in domains such as e-commerce and Web 2.0.

Acknowledgment

The authors presented an earlier and abridged version of this manuscript at the January 2007 ENTER Conference in Ljubljana, Slovenia.

References

Adamic, L. A., & Huberman, B. A. (2000). The nature of markets in the World Wide Web. *Quarterly Journal of Electronic Commerce, 1* (1), 5-12.

Babbie, E. R. (1997). *The Practice of Social Research* (8th ed.). Belmont, CA: Wadsworth Publishing.

Bagozzi, R. P. (1981). An examination of the validity of two models of attitude. *Multivariate Behavioral Research, 16* (3), 323-359.

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus, and Web of Science. *Biomedical Digital Libraries, 3* (7). Retrieved October 20, 2007 from http://www.bio-diglib.com/content/3/1/7

Boudreau, M.-C., Gefen, D., & Straub, D. W. (2001). Validation in information systems research: A state-of-the-art assessment. *MIS Quarterly, 25* (1), 1-16.

Brock, A. (2005). A belief in humanity is a belief in colored men: Using culture to span the digital divide. *Journal of Computer Mediated Communication, 11* (1), article 17. Retrieved September 18, 2007 from http://jcmc.indiana.edu/vol11/issue11/brock.html

Chatterjee, D., Grewal, R., & Sambamurthy, V. (2002). Shaping up for e-commerce: Institutional enablers of the organizational assimilation of web technologies. *MIS Quarterly, 26* (2), 65-89.

Chu, S.-C., Leung, L. C., Van Hui, Y., & Cheung, W. (2007). Evolution of e-commerce web sites: A conceptual framework and a longitudinal study. *Information and Management, 44* (2), 154-164.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychology Bulletin, 52* (4), 281-302.

Cooper, R. B., & Zmud, R. W. (1990). Information technology implementation research: A technological diffusion approach. *Management Science, 36* (2), 123-139.

Damanpour, F. (1991). Organizational innovation: A meta-analysis of effects of determinants and moderators. *Academy of Management Journal, 34* (3), 555-590.

Dinlersoz, E. M., & Hernández-Murillo, R. (2005). The diffusion of electronic business in the United States. *Federal Reserve Bank of St. Louis Review, 87* (1), 11-34.

FAQs. (2007). *The Wayback Machine: Frequently asked questions*. Retrieved April 5, 2007 from http://www.archive.org/about/faqs.php

Frew, A. J. (2000). Information technology and tourism: A research agenda. Information *Technology and Tourism, 3* (2), 99-110.

Garofalakis, J. G., Kappos, P., & Makris, C. (2002). Improving the performance of web access by bridging global ranking with local page popularity metrics. *Internet Research: Electronic Networking Applications and Policy, 12* (1), 43-54.

Gelman, L. (2004). Internet archive's web page snapshots held admissible as evidence. *Packets, 2* (3). Retrieved October 20, 2007 from http://cyberlaw.stanford.edu/packets002728.shtml

Hackett, S., & Parmanto, B. (2005). A longitudinal evaluation of accessibility: Higher education web sites. *Internet Research, 15* (3), 281-294.

Hall, C. M. (2006). The impact of tourism knowledge: Google Scholar, citations, and the opening up of academic space. *e-Review of Tourism Research, 4* (5), 119-136.

Hashim, N. H., & Murphy, J. (2007). Branding on the web: Evolving domain name usage among Malaysian hotels. *Tourism Management, 28* (2), 621-624.

Hashim, N. H., Murphy, J., & Hashim, N. M. (2007). Islam and online imagery on Malaysian tourist destination websites. *Journal of Computer Mediated Communication, 12* (3), article 16. Retrieved September 18, 2007 from http://jcmc.indiana.edu/vol12/issue13/hashim.html

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21* (5), 967-988.

Hoffman, D. L., Novak, T. P., & Chatterjee, P. (1996). Commercial scenarios for the web: Opportunities and challenges. *Journal of Computer Mediated Communication, 1* (3). Retrieved October 20, 2007 from http://jcmc.indiana.edu/vol1/issue3/hoffman.html

Howell, B. A. (2006). Proving web history: How to use the Internet archive. *Journal of Internet Law, 9* (8), 3-9.

Jacsó, P. (2005, December). *Comparison and analysis of the citedness scores in Web of Science and Google Scholar*. Paper presented at the Proceedings of Digital Libraries: Implementing Strategies and Sharing Experiences: 8th International Conference on Asian Digital Libraries, ICADL, Bangkok, Thailand.

Jacsó, P. (2006). Google Scholar: The pros and the cons. *Online Information Review, 29* (2), 208-214.

Karahanna, E., Straub, D. W., & Chervany, N. L. (1999). Information technology adoption across time: A cross-sectional comparison of pre-adoption and post-adoption beliefs. *MIS Quarterly, 23* (2), 183-213.

Kitchenham, B., Pfleeger, S. L., & Fenton, N. (1995). Towards a framework for software measurement validation. *IEEE Transactions on Software Engineering, 21* (12), 929-944.

Koh, J., & Kim, Y.-G. (2003-4). Sense of virtual community: A conceptual framework and empirical validation. *International Journal of Electronic Commerce, 8* (2), 75-93.

Kousha, K., & Thelwall, M. (2007, in press). Google Scholar citations and Google web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology, 58* (7).

Le, T. T., & Koh, A. C. (2002). A managerial perspective on electronic commerce development in Malaysia. *Electronic Commerce Research, 2* (1/2), 7-29.

Matzler, K., Pechlaner, H., Abfalter, D., & Wolf, M. (2005). Determinants of response to customer e-mail enquiries to hotels: Evidence from Austria. *Tourism Management, 26* (2), 249-259.

McMillan, S. J. (2002). Longevity of websites and interactive advertising communication. *Journal of Interactive Advertising, 2* (2). Retrieved October 20, 2007 from http://www.jiad.org/vol2/no2/mcmillan/

Mosqueira-Ray, E., & Moret-Bonillo, V. (2000). Validation of intelligent systems: A critical study and a tool. *Expert Systems with Applications, 18* (1), 1-16.

Murphy, J., Forrest, E. J., Wotring, C. E., & Brymer, R. A. (1996). Hotel management and marketing on the Internet. *Cornell Hotel and Restaurant Administration Quarterly, 37* (3), 70-82.

Murphy, J., Olaru, D., & Schegg, R. (2006). Investigating the evolution of hotel Internet adoption. *Information Technology and Tourism, 8* (3/4), 161-178.

Murphy, J., Olaru, D., Schegg, R., & Frey, S. (2003). The bandwagon effect: Swiss hotels' website and e-mail management. *Cornell Hotel and Restaurant Administration Quarterly, 44* (1), 71-87.

Murphy, J., & Scharl, A. (2007). An investigation of global versus local online branding. *International Marketing Review, 24* (3), 297-312.

O'Leary, T. J., Goul, M., Moffitt, K. E., & Radwan, A. E. (1990). Validating expert systems. *IEEE Intelligent Systems, 5* (3), 51-58.

Palmer, J. W. (2002). Web site usability, design, and performance metrics. *Information Systems Research, 13* (2), 151-167.

Park, H. W., & Thelwall, M. (2003). Hyperlink analysis of the World Wide Web: A review. *Journal of Computer-Mediated Communication, 8* (4). Retrieved October 20, 2007 from http://jcmc.indiana.edu/vol8/issue4/park.html

Pauly, D., & Stergiou, K. I. (2005). Equivalence of results from two citation analyses: Thomson ISI's citation index and Google's scholar service. *Ethics in Science and Environmental Politics, 2005,* 33-35. Retrieved October 20, 2007 from http://www.int-res.com/articles/esep/2005/E65.pdf

Pavlou, P. A., & Fygenson, M. (2006). Understanding and predicting electronic commerce adoption: An

extension of the theory of planned behavior. *MIS Quarterly, 30* (1), 115-143.

Piccoli, G., Brohman, M. K., Watson, R. T., & Parasuraman, A. (2004). Net-based customer service systems: Evolution and revolution in web site functionalities. *Decision Sciences, 35* (3), 423-455.

Raho, L. E., Belohlav, J. A., & Fiedler, K. D. (1987). Assimilating new technology into the organization: An assessment of McFarlan and McKenney's model. *MIS Quarterly, 11* (1), 47-57.

Rogers, E. M. (2003). *Diffusion of Innovations* (5th ed.). New York: Simon & Schuster.

Ryan, T., Field, R. H. G., & Olfman, L. (2003). The evolution of U.S. state government home pages from 1997 to 2002. *International Journal of Human-Computer Studies, 59* (4), 403-430.

Sekaran, U. (2003). *Research Methods for Business: A Skill Building Approach* (4th ed.). New York: John Wiley & Sons Inc.

Siguaw, J. A., Enz, C. A., & Namiasivayam, K. (2000). Adoption of information technology in U.S. hotels: Strategically driven objectives. *Journal of Travel Research, 39* (November), 192-201.

Straub, D. W. (1989). Validating instruments in MIS research. *MIS Quarterly, 13* (2), 147-169.

Straub, D. W., Boudreau, M.-C., & Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems, 13*, 380-427.

Teo, T., & Pian, Y. (2004). A model for web adoption. *Information and Management, 41* (4), 457-468.

Thelwall, M., & Wilkinson, D. (2003). Three target document range metrics for university websites. *Journal of the American Society for Information Science and Technology, 54* (6), 490-497.

Vaughan, L., & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal web sites? *Journal of the American Society for Information Science and Technology, 54* (1), 29-38.

Veronin, M. A. (2002). Where are they now? A case study of health-related web site attrition. *Journal of Medical Internet Research, 4* (2). Retrieved October 20, 2007 from http://www.jmir.org/2002/2002/e2010/

Wade, M. R., & Nevo, S. (2005-6). Development and validation of a perceptual instrument to measure e-commerce performance. *International Journal of Electronic Commerce, 10* (2), 123-146.

Wang, Y., & Fesenmaier, D. R. (2005). Identifying the success factors of web-based marketing strategy: An investigation of convention and visitors bureaus in the United States. *Journal of Travel Research, 43* (3), 1-11.

Wei, S., Ruys, H. F., van Hoof, H. B., & Combrink, T. E. (2001). Uses of the Internet in the global hotel industry. *Journal of Business Research, 54* (3), 235-241.

Zmud, R. W., & Apple, L. E. (1992). Measuring technology incorporation/infusion. *Journal of Product Innovation Management, 9* (June), 148-155.

About the Authors

Associate Professor Jamie Murphy's background includes complementary industry and academic experience. In addition to owning/managing hospitality businesses, he served as the European Marketing Manager for U.S. sports companies. Dr. Murphy's research focus is effective use of the Internet for citizens, businesses, and governments.

Address: The University of Western Australia Business School, Crawley, WA 6009, Australia

Noor Hazarina Hashim is a Ph.D. candidate in Internet Marketing at the School of Economics and Commerce, University of Western Australia. In Malaysia, she lectures in marketing at the University of Technology Malaysia. Her research interests include the evolution of website and email, and effective Internet marketing.

Address: The University of Western Australia Business School, Crawley, WA 6009, Australia

Peter O'Connor is Professor of Information Systems at Essec Business School France and serves as Academic Director of Institute de Management Hotelier International (IMHI), its specialized MBA program in hospitality management. His primary research, teaching, and consulting interests focus on the use of information technology in the hospitality sector. Previously he held a visiting position at the Cornell School of Hotel Administration and worked in a variety of international positions in hospitality management in sectors ranging from luxury hotels to contract food services.

Address: IMHI, Essec Business School, 95021 Cergy-Pontoise Cedex, France

R. Rogers (in press). *Digital Methods*, Cambridge, MA: MIT Press. (excerpt: The website as archived object.)

3

The website as archived object

That the web arrived as infrastructure awaiting content, as opposed to content awaiting infrastructure is not often appreciated. In the early to mid 1990s websites were under construction and databases were yet to be populated. Sites generally needed filling in. (The same could be said these days of people's profiles on social networking sites, a subject of chapter six. Often fields are empty.) The web's initial emptiness could account for the importance placed upon the precious 'content providers,' a phrase from the web's early period. As noted in the previous chapter, creative encouragement for putting up content came in the form of homespun awards, granted by self-appointed web editors to websites chosen for their quality (see figure one). Once granted, the seal for the site of the week (or similar) typically would be affixed to the winning frontpage, with a link back to the originating awards page. At the awards page, a surfer could view other sites that had earned the same distinction. Awards gradually would be granted by category, such as the best education site award, technical site of the day, coolest science site, shiitake enlightened site, etc.1 To bestow added distinction to them as they proliferated, awards might be branded (*Exploratorium*'s ten cool sites or *Popular Science*'s best of the web) or provided with a provenance (the *original* cool site of the day award). Over time collections of selected sites organized by category became formalized. There are annual awards, modelled after film and TV with an 'academy' that grants them in a ceremony, providing a seal, reciprocal linking as well as an actual statuette (the webby Awards).2

1 Examples taken from author's collection of web awards from the 1990s. The awards are discussed in terms of reliability graphics, and a particular web-epistemology practice, in Rogers, 2000.

2 However historically dominant, the U.S. web awards culture has been joined by other national ones; for the Danish public sector context (with a discussion of the Swedish as well as Norwegian), see Sørum et al., 2009.

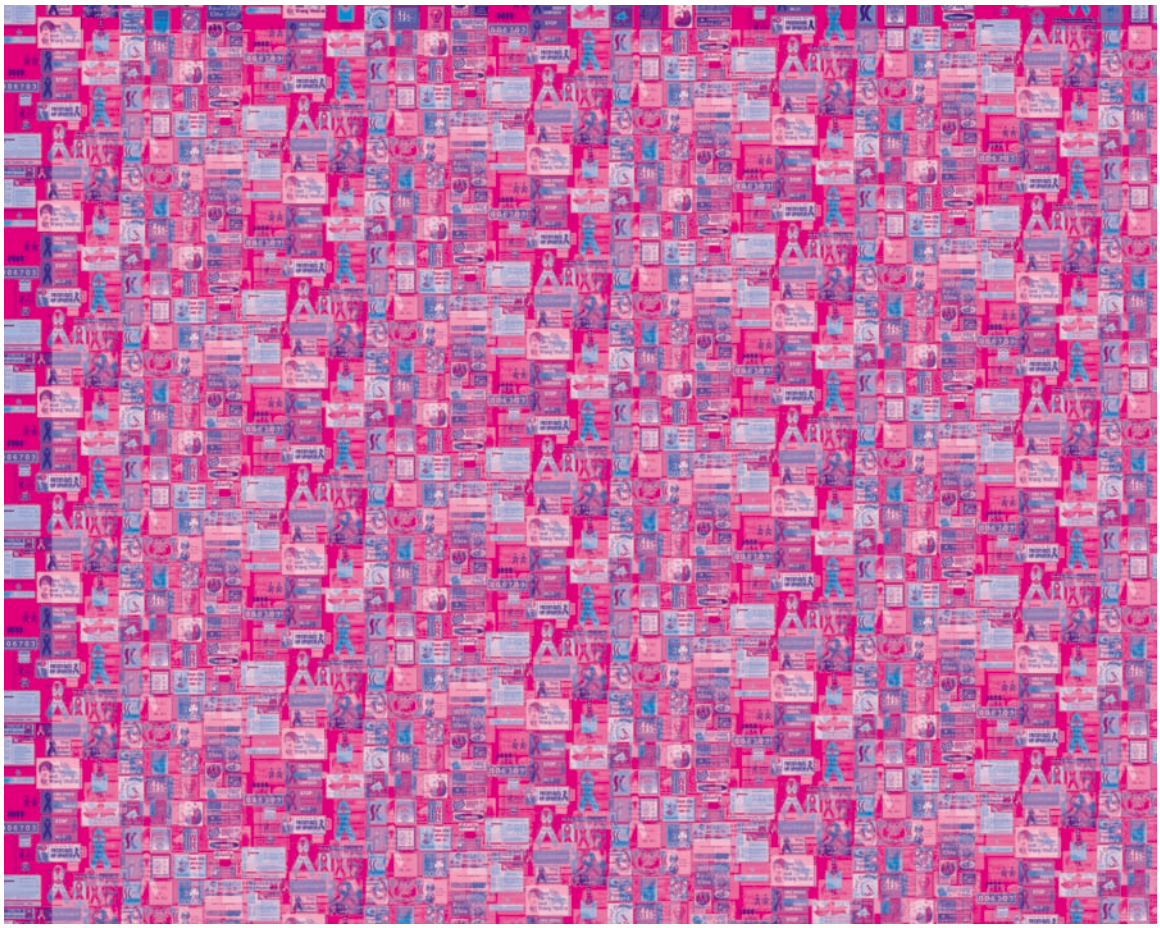


Figure one: Homespun and professional web awards, mid to late 1990s. Collection by the author. Artwork by Anja Lutz, 2000.

A second outgrowth of collections of good websites were the professional link lists (Amnesty International's list of human rights groups, for example) and directories (Yahoo and the Open Directory Project), together with particular methods of website collection-making (such as web archiving), which is the subject of this chapter. Carefully chosen link lists organized by category could be considered the first web guides, or web gazetteers if one thinks in early cybergeographic terms. In the mid-1990s one of the more important listings sites of its kind continually updated an index of worthwhile destinations per content category. One would submit a URL with description to Yahoo (originally "Jerry's Guide to the World Wide web"), so that it would be considered for placement in its directory. (Early search engines also accepted URL submissions; nowadays URL submissions made to Google

are less likely for site inclusion, and more for site removal.3) Online editors browsed and sorted websites. Yahoo as well as the Open Directory Project (originally called "NewHoo"), the volunteer-expert directory, undertook the immense editorial task of choosing, listing and keeping unbroken the links to sites per category. At the same time Yahoo Labs in particular could stake claim to having put into place a new content classification system for the web. To an "Internet cataloger" writing a well-known essay in 1998, Yahoo was making a significant contribution to newfangled online library science not only for its classification scheme but also for the means of content 'navigation' it developed.4 Yahoo's differed from that of a library where each book would be shelved by necessity in one location. At yahoo.com, the resource could be placed in multiple categories, and linked to (and located) from each.

Solliciting, evaluating and categorizing websites – not to mention developing navigation schemes to reach them – could be considered an original form of website analysis. As discussed in the next chapter, the rise of the algorithmic search engine has accompanied the demise of this activity: the large-scale collecting, hand-sorting and display of websites.5 Link list authors, Internet catalogers, directory-makers and all manner of human editors of the web have become beleagured by the search engine. Directories are ill-maintained, or over-commercialized; Amnesty's link list is gone. Googlization, though it has many connotations, could be thought of in terms of the commanding position the search engine has assumed in contemporary website analysis. Prior to that discussion, I first would like to step back and focus on one contribution to website analysis that is still editorial and undertaken at least partly by hand, web archiving.

The archived website and the privileging of content

In certain areas of web studies, the individual website is privileged over other web objects and spaces because that is where the 'content' is. Besides the hyperlink, the search engine, the sphere and the platform, the website is a fundamental organizing unit of the web. It

3 Bercic, 2005.
4 Glassel, 1998; Ellis and Vasconcelos, 1999.
5 There are exception to the overall decline of URL list-making. Whitelists for child's play and blacklists for the purposes of Internet censorship, or content filtering, are actively maintained, as are those compiled to combat spam.

could be considered what the film is to film studies, the television show to television studies. To take the analogy further along, it is the television show (not the television listings) that tends to be saved, just as websites are archived, and not the search engine results that once returned them.6 In other words, one archives the website over the references contained therein (hyperlinks), the systems that delivered them (engines), the ecology in which they may or may not thrive (the sphere) and the pages or accounts contained therein that keep the user actively grooming his or her online profile and status (the platform).

Website archiving is the preserve of old media (if that term may still be used), in the sense of what is privileged, or in fact under-privileged. Archived is the content, stripped of much else. To save its content, the web archivist usually must destroy much of the website. The website is archived without the annotations and other gloss that is written onto it, attaches to it, is embedded in it, or surrounds it. Thus comments may be left out, as they are by search engines, because the comment space often contains a 'no follow' tag instructing crawlers and other indexing devices to leave the area unharvested. Location-aware banner advertisements that are targeted to a particular market place normally are not saved. The same may be said of the more dynamic 'plugged-in' mini-modules such as a social plug-in with lists of friends or Google adwords, both of which update in more of a cascading fashion than the rotation of a billboard banner. Usually, embedded video is not retained in the archived website, for like banners and adwords it is pulled in from another content provider. Surrounding entities such as the complex of relations the website has with cookies as well as with interlacing (ad-serving or surveillance) 'websites' are not captured. One may view these complex relations as a website loads, and ultimately resolves. There may be a series of URLs involved whilst a website loads, be it a tinyurl or bit.ly to begin, redirected to the destination URL, which triggers one or more adservers and the 1x1 pixel market research 'web bugs,' placing or reading cookies and counting impressions. A list of all the URLs that load for a single website is displayed in the brower's activity log.7 These appendages to the website are not visible in the everyday browsing experience, and thus would require consideration of an

6 Weltevrede, 2009.
7 At the Piet Zwart Institute in Rotterdam, Andrea Fiore developed a means to capture and analyze third-party cookies.

archiving practice (capturing cookies, for example) that has a specific research focus (websites' advertising entanglements).

Of all natively digital objects I make mention of web bugs and cookies not to be overly obscure, but rather to point out that the question of where the website begins and ends – which is a classic one in web archiving discourses – is a piece with the media theory and historiography that accompanies the practice of archiving, as I come to shortly. Generally speaking, the archived website ends nowadays with the content put up by the site author. In the archiving, that content is freed from the commercial support system (or political economy), second and third-party material (intellectual property of others) as well as the social apparatus and the talkback (friends' recommendations and visitors' comments). In a sense, the 'new media' elements (cookies, embedded material, recommendations, comments, etc.) are eliminated for posterity, and a traditional content container, looking somewhat broken for its missing pieces, remains as the 'archived website.'

Surfing the web as it was

The web archiving scholar, Niels Brügger, has written: "[U]nlike other well-known media, the Internet does not simply exist in a form suited to being archived, but rather is first formed as an object of study in the archiving, and it is formed differently depending on who does the archiving, when, and for what purpose".8 That the object of study is constructed by the means by which it is 'tamed' and captured by method and technique is a classic point from the sociology and philosophy of science and elsewhere.9 Indeed, one may make web archives into objects of study in themselves, beginning with the first and still most significant one of its kind, the Internet archive (archive.org), and the Wayback Machine (waybackmachine.org), which is its interface, and also its primary and most well-known means of querying and navigating its contents.10 Following Brügger of importance here is how a web archive as an object, formed by the archiving process, embeds particular preferences for how it is used, and for the type of research performed with it. Which

8 Brügger, 2005.
9 Latour and Woolgar, 1986; Knorr-Cetina, 1999; Walker, 2005.
10 Kahle, 1997; Lyman and Kahle, 1998. One could make the distinction between the Internet archive (as repository) and the 'current' interface on it (the navigation). The URL for the interface is http://waybackmachine.org/.

research practices are invited by the specific form assumed by the Internet Archive, and which are precluded?

When one uses the Internet Archive (archive.org), what stands out for everyday web users accustomed to search engines, is not so much the achievement of the very existence of an archived Internet, which in itself is remarkable. Rather, the user is struck by how it is queried via the Wayback Machine. The search box contains an http:// prompt; one enters a single URL, not key words, into the search box, and returned is a list of stored pages associated with the URL from the past, either in a table with columns (in the classic version), or in a calendar mode (in the newer version). Next to a date an asterisk indicates that the archived page is different from the one previously archived (in the classic version), which is important for researchers interested in capturing and studying website evolution, as an approach to the study of the website as archived object, as I come to.

The Internet archive came into being in 1996, and its interface and content navigation system, the Wayback Machine, in 2001, though archived websites had been available for viewing earlier through the Alexa toolbar, which indicated if an archived version of a site were available when one came upon a 404 error, or page not found. In other words, originally the Alexa toolbar, in tandem with the Internet Archive, was the solution to the broken link, and to interruption in surfing. Arguably the entire means of navigation of the Internet Archive in the Wayback Machine derives from a flow principle. In keeping with the principle, it also preserves the Internet as a 'cyberspace', which one navigates seamlessly. I also would argue that the Wayback Machine's construction furnishes an experience of web history, "surf[ing] the web as it was," as its motto reads, more than it provides a means to study it. Indeed, surfing is arguably a model of web usage from the 1990s that has faded in practice, and been supplanted by search, and perhaps 'wilfing,' a British acronym for 'what was I looking for' that also references ideas about the impact of the web and search engines on cognition more generally.11 At the Wayback Machine, preserving surfing is manner of doing web history; in fact it also makes history in the sense that the surfing is sometimes smoother in the Wayback Machine than it was on the web, when links were often broken.

11 Shirky, 2005; Lewis, 2007; Carr, 2008.

The Wayback Machine embraces continuous flow (click-through) over interruption and pages not found by what I would call 'atemporal linking.'12 By atemporal linking I mean that sites linked to one another may not share the same 'periodicity,' a term for a bounded timeframe employed in scholarly web archiving circles (e.g., the few months of a media attention cycle for a major disaster, or the campaigning season for elections).13 In the event, radio buttons, animated gifs and starry night backgrounds may meet big buttons and tag clouds, all in the same surfer's path. Once the available pages of the queried URL are loaded, one may click through the pages returned, and onto other pages of other sites. When a user clicks a link, the page nearest to the date of the originating page is loaded; if there is no archived page available, the Wayback Machine will access the live web page instead. That is, the links from one site to another always 'work.'

> Not every date for every site archived is 100% complete. When you are surfing an incomplete archived site the Wayback Machine will grab the closest available date to the one you are in for the links that are missing. In the event that we do not have the link archived at all, the Wayback Machine will look for the link on the live web and grab it if available.14

By loading pages closest in date to the ones surfed away from or by connecting to the live web, the Wayback Machine, with its atemporal linking, 'jump-cuts' through time, thus providing the continuous flow of surfing, and preserving the web as cyberspace (and improving upon the 'old' cyberspace). What else may one do with the Internet archive apart from surfing it, and thus reliving the web as cyberspace?

Website biography as historiographical approach embedded in the Wayback Machine

I would like to introduce thought about the Internet Archive (and particularly the Wayback Machine) as not only presenting a particular history of the web, but also representing a specific historiography: the single-site history, or the site biography. In effect, the Internet

12 Galloway, 2004; Sterling, 2010.
13 Schneider et al., 2003.
14 Internet Archive, 2008.

Archive, through the interface of the Wayback Machine, has organized the story of the web, for the researcher, into the histories of single websites. With the current form assumed by the Wayback Machine, one can study the evolution of a single page (or multiple pages) over time, for example, by collecting snapshots from the dates that a page has been indexed, and playing them back like time-lapsed photography. (The outcomes of such an approach are discussed briefly below.)

One also can go back in time to a page for evidentiary purposes, which appears to be a primary use case, according to the literature.15 Scenarios of using the Internet archive in the evidentiary arena include instances of intellectual property infringement as well as trademark infringement through practices as so-called cybersquatting and typosquatting. In patent cases, alleged novelty may be harmed by prior art found online.16 The archive also would aid in retrieving missing web citations in law as well as medical journals, a lament with a literature describing the decay rate of links in recent journal articles, also known as accelerating link rot.17

Outside of the evidentary arena, what would comprise a website biography? One could peruse the public records for ownership (a genealogical approach), and begin with the birth of the website, and follow its life as documented records. Sites are not only *sui generis*, but may be adopted, poached, sold and resold; they also may be vandalized, attacked and put out of service. URLs may have had websites that violated guidelines of search engines, or countries practicing censorship, and were blacklisted. They may have been purchased, and never used. Of interest in this context is Constant Dullaart's hand-made collection of parked websites, with generic templates and content, awaiting owners; suggestedomain.com is where his repository of parked sites loops.18 Obtaining the log files for a site may be of interest to researchers desiring to know about the patterns of visitation; by default hit and referral logs are often erased monthly and a site owner may have only the past twelve months on file.

15 Howell, 2006.
16 Rogers, 2007.
17 Rumsey, 2002; Carnevale and Aronsky, 2007.
18 Dullaart, 2010.

Thus in the genealogical approach, sites come furnished with (historical 'who was') records as well as (short-lived) analytics data in the form of logs.

One could take a 'layer' approach (in the sense of graphical or image editing software), akin to "web2diZZaster" (2007) by the media artist, sumoto.iki, who stripped webpages of their content so that only the underlying templates and formats remain (see figure two). It is critical work in that sumoto.iki evacuated web 2.0 sites of their user-generated content, revealing the sites' emptiness without "users like you."19 More radically it shows the effects of the dying out of the bees (the 'disaster' in the title of the work) in what has come to be known as the "worker bee economy" that is web 2.0.20 Other artistic research on the anatomy of a website is Hendrik-Jan Grievink's "Template Culture: Form Follows Format" (2010), an exhibition of well-known company sites, reduced to their templates (see figure three). Here one peels websites, like proverbial onions, revealing the commonalities in form and structure, and in the critical mode, an underlying sameness or blandness.

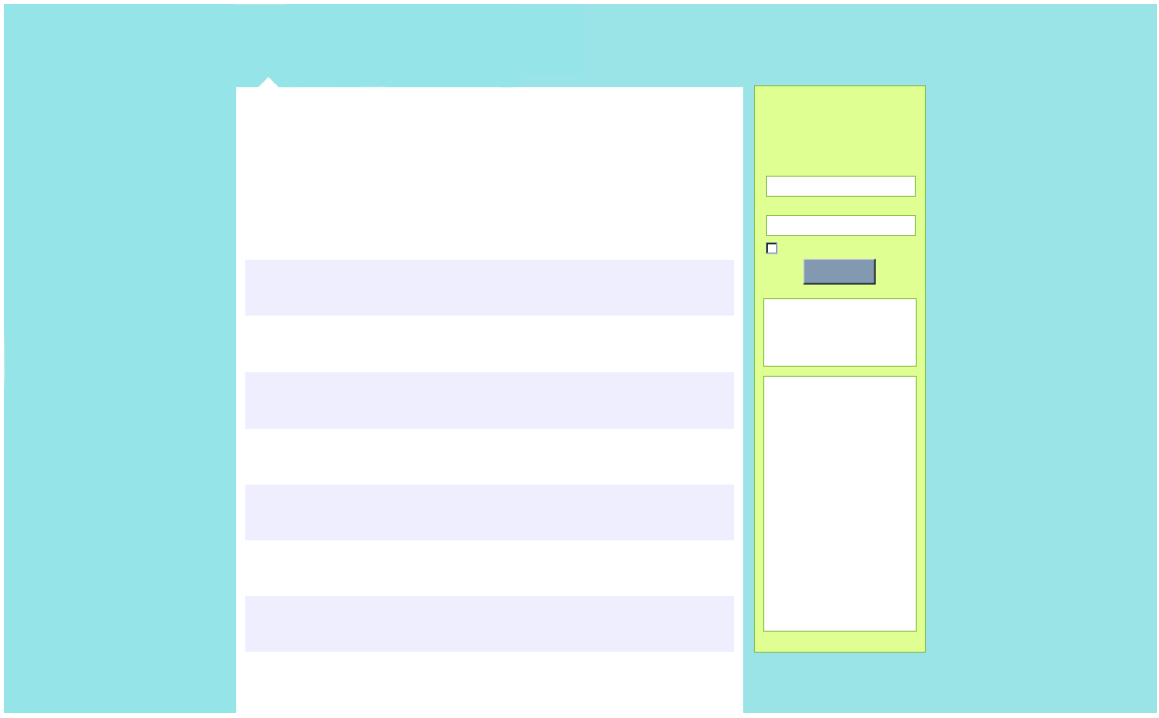


Figure two: Twitter stripped to template. web2diZZaster by sumoto.iki, 2007.

19 Gehl, 2010; van Dijck, 2009. Sumoto.iki depopulated the following of content: Delicious, Digg, Last.fm, Technorati, YouTube, Myspace, 43Things, Twitter, Facebook and Netvibes.
20 Moulier-Boutang, 2008.

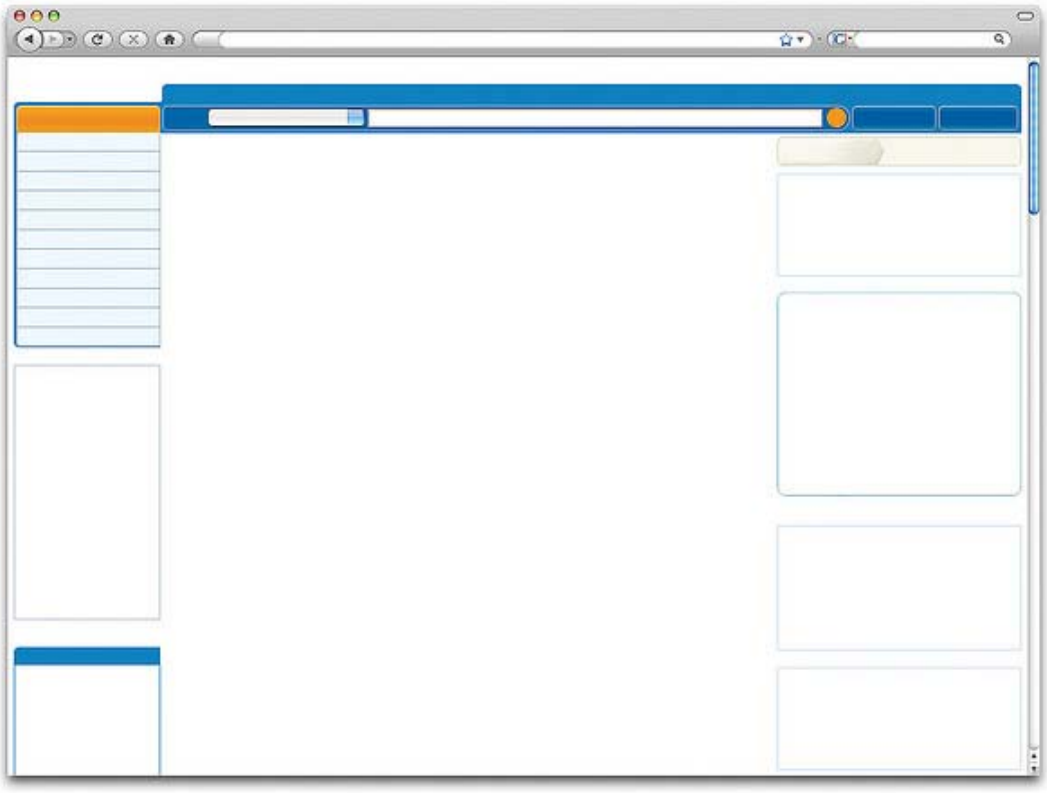


Figure three: Selection from *Template Culture: Form Follows Format* by Hendrik-Jan Grievink, 2010. Template shown is Amazon.com's.

A related, albeit more social scientific, approach to the manual study of the website is 'feature analysis,' where one creates a codebook of all or as many possible website features, and checks a set of sites for presence or absence of them, creating a features matrix.21 Sites are scrutinized for the prominence or obscurity of features, too. As mentioned in the opening chapter, eye-tracking shows that western readers are attracted to the upper left portion of a website, so prominence may be thought of in terms of placement on the page; any features residing 'below the fold,' which is beneath the browser window and reachable only by scrolling, are considered obscured. A web page's advertisement real estate provides a guide to placement and prominence analysis. Here one may compare traditional newspaper analysis (units such as headline size, column inches) to their counterparts online.

21 Ryan et al., 2003.

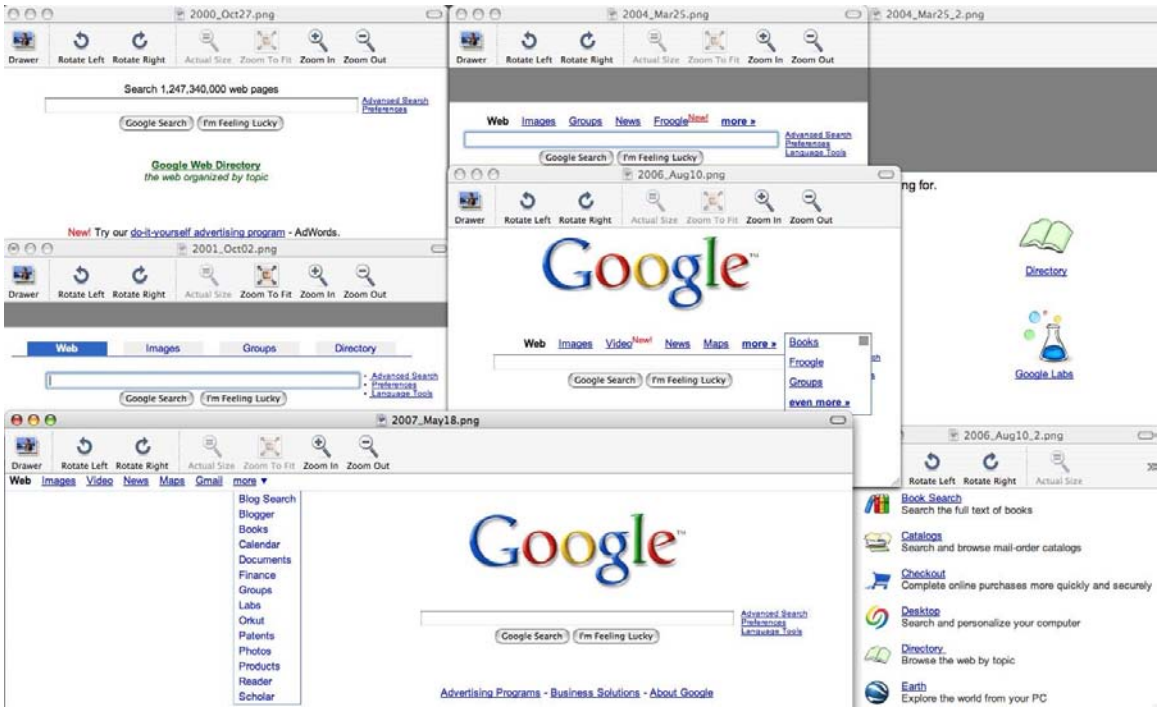


Figure four: The Demise of the directory: web librarian work gradually demoted in Google, 1998-2007. Screenshot collection and analysis by author, Digital Methods Initiative, 2008. Source: Wayback Machine at web.archive.org.

Once time is introduced to the above types of analysis (and others), the Wayback Machine becomes compelling for website biography. The practice of making a movie of a website, in the style of time-lapsed photography, originates with the pioneering "Heavy Metal Umlaut," which is the story of the evolution of a Wikipedia entry, and likely one of the earliest "documentary screencasts."22 It is instructive for its narrative, beginning as it does with the overall story of the growth and professionalization of a once amateurish encyclopedic entry (on a subcultural practice), and subsequently focusing on a few storylines, including the struggle to 'typeset' the heavy metal umlaut online, and the vigilance of the article authors when page vandalism strikes. The movie was made by screencapturing the history of the revision edits to the article (clicking through Wikipedia itself), and providing a voiceover track.23 In the following, I relate (briefly) the making of a screencast documentary, not of a

22 Udell, 2005a.
23 Udell, 2005b.

Wikipedia entry, with its history conveniently embeded in the wiki, but of a website, using the output of the Wayback Machine.

"Google and the politics of tabs," the movie, is an alternative history to Google's own 10-year anniversary timeline.24 As I noted earlier, it is the story of the demise of the Internet cataloguer, and the human editors of the web, which can be dated March 2004, when the once well-placed "directory" was removed from the frontpage real estate at google.com. The movie also provides a method for using the output of the Wayback Machine (the google.com pages bearing an astericks). In doing so it follows the dominant medium device (organizing the web into single site histories), and repurposes its output for social study (demise of the online librarian).

All the available and unique pages from http://www.google.com were captured from the Wayback Machine and made into a movie as well as an info-graphic.25 The analysis focused on the area of the interface above the search box – the tabs – examining which search services (web, images, maps, news, etc.) have been privileged by Google over time on its front-page tabs, where the further to the left the more preferred the placement of the service. It was found that the "directory," the human-edited project by the Open Directory Project (dmoz.org), enjoyed front-page status (third tab from the left) on Google from March 2000 until March 2004, when it was degraded and placed under the "more" button. By August of 2006 the directory had been moved from under the "more" button to under "even more," and in May 2007 it was removed entirely from the menu of search services, which by that time had moved upper left on the Google frontpage. One had to search Google to find Google's directory, as the movie concludes. The history, or screencast documentary, provides a long view (a decade in web history) of the decline of the significance of Internet cataloguers and web librarians, generally, and the rise of web information organization by algorithm rather than by hand, which is the subject of the next chapter.

24 Google, 2008.
25 Digital Methods Initiative, 2008.

In all, the Wayback Machine makes the Internet Archive, and the web, into surfable history, with pages atemporally linked from multiple points in time so as to preserve surfing the web as it was, and providing a solution to broken links, the 404 error. Arguably the archive has a content management system (broadly speaking) with more of a web user, than a web researcher in mind. For the latter, it invites research that captures and interprets single-site histories, or significant changes to single pages of websites. One now can replay sites at waybackmachine.org, and through the method described above turn them into a movie (screencast documentary), narrating the history of the website as the history of the web, in the life and times approach of the biographical tradition, among other pursuits, as I come to.

Apart from whois geneologies, anatomies, features analysis and interface politics and epistemology, one may capture and interpret changes in substance on a website, that is, shifting priorities and commitments of the individual, group, organization or institution that runs the site. Here it is not structures or features that are analyzed but rather the substance of the main menu – lists of issues, campaigns, missions, slogans, services, products, etc. that reside on the front page and organize the content of the website. For example, we have captured and loaded into a movie the historical homepages of whitehouse.gov, concentrating in particular on the issue list, which is one of the substantive menu items. It is a study of the gradual appearance of the word 'security' in the issue language used after 9/11, reaching its height one year later in September 2002, when all issues on the White House agenda (as seen on whitehouse.gov) were security ones: 'national security,' 'homeland security,' and 'economic security' (see Table One). All remaining issues were under placed under a 'more' button, showing their demotion in standing at that time by the U.S. White House under the George W. Bush administration (2000-2008).

Table One. Up and Down with 'Security' as Prominent Issue Language at Whitehouse.gov, September 2001 – September 2009

September 28 2001	Faith-Based and	Homeland Security
Education	Community	Economic Security
Tax Relief		More Issues
Defense		
Social Security	28 September 2002	1 October 2003
Medicare	National Security	Medicare

Iraq
National Security
Economic Security
Homeland Security
More Issues

28 September 2004
Economy
Iraq
Education
National Security
Homeland Security
More Issues

28 September 2005
Hurricane Relief
Homeland Security
Judicial Nominations
National Security
Renewal in Iraq
Jobs and Economy
Social Security
More Issues

29 September 2006
Budget Management
Education
Energy
Health Care
Homeland Security
Hurricanes
Immigration
Jobs and Economy
Judicial Nominations
Medicare
Middle East
National Security
Pandemic Flu
Patriot Act
Renewal in Iraq
Social Security
More Issues

26 September 2007
Budget Management
Defense
Economy
Education
Energy
Environment
Global Diplomacy
Gulf Coast
Health Care
Homeland Security
Immigration
Iraq
Judicial Nominations
Medicare
National Security
Pandemic Flu
Patriot Act
Veterans
More Issues

2 October 2008
Afghanistan
Africa
Budget Management
Defense
Economy
Education
Energy
Environment
Global Diplomacy
Health Care
Homeland Security
Immigration
International Trade
Iraq
Judicial Nominations
Middle East
National Security
Veterans
More Issues

27 September 2009
Civil Rights
Defense
Disabilities
Economy
Education
Energy and Environment
Ethics
Family
Fiscal Responsibility
Foreign Policy
Health Care
Homeland Security
Immigration
Poverty
Rural
Seniors and Social
Security
Service
Taxes
Technology
Urban Policy
Veterans
Women
Additional Issues

In the opening chapter, I discussed the research practice of reading websites, which in the discussion so far only menu substance analysis approximates. The reference earlier was to the use of the Internet archive by Dutch investigative journalists, who hand-picked over one hundred right-wing and right-wing extremist websites via the Wayback Machine, and read the changes to their contents in the past ten years. Their approach was word choice analysis;

given a range of equivalents, was the harsher term employed, one that was more extremist? They found that the right-wing sites gradually began to align in tone and sentiment with the right-wing extremist sites. Dutch society appeared 'hardening' over the course of the years since the assassinations of Pim Fortuyn and Theo van Gogh, and that impression was made more solid not through 'going native,' visiting the pamphlet library or interviewing extremism experts (however valuable), but rather through compiling a list of websites and manually analyzing them. As I mentioned at the outset, that the Internet could be used to ground a claim about a societal condition was not only surprising for those of us familiar with its study as cyberspace and cyberculture, and with ideas of virtual life as distinctive and separate, however much they have been contested empirically. The analysis that confirmed a shift in the language of the right-wing towards extremism also led to the notion of 'online groundedness'; one could ground claims through web website analysis, and seek to apply them beyond the legal (evidentiary) arena. It is worthwhile to emphasize that the analysis was performed by making a list of websites – from the past. Indeed, one could think of it as a new kind of link list to the archive, and imagine it as a web compilation or even a special collection, entitled Dutch right-wing and right-wing extremist sites, 1997-2007. In other words, as an analytical strategy one would make a list of thematic or period websites *already archived*, and provide a means of accessing, querying and otherwise analyzing them – an approach to the website as archived object with which I will conclude. Before discussing results from analyses made through collections of previously archived websites, and especially from conjuring up a "past state of the web" – which is the specific contribution made here – I would like to consider the larger question of website special collections, a fledgling area with a manual approach to website analysis.

From biographical to event-based and national historiographies

The suggested citation for the collections of web archives at the Library of Congress (LOC), "Archived in the Library of Congress web Archives," returned very few results in Google Scholar, Google web or Google book search, with the exception of pages from the Library of Congress website itself and a smattering of other sites.1 Virtually no one references the

1 On 19 April 2011 in Google book and web search the sole scholarly reference to the LOC's web archives is Self, 2009. No references were returned in Google scholar. Apart from that of Kirsten Foot, Steven Schneider and colleagues, notable research which has

dozen special web archives as primary source material in their scholarly or non-scholarly publications, at least in the LOC's preferred style, according to the dominant search engine. The problem posed at the opening of this chapter from the early days of the web has been inverted; there is precious content now awaiting users, like books in libraries awaiting borrowers.

While the web archives are under referenced, the Internet Archive itself as well as its Wayback Machine are well-cited. Queries for the "Wayback Machine" and the "Internet Archive" in the search engine return copious results. The vast majority of the references is to information and library science pieces about the methods and techniques of web archiving, including (on occasion) to certain critiques of their biases towards western sources and subject matters – an observation made of Wikipedia, too.2 web archiving infrastructure receives scholarly and non-scholarly attention; the archived materials – the primary source material – gain less notice.

The question of the lack of "researcher engagement with web archives" has been taken up by web archiving scholars, where one of the more poignant observations concerned the kind of web to be archived in the first instance, and in future, so that the materials would be used.3 According to one observation, archives may be more attractive (to humanities scholars) if they were made up of digitized materials, e.g., websites with photographs, personal letters and other materials from World War II. In the event, websites containing primarily digitized materials have been archived. Here history and web history divide, or become separate objects of study. The web becomes a delivery mechanism for 'old media' – albeit with vintage html code enframing it.

As an approach to the selection of materials to be archived, saving websites containing digitized historical media has its practitioners. In the event, there is a "single site" collection

made use of the Internet Archive includes Ryan et al., 2003; Brock, 2005; and Hacket and Parmanto, 2005.
2 Thelwall and Vaughan, 2004.
3 Dougherty et al., 2010. For an earlier effort, see Arms et al., 2006.

at the Library of Congress web archives, one of the special collections of web archives.4 Saved in this collection are 23 individual websites, many of which are themselves online archives of military history materials, making the special collection into a double container. These are website archives of digitized archival materials.

Apart from the single site set, the special collections of web archives at the Library of Congress include ones on (in alphabetical order) the Crisis in Darfur, Sudan, 2006; Iraq War 2003; Papal Transition 2005; September 11, 2001; United States 107th Congress; 108th Congress; United States Election 2000; 2002; 2004; 2006; and 2008. At the Internet archive there are four other special collections, on the Asian Tsunami (2004-2005), Hurricanes Katrina and Rita, U.K. national archives as well as web Pioneers, all of which were undertaken by the Internet archive without collaboration with the Library of Congress, and at the time of writing appear somewhat abandoned with 404 page not found errors when loading the Asian Tsunami as well as web Pioneers collections. Archiving activity reaches only to 2006, when many of the LOC's collections also end. Links to archives made so that links do not break themselves are broken.

The U.K. national archives pointer links through to the more recent special collections at the U.K. governmental site, including Volcanic ash cloud (2010), U.K. national budgets (March and June 2010), Financial crisis (2008), Swine influenza (2009) and the 2012 Olympic and Paralympic Games and Cultural Olympiad. If one were to characterize the special collections generally, they appear to embody a second historiographical approach to web archiving: event-based history. Indeed, to a leading handbook on web archiving, it has become an established pursuit to capture "events of importance, such as elections or disasters."5 This historiographical commitment derives from the work of the pioneering webarchivist.org project by Steven Schneider and Kirsten Foot, who, together with collaborators, have created a series of special collections of websites ("web spheres"), beginning with the 2000

4 Passion and eclecticism are also on display in the same collection of websites. Apart from those on military as well as African-American history, two of the 23 single sites saved are those of the Hungarian national bank and the coins and currencies collection at Notre Dame University.
5 PoWR, 2008: 19.

U.S. elections.6 "September 11, 2001," as the Library of Congress lists it, is perhaps the most well-known of the collections, and together with their efforts in archiving the 2002 U.S. elections and the Asian Tsunami of late 2004, established the web archiving tradition of histories of elections and disasters.7 Events arguably pose the greatest challenges for archivists, and at the same time also create the "archive fever" for the urgency of the undertaking, as content is continuously being lost to posterity through the combination of the ephemeral nature of web content generally and rapidly changing websites during events more specifically.8 Without rapid steps taken, content is forever lost. In the case of the September 11 archive, the archivists were putting the necessary pieces together to archive the 2002 U.S. national election websites, when the attacks on the World Trade Center took place. They were well positioned so as to begin the special practice of creating what they call a "web sphere," which is treated in multiple articles by the webarchivist authors as well as a small circle of scholars engaged in the specialty area of web archiving. Foot and Schneider are remarkably consistent in their definition of a websphere, which is also a method and research practice. In the original piece of scholarship, they write that "a web sphere [is] a hyperlinked set of dynamically defined digital resources spanning multiple web sites relevant to a central theme or 'object.' The boundaries (…) are delimited by a shared object-orientation and a temporal framework."9

In the seminal as well as in successive articles, the research practice is also laid out.10 The web sphere crucially is dynamic in two senses, for the archivists continually locate new websites (or web resources) to be included, and websites continually point to other websites (either new ones or previously unknown ones) which are relevant to the theme. The web sphere is bounded by the theme as well as by a temporal dimension ("periodicity"), which could be thought of as its coverage span or attention cycle (in traditional terms) of the event. The actual research practice of collecting the websites could be characterized as a snowball method, updated for the web. Editors find URLs through searching and surfing the links

6 Foot and Schneider, 2002.
7 Foot et al., 2003; Foot and Schneider, 2004; Wu and Heok, 2006.
8 Derrida, 1996; Veronin, 2002.
9 Foot and Schneider, 2002: 225.
10 Schneider and Foot, 2002; Foot et al., 2003; Schneider and Foot, 2004; Foot, 2006; Foot and Schneider, 2006.

between the thematically related websites; URLs are also recommended to them through crowd-sourcing, and checked for inclusion. websites are subsequently tagged or otherwise annotated so as to create metadata. They are also categorized into site types, and analyzed for features.

The radical nature of their approach to the selection of materials to be archived (the dynamically evolving collection) is to be appreciated when contrasted with a third archiving method and embedded historiography. In the list of U.S. and U.K. special collections above, one also may take note of the emergence of a normal archiving practice (in the Kuhnian sense of normal science), now applied to the web: the keeping of records for the purposes of national history. Indeed, as the Internet Archive as well as special collection-makers using the web sphere method cede their position as the major archivists of the web, in terms of sheer number of projects, national libraries are creating lists of websites to be saved. At the time of writing, the National Library of the Netherlands, for example, is regularly archiving 998 websites.11 The actual quantity of websites archived, approximately 1,000 which is up from the original 100 that were being saved, is an artificial round number that opens up questions of how to pick and hand-sort the websites to be kept for national history purposes, not to mention how many sites to keep. (The web sphere approach would not result in not in round but in squiggly numbers.)

To begin with, the criteria of what constitutes a Dutch website are of interest here, in order to appreciate why websites are still analyzed manually. Following similar definitions of a national website from archiving projects in other European countries, the National Library defines a website as Dutch if it meets certain tests. What is a 'Dutch website'? It is a Dutch website, if it is:

1) Dutch language, and registered in the Netherlands;
2) Any language, and registered in the Netherlands;
3) Dutch language, registered outside the Netherlands; or
4) Any language, registered outside the Netherlands, with subject matter related to the

11 Personal correspondence with Caroline van Wijk, National Library of the Netherlands, 27 May 2009.

Netherlands12

There are national registrars of country domain names, so that each website registered in the Netherlands as .nl is known, in principle. There are libraries (in a software sense) for detecting automatically the language of a website, so one could differentiate for sorting purposes a site in Dutch and a site not in Dutch (that is, between the first and second criteria). Given a very large collection of websites (for example the Internet Archive's collection as well as the French or Danish National Library's open-ended trawls), one could detect Dutch-language sites outside of the .nl domain, the third criteria, and filter out Belgian (and Flemish) sites if they are .be.13 To classify those remaining Dutch/Flemish language sites (which are neither .nl nor .be) would require a manual intervention. Indeed, that is where the automated identificaiton and sorting would end. To identify for archiving purposes a website of any language, registered outside of the Netherlands, with a subject matter related to the Netherlands requires reading websites.

In the realm of web archiving at least, the Internet cataloguers, web librarians, link list builders and other web editors have defined the Dutch website – their object of archiving, in the sense of Brügger above – so as to require a manual approach. The web archiving handbook I referenced above recommends the formulation of a collection policy and a collection list, which contextualizes further the example of the Netherlands above, where the definition of a Dutch website would be related to the collection policy of archiving the Dutch web, and the 998 sites would be the selection or list (see appendix for the list of URLs).14 The sites that are typically archived are governmental, national cultural and higher education – a kind of pre-web establishment.

One purpose of thinking through the consequences of manual practices of website analysis concerns the kind of webs we are left with once archived, and the kind of research we are able to perform with them, as I have discussed in historiographical terms: single-site histories or biographies, event-based history and national history. When critiquing the practice of

12 Weltevrede, 2009.
13 There is also language-specific crawling. See Somboonviwat et al., 2006.
14 Pinsent et al., 2008.

Dutch web archiving, as a scholar in the Netherlands, and particularly the actual results (998 websites archived out of 3.5 million .nl sites and an unknown number of the other 'Dutch' sites), I would like to recall the nary-a-care archiving by the Internet archive in the 1990s and early 2000s.15 As Brewster Kahle, the founder of the Internet Archive, put it in a 1996 *Wired* article: "I usually work on projects from the you've-got-to-be-crazy stage," by which he meant envisaging to archive the entire web, or as much of it as possible. As I argued in the opening chapter, the end of cyberspace as virtual realm apart and the rise of the institutional and regulatory frameworks for the Internet have not been kind to web archiving. They also have 'damaged' the Archive. To process the quantity of requests to be removed from the Archive (and the Wayback Machine), the decision was taken to interpret robots.txt, the robot or crawler exclusion code that may be built into a website, to mean that the site prefers to be left out of the Internet Archive all together, even those pages that were previously in there, prior to the placement of robots.txt, or prior to the current ownership of the website domain. As I noted above, there are far more requests these days (also to Google) to have websites removed from storage (even in temporary caches) than for them to be included (as in the web directories, guides and awards pages of old).

Conjuring up a past state of the web

In keeping with the overall digital methods principles, colleagues and I approached the Internet Archive by considering how to repurpose its ordering device (the Wayback Machine) for social research. As discussed, the first outcome followed the output of the Wayback Machine (lists of pages of a single site from the past). Site histories are captured by retaining only the pages with changes to them (the ones marked with astericks), and loading them for playback in a movie. What could one learn from the history of a single website, apart from seeking evidence from it for legal proceedings? With the screencast documentary making (or Wayback Machine movie-making), a time dimension is added to website analysis.

---

15 The critique I refer to is a short speech I gave about Dutch national web archiving on the occasion of the retirement of Eric Ketelaar, the Professor of Archive Science at the University of Amsterdam, in May 2009. It was entitled "998 websites," which was the number of Dutch websites archived to date by the National Library of the Netherlands. An accompanying slide showed this math: 998 / 3364922 = 0.000296 of .nl websites archived to date. With that I posed the question, whatever happened to the spirit of website collecting exhibited by the Internet Archive?

It also makes explicit what the Wayback Machine implies, with its invitation to tell the history of a website and through it the history of the web – the life and times of Google as being the life and times of a decade of web history, in the example discussed. website biographies now could stand beside the event-based histories of the special collections (which in the U.S. at least appear in need of reinvigoration) and the national histories of the national web records (as they continue to be built).

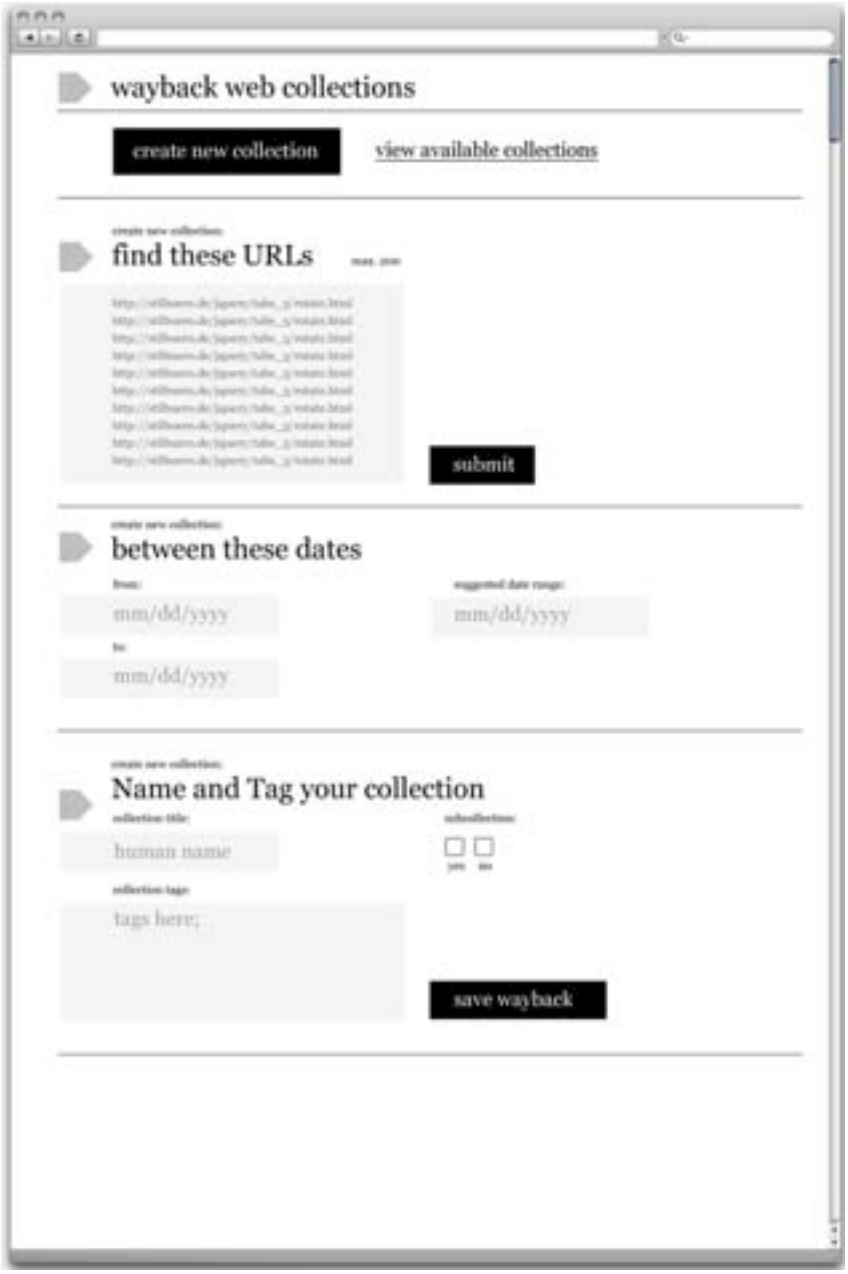


Figure three: Mock-up of website collection-maker. Make collection of already archived websites from archive.org. Digital Methods Initiative, Amsterdam, 2009.

The second outcome of applying the digital methods principles to the Internet archive and the Wayback Machine was to build a collection-maker of already archived websites, as mentioned above (see Figure three). In a sense such a collection-maker would be in keeping with a trend in web archiving towards providing tools for users to archive the web themselves (the Archive-IT project), instead of providing archives in search of users and researchers. The impetus was a desire to add another historiographical approach that also would be sensitive to the needs of web history. To the biographical, event-based and national historiographies on offer to date, colleagues and I sought to offer a past state of the web, or a portion thereof, which could be reconstructed and studied. The early blogosphere was chosen for its significance in web history, and the Eatonweb, the most complete blog directory of its day. The Eatonweb was used to date the end of the early blogophere: the day, or close to it, that Eaton no longer could keep up with his list of all blogs online, and thus when the blogosophere as sphere ceased to exist. In chapter four I treat the notion of the sphere in blogosphere as being held together by at least one link list, or core directory site, that links to all sites in it, so that in theory each site is equidistant from the core, in the classic geometrical form of the sphere. The last 'complete' list of the blogosphere at Eatonweb (15 August 2000) serves as the list of URLs for the nominal early blogosphere. Each URL is queried in the Wayback Machine, and the percentage of the early web that is archived is established. A remarkable 70% of the early blogosphere, as defined by Eaton, is available in the Internet Archive; a small percentage of websites could be added to that figure, if robots.txt code were removed from certain sites that were once significant in the early blogosphere. These sites are still online but are now parked and owned by domain resellers. (Michael Stevenson, heading up the project to conjure up the early blogosphere with the Wayback Machine, has considered purchasing the parked sites and removing the robots.txt code, thereby reactivating or reanimating the once missing websites in the Internet Archive.)

Each of the archived websites from the early blogosphere is crawled, and its outlinks captured. Using hyperlink mapping software, we created a cluster graph (or map) of the early

blogosphere, which includes on it not only those sites that are in the archive, but also the sites that are missing from it. Still lost, these missing blogs from the early blogosphere now reappear by name on the map, and the links to them and from them are visible, providing them with a context from the time that had been invisible in the single-site output of the Wayback Machine (or the categorizations of site types in special collections). The map of the early blogosphere, showing interlinkings between archived and non-archived sites, is a means of conjuring up a past state of the web. Among other things, it shows a sense of the relevance of the site at the time, and also thus the relevance of the sites in the collection (and those missing). Perhaps it also could put a value on the missing sites so as to aid with their recovery.16

16 By value I refer to placement of a site among others vying for prominence in the sense of a hyperlink economy. See Rogers, 2002.

S. Schneider and K. Foot (2004). “The web as an object of study.” *New Media and Society*. 6(1): 114-122.

new media & society

London, Thousand Oaks, CA and New Delhi
Vol6(1):114–122 DOI: 10.1177/1461444804039912
www.sagepublications.com

ARTICLE

The web as an object of study

STEVEN M. SCHNEIDER
SUNY Institute of Technology, USA

KIRSTEN A. FOOT
University of Washington, USA

INTRODUCTION

Over the last decade, as email, the world wide web and various digital technologies have emerged, scholars of new media have employed a variety of methodological strategies to explore the social, political and cultural phenomena associated with the growth of these applications. Several recently-published edited volumes highlight the range of methods employed in research regarding social dimensions of internet technologies (Gauntlett, 2000; Howard and Jones, 2003; Jones, 1999; Mann and Stewart, 2000). These collections, along with recent issues of scholarly journals, demonstrate that traditional methods of social research, such as ethnography (e.g. Hakken, 1999; Hine, 2000; Markham, 1998), textual analysis (e.g. Crowston and Williams, 2000; Mitra, 1999; Mitra and Cohen, 1999), focus groups (e.g. Price and Capella, 2001, Stromer-Galley and Foot, 2002), surveys (e.g. Parks and Floyd, 1996; Schmidt, 1997; Smith, 1997; Yun and Trumbo, 2000) and experiments (e.g. Iyengar, 2002) have been adapted for use online in order to investigate both online and offline phenomena. In addition, some scholars have found it useful to employ internet applications as bases for studies of purely offline phenomena (e.g. Witte et al., 2000). However, our focus is on the development of methods for studying the social dimensions of the internet itself, and in particular, the web.

As the web has emerged as a distinct media form in the past 10 years, it has been viewed increasingly as an object of study by social researchers. The ongoing evolution of the web poses challenges for scholars as they seek to develop methodological approaches that permit robust examination of web

phenomena. Some of these challenges stem from the nature of the web, which is a unique mixture of the ephemeral and the permanent. There are two aspects to the ephemerality of web content. First, web content is ephemeral in its transience, as it can be expected to last for only a relatively brief time. From the perspective of the user or visitor (or researcher), specialized tools and techniques are required to ensure that content can be viewed again at a later time. Second, web content is ephemeral in its construction – like television, radio, theater, and other 'performance media' (Hecht et al., 1993; Stowkowski, 2002). Web content, once presented, needs to be reconstructed or represented in order for others to experience it. Although webpages are routinely reconstructed by computers without human intervention (when a request is forwarded to a web server), it nevertheless requires some action by the producer (or the producer's server) in order for the content to be viewed again. In other words, the experience of the web, as well as the bits used to produce the content, must be intentionally preserved in order for it to be reproduced (Arms et al., 2001). Older media – including printed materials, film and sound recordings, for example – can be archived in the form in which they are presented; no additional steps are needed to recreate the experience of the original.

At the same time, the web has a sense of permanence that clearly distinguishes it from performance media. Unlike theater, or live television or radio, web content must exist in a permanent form in order to be transmitted. The web shares this characteristic with other forms of media such as film, print, and sound recordings. However, the permanence of the web is somewhat fleeting. Unlike any other permanent media, a website may destroy its predecessor regularly and procedurally each time it is updated by its producer; that is, absent specific arrangements to the contrary, each previous edition of a website may be erased as a new version is produced. By analogy, it would be as if each day's newspaper was printed on the same piece of paper, obliterating yesterday's news in order to produce today's.

The ephemerality of the web requires that proactive steps be taken in order to allow a recreation of web experience for future analyses. The permanence of the web makes this eminently possible. Although saving websites is not as easy as, for example, saving editions of a magazine, archiving techniques are evolving in such a way as to facilitate scholarly research of websites. As distinct from other ephemeral media, the web can be preserved in nearly the same form as it was originally 'performed' (Kahle, 1997; Lyman, 2002; Lyman and Kahle, 1998) and analyzed at a later time. Web archiving enables more rigorous and verifiable research, as well as developmental analyses that are time sensitive (e.g. Foot et al., 2003).

APPROACHES IN WEB STUDIES

Some of the broad questions currently under investigation by web scholars include the following.

- What forms of communicative actions are being inscribed on the web, and how do they change over time?
- How do the actions of web producers enable and/or constrain the potential actions of web users?
- What kinds of user experiences are potentiated on, and between, particular websites?
- How are relations between web producers, as well as between producers and users, enacted and mediated via web texts and links?

These kinds of research questions, along with the increasingly complex web applications that are altering traditional relationships between media form and content, challenge traditional approaches to social research. Web-based media require new methods of analyzing form and content, along with processes and patterns of production, distribution, usage and interpretation.

We identify three sets of approaches that have been employed in web-related research over the last decade. These approaches are not necessarily mutually exclusive, and some studies cited below employed more than one approach. Distinguishing between these approaches helps to establish the trajectory of web studies; highlighting the strengths and weaknesses of each focuses attention on the methodological challenges that are associated with the field of web studies.

The first set of approaches that we identify employ *discursive* or *rhetorical* analyses of websites; it is more concerned with the content of a website than its structuring elements. Studies employing these approaches focus on the texts and images that are contained on webpages, and/or on webpages/websites as texts in a Foucauldian sense (e.g. Baym, 1999; Benoit and Benoit, 2000; Sillaman, 2000; Warnick, 1998). Studies using a discursive/rhetorical approach, especially those that take broad views of what constitutes text, contribute significantly to our understanding of communicative phenomena on the web. However, we contend that the classic arguments regarding the inseparability of form and content in traditional media are especially applicable to the web, and thus studies of web 'content' that overlook the structuring elements of a webpage or site are also limited. Another limitation within this set of approaches is the paucity of analytical tools for making sense of the links among webpages and between websites. Some studies of hypertext intertextuality include analyses of cross-site linking, (e.g. Mitra, 1999; Warnick, 2001), but most content-focused studies of the web tend to reflect and perpetuate what we

believe is an inadequate construction of the web as merely a collection of texts. As Burbules and Callister (2000: 83) observe, 'people usually see points or texts as primary, and the links between them as mere connectives'. We agree with their claim that links are 'associative relations that change, redefine and provide enhanced or restricted access to the information they comprise', and we support the argument offered by Berners-Lee (2000), Mitra (1999), Odlyzko (2001) and others that, on the web, connectivity matters as much as content.

We characterize the second set of approaches as *structural* or *feature* analyses. Studies in this genre tend to use individual websites as their unit of analysis, focusing on the structure of the site, such as the number of pages and their hierarchical ordering, or on the features found on the pages within the site, for example, the presence of a search engine, privacy policy, or multiple navigation options (Benoit and Benoit, 2000; D'Alessio, 1997, 2000; Hansen, 2000; McMillan, 1999). Although understanding the structural and feature aspects of a particular site is important, our primary concern with these approaches is that they do not afford systematic analysis of an individual site's situatedness in the larger web, that is, the external pages to which it links and are linked to it. Another type of structural analysis employs computer-assisted, macro-level network analysis methods for mapping linking patterns (e.g. Jackson, 1997; Park, 2003; Park and Thelwall, 2003; Rogers and Marres, 2000, 2002). Studies of this type enable understanding of network structures on the web, but inferring the meaning or 'substance' of those network structures can be difficult to infer from large-scale mapping studies.

More recently, a third set of approaches to web analysis has emerged that takes hyperlink relationality into account in more nuanced ways. We refer to this set of approaches for analyzing multi-actor, cross-site action on the web as *sociocultural* analyses of the web (see several examples in Beaulieu and Park, 2003). Lindlof and Shatzer (1998) point in this direction in their article calling for new strategies of media ethnography in 'virtual space'. Hine (2000) presents a good example of sociocultural analysis of cross-site action on the web. Similarly, Howard's (2002) conceptualization of network ethnography reflects methodological sensitivity to processes of web production. By appropriating the term 'sociocultural' to describe this set of approaches, we seek to highlight the attention paid in this genre of web studies to the hyperlinked context(s) and situatedness of websites – and to the aims, strategies and identity-construction processes of website producers – as sites are produced, maintained and/or mediated through links.

WEB SPHERE ANALYSIS

Our own work has benefited from the methodological groundwork that has been established by our colleagues in web studies. We are developing a

multi-method approach called 'web sphere analysis' that enables analysis of communicative actions and relations between web producers and users developmentally over time. We conceptualize a web sphere as not simply a collection of websites, but as a hyperlinked set of dynamically-defined digital resources that span multiple websites and are deemed relevant, or related, to a central theme or 'object'. The boundaries of a web sphere are delimited by a shared object-orientation and a temporal framework. Web sphere analysis is an analytic strategy that includes relations between producers and users of web materials, as potentiated and mediated by the structural and feature elements of websites, hypertexts, and the links between them (Foot and Schneider, 2002; Foot et al., 2003).

The most crucial element in this definition of web sphere is the dynamic nature of the sites to be included. This dynamism comes from two sources. First, the researchers involved in identifying the boundaries of the sphere are likely to find continuously new sites to be included within it. Second, the notion of defining a web sphere is recursive, in that pages that are referenced by other included sites, as well as pages that reference included sites, are considered as part of the sphere under evaluation. Thus, as a web sphere is analyzed over time (ideally via an archive that enables retrospective analysis), its boundaries are dynamically shaped by both researchers' identification strategies and changes in the sites themselves.

The web sphere can function as a macro unit of analysis, by which historical and/or inter-sphere comparisons can be made. For example, the web sphere of the 2000 elections in the United States can be analyzed comparatively with the electoral web sphere of 2002 and those that develop in later years, as well as with electoral web spheres in other countries. Alternatively and/or simultaneously, other, more micro units such as texts, features and/or links can be employed in analyses within a web sphere (Schneider and Foot, 2002; Schneider and Foot, 2003).

Web sphere analysis is an analytic strategy that, when fully implemented, includes analysis of the relations between producers and users of web materials as potentiated and mediated by the structural and feature elements of websites, hypertexts, and the links between them. In a nutshell, the multi-method approach of web sphere analysis consists of the following elements. Websites related to the object or theme of the sphere are identified, captured in their hyperlinked context, and archived with some periodicity for contemporaneous and retrospective analyses. The archived sites are annotated with human and/or computer-generated 'notes' of various kinds, which creates a set of metadata. These metatdata correspond to the unit(s) and level(s) of analysis anticipated by the researcher(s). Sorting and retrieval of the integrated metadata and URL files is accomplished through several computer-assisted techniques. Interviews of various kinds are conducted with the producers and users of the websites in the identified

sphere, to be triangulated with web media data in the interpretation of the sphere.

CONCLUSIONS

The emergence of the internet, and especially the web, has challenged scholars conducting research to both adapt familiar methods and develop innovative approaches that account for the unique aspects of the web. This uniqueness includes both the nature of the communicative processes that it engenders, and the challenges that are posed in order to create research repositories allowing robust analyses (that are representative and reproducible) to proceed.

Methodological innovations have emerged in correspondence with the properties of these new media applications. This analysis has highlighted some methodological trends. Earlier studies of the internet tended to focus either on users and/or usage patterns, or on media and production characteristics. Within the user studies genre, the predominant methods were various forms of textual or discourse analysis and participant observation, with online surveys and experiments emerging later. Overarching these trends is a shift toward methods that recognize the co-productive nature of new media – thus the duality of users and producers – and the potential for digital media productions to be simultaneously inscriptions of communicative action and structures for action, especially on the web.

The emergence of web archiving techniques that are designed to facilitate scholarly analysis integrates researchers into archiving activities. Traditionally, the work of archivists proceeded largely independently of the scholars who would be expected to use the archived materials as the basis of research work at a later time. Given the cost and complexity of web archiving, an alternative approach is emerging that attempts to integrate researchers into archiving activities. If scholars join with archivists to identify web objects of interest, and to delineate strategies for building archives that support scholarly activities, the basis of future research efforts is likely to be enhanced, and new methods for web studies developed.

References

Arms, W., R. Adkins, C. Ammen and A. Hayes (2001) 'Collecting and Preserving the Web: the MINERVA Prototype', *RLG DigiNews*, 15 April, 5(2), URL (consulted August 2003): http://www.rlg.org/preserv/diginews/diginews5-2.html

Baym, N. (1999) *Tune in, Log On: Soaps, Fandom and On-line Community*. Thousand Oaks, CA: Sage.

Beaulieu, A. and H.W. Park (eds) (2003) 'Special Issue: Internet Networks: the Form and the Feel', *Journal of Computer-mediated Communication* 8(4), available online: http://www.ascusc.org/jcmc/vol8/issue4.

Benoit, W.J. and P.J. Benoit (2000) 'The Virtual Campaign: Presidential Primary Websites in Campaign 2000', *American Communication Journal* 3(3), available online: http://acjournal.org/holdings/vol3/Iss3/curtain.html#4

Berners-Lee, T. (2000) *Weaving the Web: the Original Design and Ultimate Destiny of the World Wide Web.* New York: HarperCollins.

Burbules, N.C. and T.A. Callister (2000) *Watch IT: the Risks and Promises of Information Technologies for Education*. Boulder, CO: Westview Press.

Crowston, K. and M. Williams (2000) 'Reproduced and Emergent Genres of Communication on the World Wide Web', *The Information Society* 16(3): 201–15.

D'Alessio, D. (1997) 'Use of the World Wide Web in the 1996 US Election', *Electoral Studies* 16(4): 489–500.

D'Alessio, D. (2000) 'Adoption of the World Wide Web by American Political Candidates, 1996–1998', *Journal of Broadcasting and Electronic Media* 44(4): 556–68.

Foot, K.A. and S.M. Schneider (2002) 'Online Action in Campaign 2000: an Exploratory Analysis of the U.S. Political Web Sphere', *Journal of Broadcasting & Electronic Media* 46(2): 222–44.

Foot, K.A., S.M. Schneider, M. Dougherty, M. Xenos and E. Larsen (2003) 'Analyzing Linking Practices: Candidate Sites in the 2002 U.S. Electoral Web Sphere', *Journal of Computer-mediated Communication* 8(4), available online: http://www.ascusc.org/jcmc/vol8/issue4/foot.html

Gauntlett, D. (ed.) (2000) *Web.studies: Rewiring Media Studies for the Digital Age.* New York: Oxford University Press.

Hakken, D. (1999) *Cyborgs@Cyberspace? An Ethnographer Looks Into the Future.* London: Routledge.

Hansen, G. (2000) 'Internet Presidential Campaigning: the Influences of Candidate Internet Sites on the 2000 Elections', paper presented at the National Communication Association, Seattle, WA, November.

Hecht, M.L., S.R. Corman and M. Miller-Rassulo (1993) 'An Evaluation of the Drug Resistance Project: a Comparison of Film Versus Live Performance Media', *Health Communication* 5(2): 75–88.

Hine, C. (2000) *Virtual Ethnography.* Thousand Oaks, CA: Sage.

Howard, P. (2002) 'Network Ethnography and Hypermedia Organization: New Organizations, New Media, New Myths', *New Media & Society* 4(4): 550–74.

Howard, P.N. and S. Jones (eds) (2003) *Society Online: the Internet in Context*. Thousand Oaks, CA: Sage.

Iyengar, S. (2002) 'Experimental Designs for Political Communication Research: from Shopping Malls to the Internet', paper presented at Workshop in Mass Media Economics, Department of Political Science, London School of Economics, 25–26 June, available at: http://pcl.stanford.edu/common/docs/research/iyengar/2002/expdes2002.pdf.

Jackson, M. (1997) 'Assessing the Structure of the Communication on the World Wide Web', *Journal of Computer-mediated Communication* 3(1), available online: http://www.ascusc.org/jcmc/vol3/issue1/jackson.html

Jones, S. (ed.) (1999) *Doing Internet Research: Critical Issues and Methods for Examining the Net*. Thousand Oaks, CA: Sage.

Kahle, B. (1997) 'Preserving the Internet', *Scientific American* 276(3): 82–3.

Lindlof, T.R. and M.J. Shatzer (1998) 'Media Ethnography in Virtual Space: Strategies, Limits, and Possibilities', *Journal of Broadcasting and Electronic Media* 42(2): 170–89.

Lyman, P. (2002) 'Archiving the World Wide Web', *Building a National Strategy for Digital Preservation*, April, URL (consulted August 2003): http://www.clir.org/pubs/reports/pub106/web.html

Lyman, P. and B. Kahle (1998) 'Archiving Digital Cultural Artifacts: Organizing an Agenda for Action', *D-Lib Magazine*, URL (consulted August 2003): http://www.dlib.org/dlib/july98/07lyman.html

McMillan, S.J. (1999) 'Health Communication and the Internet: Relationships Between Interactive Characteristics of the Medium and Site Creators, Content, and Purpose', *Health Communication* 11(4): 375–90.

Mann, C. and F. Stewart (2000) *Internet Communication and Qualitative Research Online: a Handbook for Researching Online*. London: Sage.

Markham, A.N. (1998) *Life Online: Researching Real Experience in Virtual Space*. Walnut Creek, CA: Altamira Press.

Mitra, A. (1999) 'Characteristics of the WWW Text: Tracing Discursive Strategies', *Journal of Computer-mediated Communication* 5(1), available online: http://www.ascusc.org/jcmc/vol5/issue1/mitra.html.

Mitra, A. and E. Cohen (1999) 'Analyzing the Web: Directions and Challenges', in S. Jones (ed.) *Doing Internet Research: Critical Issues and Methods for Examining the Net*, pp. 179–202. Thousand Oaks, CA: Sage.

Odlyzko, A. (2001) 'Content is Not King', *First Monday* 6(2), URL (consulted August 2003): http://firstmonday.org/issues/issue6_2/odlyzko/index.html.

Park, H.W. (2003) 'What is Hyperlink Network Analysis? New Method for the Study of Social Structure on the Web', *Connections* 25(1): 49–62.

Parks, M. and K. Floyd (1996) 'Making Friends in Cyberspace', *Journal of Computer-mediated Communication* 1(4), available online: http://www.ascusc.org/jcmc/vol1/issue4/vol1no4.html.

Park, H.W. and M. Thelwall (2003) 'Hyperlink Analyses of the World Wide Web: a Review', *Journal of Computer-Mediated Communication* 8(4), available online: http://www.ascusc.org/jcmc/vol8/issue4/park.html

Price, V. and J.N. Capella (2001) 'Online Deliberation and its Influence: the Electronic Dialogue Project in Campaign 2000', paper presented at the American Assocation for Public Opinion Research, May, Canada.

Rogers, R. and N. Marres (2000) 'Landscaping Climate Change: a Mapping Technique for Understanding Science and Technology Debates on the World Wide Web', *Public Understanding of Science* 9(2): 141–63.

Rogers, R. and N. Marres (2002) 'French Scandals on the Web, and on the Streets: a Small Experiment in Stretching the Limits of Reported Reality', *Asian Journal of Social Science* 30(2): 339–53.

Schmidt, W.C. (1997) 'World Wide Web Survey Research: Benefits, Potential Problems, and Solutions', *Behavior Research Methods, Instruments & Computers* 29(2): 274–9.

Schneider, S. M. and K.A. Foot (2002) 'Online Structure for Political Action: Exploring Presidential Web Sites from the 2000 American Election', *Javnost (The Public)* 9(2): 43–60.

Schneider, S.M. and K.A. Foot (2003) 'Crisis Communication and New Media: the Web After September 11', in P.N. Howard and S. Jones (eds) *Society Online: the Internet in Context*, pp. 137–54. Thousand Oaks, CA: Sage.

Sillaman, L. (2000) 'The Digital Campaign Trail: Candidate Images on Campaign Websites', master's thesis, Annenberg School for Communication, University of Pennsylvania.

Smith, C. (1997) 'Casting the Net: Surveying an Internet Population', *Journal of Computer-mediated Communication* 3(1), available online: http://www.ascusc.org/jcmc/vol3/issue1/smith.html

Stowkowski, P.A. (2002) 'Languages of Place and Discourses of Power: Constructing New Senses of Place', *Journal of Leisure Research* 34(4): 368–82.
Stromer-Galley, J. and K.A. Foot (2002) 'Citizen Perceptions of Online Interactivity and Implications for Political Campaign Communication', *Journal of Computer-mediated Communication* 8(1), available online: http://www.ascusc.org/jcmc/vol8/issue1/stromerandfoot.html
Warnick, B. (1998) 'Appearance or Reality? Political Parody on the Web in Campaign '96', *Critical Studies on Mass Communication* 15(3): 306–24.
Warnick, B. (2001) *Critical Literacy in a Digital Era: Technology, Rhetoric, and the Public Sphere.* Mahwah, NJ: Lawrence Erlbaum Associates.
Witte, J., L. Amoroso and P.N. Howard (2000) 'Method and Representation in Internet-based Survey Tools: Mobility, Community, and Cultural Identity in Survey 2000', *Social Science Computer Review* 18(2): 179–195.
Yun, G.W. and C. Trumbo (2000) 'Comparative Response to a Survey Executed by Post, E-mail and Web Form', *Journal of Computer-mediated Communication* 6(1), available online: http://www.ascusc.org/jcmc/vol6/issue1/yun.html

STEVEN M. SCHNEIDER is a visiting associate professor in the Department of Language, Literature and Communication at Rensselaer (Troy, NY) for the 2003–2004 year, and is Associate Professor at SUNY Institute of Technology. He is currently co-editor of politicalweb.info, and co-director of WebArchivist.org. Schneider has been studying the relationship between political life and the internet since 1988, and received a PhD in Political Science from Massachusetts Institute of Technology.
Address: SUNY Institute of Technology, Utica, NY 13504, USA. [email: steve@sunyit.edu]

KIRSTEN FOOT is an assistant professor in the Department of Communication at the University of Washington. As co-director of the WebArchivist.org research group, she is developing new techniques for studying social and political action on the web. Her research interests include co-production and mobilization on the web, and online campaigning practices. She co-edits the *Acting With Technology* series at MIT Press, and her work has been published in *Communication Theory*, *Journal of Broadcasting and Electronic Media*, and *Journal of Computer-Mediated Communication*.
Address: Department of Communication, Box 353740, University of Washington, Seattle, WA 98195, USA. [email: kfoot@u.washington.edu]