Comparing Issue Crawling and Just-in-Time Texts about Information Design and Visualization Research

Stan Ruecker and Zachary Devereaux
University of Alberta, Edmonton AB Canada

Abstract

We examine two technologies for harvesting web information—issue crawling and web scraping—and apply them to the question of how the information design and visualization research communities can be understood through their presence online. We note the participation of the military-industrial complex, as well as the changing appearance of the academic community at different times of the year. The text analysis component suggests a significant difference between the way this group is discussed in typical web sites as compared to its presence in the weblogs. We also look at how these research technologies, which are currently in their infancy, might be extended in directions that would be more useful. These suggestions for enhancement examine in detail the functionality that is available and might be made available, and constitute a potential area for future research by information designers and visualization researchers.

Keywords

Text analysis, Issue Crawling, Information Design, Visualization Research

Introduction

As design researchers attempt to theorize what they do and how they contextualize their activities, it can be difficult to understand from within the community how the various issues and debates are translated to a wider audience. However, a strategic combination of technologies for harvesting web information can allow researchers to take snapshots of the internet presence of a particular issue. This snapshot can also be applied to the issue network of the design community which is brought into being around the subject of information design (Marres 2002).

This paper discusses the techniques, methods, and technologies involved in internet issue crawling and the related text analysis of just-in-time documents that deal with the topics of information design and visualization research. The purpose of the project is to compare results from two of the different methods currently available, in order to see how different research strategies in this area produce different formulations of the topic under evaluation. An additional

result is to show how these subcategories of design research are being presented as issues in the public information space of the internet.

Previous work in the visualization of online issue networks has examined a wide range of issues, most often of a political nature. Current projects that applying the same crawler we used deal with several issues intricately caught up in expanding our understanding of international civil society as it is present on the internet. To this end crawler-related study has dealt with a wide spectrum of issues running from genetically modified foods to child labour, the wireless spectrum debate to media ownership, HIV / AIDS networks, corporate accountability, and international conflict (see www.govcom.org 'Social life of Issues workshop series' and www.issuenetwork.org). Theoretically, research in this field has focused on conceptions of network society (Castells 2000), web epistemology (Rogers et al. 2000) and the social consequences of such developments (Elmer 2004).

Computer-based text analysis also has a long pedigree, having been practiced for almost as long as there have been computers. Classic studies include the use of Bayesian statistics to ascribe authorship to the contested Federalist Papers (Mosteller and Wallace 1964) and the use of principal component analysis in the study of dialect in the novels of Jane Austen (Burrows 1987). More recently, the emphasis among text analysis researchers has shifted to the need to introduce the best practices in the field to conventional literary scholars as having a valuable role to play in interpretive work. Text analysis can assist the reader in identifying patterns in textual material, and in isolating areas for further study (Sinclair 2004; Ramsay 2004).

Indeed while the euphoric rhetoric of the 'ICT revolution' and the Schumperterian burst of the 'tech bubble' can both be considered as past their point of apex, the importance of the internet as an object of study is beginning to come into its own. New theory compels us to apply snapshot making tools by explaining the importance of remediation when it comes to an issue or community; the more media the issue network occupies and can be measured in, the greater its importance (Bolter and Grusin 2000). Thus our study engages reflexively insofar as it applies virtual cartography and automated text analysis to the information design research community as we find it articulated in new media.

Methods

The texts are generated using two different approaches: automated issue crawling for links, and the concatenation of text pages from search results. The result from the issue crawls is a fluid mass of information relating to design research, captured from thousands of internet sources. The concatenated results of web searches, on the other hand, combine far fewer documents, but yield a richer body of information. In order to discover how information design and visualization research can be characterized on the internet using these different approaches, the returns from the issue crawl are studied using a methodology that includes a network visualization system that displays the results of a co-link analysis, while the concatenated document is analyzed with online word frequency and collocation tools that show trends in occurrence and co-occurrence of significant terms.

Our issue crawling was performed with the aptly named IssueCrawler server-side web tool. The IssueCrawler sends a robot spider onto the web in order to follow and retrieve the colinkage network of the actors studied[1]. The colinkage network retrieved is articulated with cluster-mapping visualization, rendering an interactive map of the issue network[2].

Issue crawls require the selection of starting URLs that can serve as reasonable points of departure. These might be characterized as institutional embodiments of the issue network. An important point to note is that because of the nature of the internet to interact is to leave a trace. For this reason virtual cartography can be applied across language, and of course geographic barriers in the operationally closed system of the network.

We selected nine URLs as starting points, divided into three categories; significant research institutes, significant corporate sponsors, and significant professional organizations. This strategy was informed by the systems theory insight that interactions between different parts of the network drive innovation forward (Leydesdorff 2000).

The three significant institutes dedicated to information design in our starting points were: CRIA, the visualization center at the University of Maryland, and MIT's New Media Lab. Exploration of the MIT New Media Lab's web presence led us to the consideration of corporate sponsorship in this research sector. Taken from their sponsor links page, we included the research and education links pages at three major corporations: HP, IBM and Nokia. To round out our set of starting points, we included three professional organizations directly engaged in information design research: IIID, SIGGRAPH, and the GDC

Findings related to our first Information Design Research Network map

A clear aspect of the first map, generated in the summer, is that the presence of .edu, i.e. universities, is extremely limited (Figure 1). Only one .edu showed up in the information design research network, and following the node link of 'cs.unc.edu' leads to a SIGGRAPH related ACM Workshop on General Purpose Computing on Graphics Processors. But if one looks at the sponsors of the event and some of their slogans, such as "All But War is Simulation" a second characteristic of the information design network becomes clear, and is reinforced by the larger map on the whole.

---

[1] IssueCrawler was developed by the Govcom.org Foundation, Amsterdam. The tool has been applied to multiple issue networks in workshop settings over the last 5 years. See: www.govcom.org & www.issuenetwork.org

[2] Cluster mapping technology has been developed by Aguidel, SA, Paris, and has also been applied to a wide array of issues independently and in conjunction with the Govcom.org foundation. See: www.aguidel.com

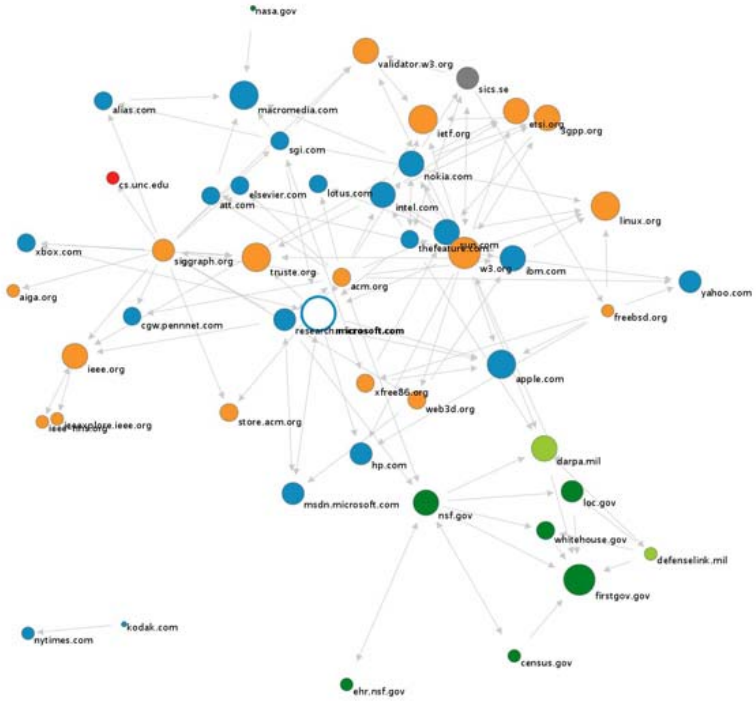

Figure 1. Our first issue crawl visualization for information design and visualization research showed a noticeable lack of university sites (there is one, top left). However, the military-industrial complex was present (lower right), even though none of the starting points included these nodes.

Clearly delineated in the first information design research issue network, we find a nexus of the US-led military-industrial complex (lower right-hand quadrant). Somewhat surprisingly, NASA is also present, through its connections to Macromedia. The constituents of the military-industrial nexus are interesting in their own right, including organizations such as the National Science Foundation, the White House, DARPA and Defenselink.

Thus the clear identification of the 'military-governmental' subset within the information design network was possible through Issue Crawling. And our interaction with the map and exploration of the military-governmental constellation led to more technical considerations of the tool itself which are elaborated upon below.

The bulk of the rest of the network in our first map is made up of a highly interconnected and relatively balanced mix of NGOs (.orgs) and corporations (.coms). Both groups are diverse in nature and colinkage is an activity availed of by many of the nodes in the network. In this issue network linking is business as usual. Thus the community is comfortably remediated.

Findings from the Newer Map

In order to see how this information changed over time, we re-submitted the same starting URLs for another crawl, three months after the original. In this version, the universities clearly appear (top and top right). One possible explanation for their sudden reappearance is that the school year represents a

period of relative stability for university sites (the second map was done in October), while the summer is a period of site update and revision (the first map was from early July) (Figure 2). Another possibility is that the crawling software met with a condition that prevented it in the former case from making crucial links to the .edu cluster. As is often the case with algorithmic retrieval systems, there is really no satisfactory way for the user to tell what happened, any more than there would be a reasonable way to understand how a system like Google obtained the results it did. The algorithms are proprietary, and even in cases where they have been published, they do not fall within the domain expertise of many researchers.

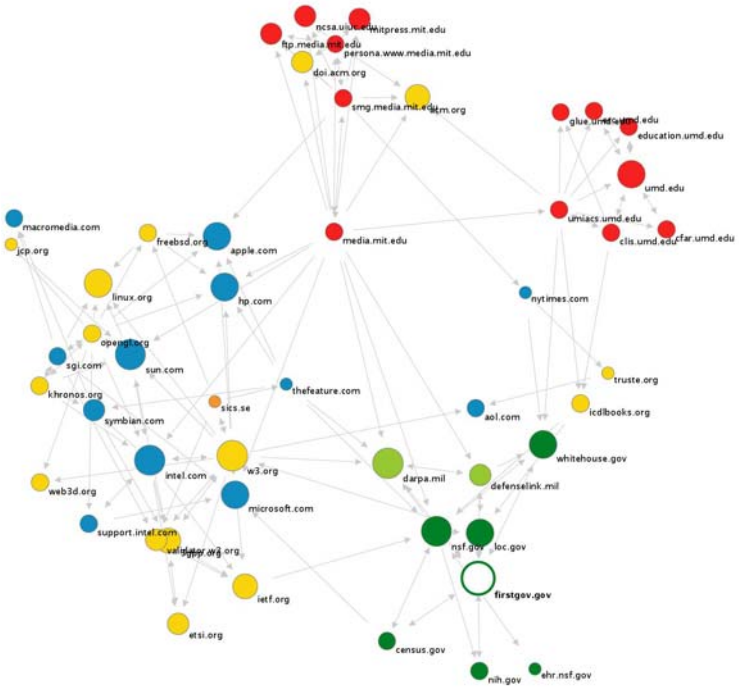

Figure 2. The universities reappear (top and top right) in the second issue crawl for information design and visualization research, produced three months after the first.

Technical Considerations of Visualization Interfaces

Tufte (2003) talks about the cognitive style of applications, emphasizing particularly that some are better than others. Sinclair (2003) talks about the importance of exploration when it comes to humanities computing. And these two concepts are key for understanding the value of the techniques we applied to the information design research community.

We can only "really" explore the information design community as it is rendered by IssueCrawler when we have the SVG online on a computer. And we can only access Just-in-Time texts in the normal way through online interactivity (and this says nothing of the importance of timeliness, the evolutionary disposition of shifts in the information landscape). This is a subtle but important point and we find ourselves tempted to analyze it like Innis

(1972); the media, here, has a strong effect on our analysis. Really it means we have to gaze and that Bush (1945) was unbelievably prescient in terms of foreseeing that in the age of computers explaining the route that one took to a conclusion or finding would be incredibly, if not predominantly important.

The wikipedia is another case where one sees clearly that the fun actually is in getting there (Viégas et al. 2004). As Rockwell et al (1999) note, the platforms we use for the production of knowledge do tell a story in and of themselves. Indeed metadocuments, which describe the data, are particularly important for explaining the research in cases such as ours, which combine results from several different technologies. Automated text analysis and issue network cartography are aimed at enabling the researcher to deal with digital online information; these approaches also make possible the analysis of online phenomenon as they develop. Nonetheless, the inherently empowering, constraining and even political nature of design comes to the fore in the form of formatting issues (Rogers 2004).

The point is that the interface matters to the researcher's comprehension of the network. The layering of information, such as displaying links, can only be achieved in print through an exhaustive and finally non-interactive manner, introducing considerable barriers to understanding. These barriers are similar to Tufte's example of the problem with PowerPoint printouts; the point is that not only is print unwieldy for network exploration, it is essentially disconnected.

Indeed we can illustrate this point in abbreviated form when it comes to our conclusions about the presence of the military (and industrial) complex in the information design research network.

To do this and exemplify (after the style of Tufte) we need to present. layer by layer, the capabilities of the .SVG map as they relate to our topic: the military-governmental constellation. In order to follow the informational threads we have available through this visualization system, we need to examine the entire map, the links to the nodes in the constellation, and the links from the included nodes.

In this way we can identify the gateways to the military cluster. For example, one question we want to answer is: how does the military-governmental constellation relate to the research network? One of the principal gates, although not the only one, is the MIT Media Lab (Figure 2). There are also connections to Apple and the ACM (Figure 1).

Thus design is political, for the visualization and interactivity processes structure our capability to identify networks such as the military-governmental constellation. Empowering, yes, as otherwise how could we have discovered the gatekeepers and swing nodes within the information design research community? But limited as well, because we need to articulate constellations for meaning, and this work could be made more accessible and efficient with better design.

And this leads us to a limitation of the interface: We can view our constellation but we cannot label it, categorize it, or 'deeply capture' its nature beyond the map. These problems with selecting, labeling and manipulating cluster groups, or 'constellations' which are of consequence to the research conclusions persists in ReseauLu as well, and are thus an issue in common across these two platforms. There is another component we wish to emphasize in terms of exploration, and that component is the chronological and morphological capacity of cluster mapping visualization techniques. Plans are currently in place for the development of a new set of tools that will allow a simpler display of information as it changes over time. Indeed 'time is of the essence,' which leads to our consideration of the Just-in-Time Text analysis below.

But before we over-emphasize the limitations of the exploratory function in relation to formatting issues inherent in current cluster map cartography, we should take note that IssueCrawler specifically has a high degree of interactivity. We can turn nodes on and off and we can also eliminate specific types of nodes. These are capacities in the right direction, and they aid exploration through complexity. Indeed the IssueCrawler information design research network map successfully differentiated between .mil and .gov while still embedding their interactivity as a group, giving us a greater understanding of this complex in its relation to the larger community.

On the topic of complexity there is another aspect of cluster mapping as it is presented in the case of ReseauLu and IssueCrawler that has consequences for our argument. That aspect is the 'tipping point' of overwhelming complexity. On the one hand a standout strength of these programs is the fact that they can coherently and comprehensively articulate huge networks with incredible complexity. These, indeed, are nearly impossible to explore in the printed paper format, and our sense of limitation for such means of visualization as expressed earlier hold in this regard. But again the advances and gains of the program and its visualization techniques renders us sensitive to the potential for greater opportunity in the cluster map environment.

Namely, the capacity to zoom comes into play, as the higher level of complexity renders the examination of linkages and relations more difficult. The colloquial experience of this often comes when a cartographer deals with a map projected digitally on a wall for the first time. The exploratory experience is profoundly different than the smaller scale. Here there is certainly no limitation of cognitive style for the communal gaze.

Findings from the Just-in-time Text Analysis

For the purposes of this study, we captured and analyzed two bodies of text that relate directly to the material harvested through the issue crawler. However, whereas the issue crawler works by following links, the text scrapers work by supplying search terms to retrieval engines. We therefore decided to use the expanded forms of the nine starting points for the issue crawl as search terms, first in a scrape of the top 100 summaries for each key phrase in Google; then for the top 100 summaries from a similar scrape of Blogstreet, which is limited to searching weblogs. We then analyzed these two composite just-in-time texts using an online word frequency generator that is part of the

suite of tools at TAPoRware, followed by a series of manual tasks where we sorted words into categories for further study.

One immediate result we obtained was that a third scrape, using Google News, provided no hits from our nine key phrases. Google News at the time of this scrape indexed more than 4,000 media sources. We tested some other phrases to make sure the retrieval system was working, which it was. The problem was that the contents of Google News represent only a three-month window of coverage, and it appears that within that media space, the information design and visualization research community as we construed it for the purposes of this project was not an active presence in these media sources during that time.

The returns from Google, on the other hand, produced more than 19,000 words, in the form of 20-word excerpts from several hundred sites. The Blog scrape yielded a sample only half that size, at approximately 11,000 words (see Figure 3). By comparison, a typical 250-page paperback novel might contain 80,000 to 100,000 words. Reading through 30,000 words consisting of 20-word extracts from thousands of sources is the kind of task that many people would find unappealing.

Scrape Summary
Blog scrape
11072 words
2777 unique words
145 content words occurring 7 or more times
Google scrape
19084 words
3885 unique words
286 content words occurring 7 or more times

Figure 3. The Blog scraper produced 145 frequent content words, while the larger sample obtained from Google scraping yielded twice that many.

For both scrapes, a cutoff point for the frequency lists was set at 7 instances (tokens) of a given word (type). Since the purpose of using frequency lists is to identify terms of interest, it is necessary to choose a cutoff point that includes enough significant terms without including everything. The smaller word list thus obtained was then further reduced by selecting the content words (usually nouns and verbs) and setting aside the function words (such as articles, prepositions, and conjunctions). Once the content word lists were established, they were manually sorted into various emergent categories, such as organizations, names of people, and words to pursue further.

The resulting sub-lists showed some interesting patterns, just as Ramsay (2004) suggests they might. The picture painted by Google is one that the information design and visualization research communities would probably recognize, and to some extent endorse. There is a strong presence of words relating to graphics, illustration, and arts, which signals the historical connection between the graphic arts and visual communication design. There

are significant figures associated with the centres of research, such as CRIA's David Sless, the MIT Media Lab's Robert Jacobson and Benjamin Fry, as well as the well-known theorist Edward R. Tufte. A cursory glance at the scrape file itself, however, reveals that more names occur than were present above the cutoff line of seven tokens. A closer reading for name lists might therefore be something worth considering.

From the perspective of the Blog scrape, on the other hand, the information design and visualization research community has a somewhat different profile. First of all, there is no significant presence of terms relating to the graphic arts: the Blogs seem for the most part to be written not by people who are talking about art and design, but by people who are talking about technology and design. Secondly, none of the names from the Google list are present in the Blog list. Instead, we find that, much as the research centres kept the names of their researchers present in the Google returns, the successful branding of the Blogs yields a list consisting primarily of the names of the bloggers, such as Russell Beattie, Roland Piquepaille, and Rajesh Jain. However, also present were names of hot authors such as Howard Rheingold, whose book *Smart Mobs* (2002) deals with technology-enhanced forms of social activism.

There were also some problems with the Blog scraper as a technology. Within the scrape results, some summary texts were included multiple times. For example, a summary text from NathanNewman.org was repeated 7 times, which meant that some terms from this single text fragment were present above the seven-token cutoff point. The repetition factor in the blogsphere results in significant questions about the ‘copy-paste’ culture of what is popularly portrayed as a liberating medium (http://www.issuenetwork.org/node.php?id=59).

Implications for Information Design and Visualization Research

Issue crawling and web scraping followed by text analysis are two alternative research strategies for harvesting web information. Both of these techniques are highly dependent on the details of the implementation of the technology. They also yield somewhat different information, to the extent that it is possible to consider them as complementary rather than conflicting technologies. While the issue crawl provides an overview of the linkages between significant locations on the web, the web scrape and text analysis provides a detailed view of topics at the level of analysis of the individual word. Surely the combination of a general picture of who is involved with a detailed examination of what they have to say is a powerful research strategy, although it is dependent for its effectiveness on the functionality available through the online research tools.

In what ways do issue crawling and just-in-time text analysis provide an improvement over simply searching the internet with a retrieval engine and browser and reading through the results? At this point, the issue crawler and its subsequent visualization founded on ReseauLu clearly provide information that would otherwise be very difficult, although not impossible, to assemble. The interactive capacities of the system also allow the results to be studied in some detail: clicking an individual node, for instance, results in a legend of

network statistics being displayed on the side of the screen (Figure 4) and clicking on a node title will call up the linked site.

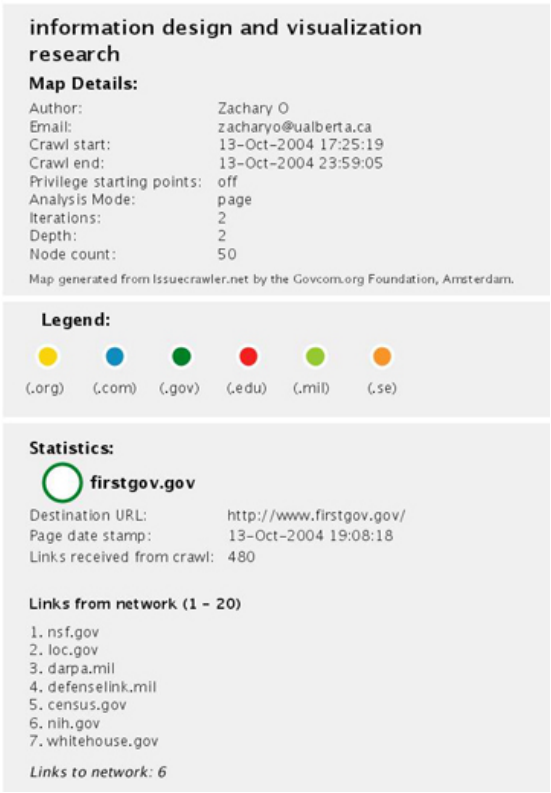

Figure 4. The SVG visualization of the results of the issue crawl includes a network statistics legend that changes as the user clicks the various nodes in the display.

However, a significant aspect of this research area is that, while the tools involved rely heavily on visual information design methods, they are in their current implementations still comparatively primitive. There are a number of areas in which existing tools could be enhanced by visual communication designers; there are also several areas where entirely new tools remain to be designed.

Future Directions for the Technologies

For web scraping, some additional functionality would be helpful to allow the user to control the choice of the sample size. The scrapers we used retrieved the summary texts only, which are approximately twenty words from the top of the home pages. It is difficult to say in what ways results would change with different sample sizes, or with the additional capacity to scrape texts from a site's internal links. A combination crawler and scraper might allow the researcher to take pieces from each page in a site, resulting in a better representation of the whole. The conundrum, really, is that the ‘top 100’ technique shows ‘proximity to the top,’ the same as the difference between a front page story and one buried in the back sections. It is a useful metric (the front page matters) but there remains a larger scale to be explored.

Another useful capacity would be an automated method of separating content words from function words. A stop list on the frequency counter would achieve this nicely, although a sorting mechanism would allow subsequent review of the two groups. Similarly, it would also be helpful to have an automated strategy for collecting proper nouns, such as those used for people's names and the names of organizations. For this project, these steps had to be carried out manually, which is unnecessarily time consuming.

Finally, the text analysis results could usefully be the basis for another visualization system that would provide a graphic means for studying selected terms. What is required here is a way of examining patterns of relationship between specific words, combined with a means of looking quickly at greater detail. In short, it would be useful to have a rich-prospect kind of browsing interface (Ruecker 2003) for scrape results that have been subjected to text analysis processes.

In terms of the issue crawls, we have mentioned the importance of interaction through the visualization tools. One of the valuable current features is the ability to zoom on selected portions of a cluster map. But this capacity could be expanded if the user were also given the capacity to zoom in and maneuver through more than one focus point of the map at a given time. The desire in this case is to render the more complex maps more intelligible. And again, multiple zooming, be it with several fisheyes or some other means, exemplifies as a concept the importance of exploration to the generation of meaning.

Similarly, the enabling of active and interactive link highlighting would allow the viewer to select multiple connections to examine in greater detail.

It is with the larger scope of exploration that the potential to fruitfully combine text analysis and cluster map visualization comes to the fore. And in this regard we must pay due respect to the myriad component forms of the internet. This is one of its great and incredible capacities, as well as one of its most frustrating.

References

Bolter, J.D. and R. Grusin. *Remediation: Understanding New Media.* Cambridge, MA: MIT Press, 2000.

Burrows, J. *Computation into Criticism: A Study of Jane Austen's Novels, and an Experiment in Method.* Oxford: Oxford University Press, 1987.

Bush, V. “As We May Think.” *The Atlantic Monthly*. July 1945. 176:101–108. http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm

Castells, Manuel. *The Rise of the Network Society.* Oxford ; Malden, MA : Blackwell Publishers, 2000.

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

Innis, H. A. *Empire and Communication*. Toronto: University of Toronto Press, 1972.

Leydesdorff, L. *A Sociological Theory of Communication: The Self-Organization of the Knowledge-Based Society*. Macquarie Park, NSW: Universal Publishers, 2000.

Marres, N. "May the true victim of defacement stand up! On reading the network configurations of scandal on the Web." In Bruno Latour and Peter Weibel, eds., *Iconoclash.* Cambridge, MA: MIT Press, 2002.

Mosteller, F. and D. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1964.

Rheingold. H. *Smart Mobs: The Next Social Revolution*. NY: Perseus Books Group, 2002.

Rockwell, G., J. Bradley and P. Monger. “Seeing the text through the trees: Visualization and interactivity in textual applications.” *Literary and Linguistic Computing*. 1999. 14.1:115–30.

Rogers, R. *Information Politics on the Web*. Cambridge, MA: MIT Press. 2004.

Rogers, R. ed., *Preferred Placement - Knowledge Politics on the Web*. Jan van Eyck Editions, Maastricht, 2000.

Ruecker, S. *Affordances of Prospect for Academic Users of Interpretively-tagged Text Collections*. Interdisciplinary Ph.D. in Humanities Computing. Edmonton: Departments of English and Art and Design, University of Alberta, 2003.

Sinclair, S. "Computer-Assisted Reading: Reconceiving Text Analysis." *Literary and Linguistic Computing*, 2003. 18.2:175-184.

Tufte, E. *The Cognitive Style of PowerPoint*. Cheshire, Connecticut: Graphics Press LLC, 2003.

Viégas, F. B., M. Wattenberg and K. Dave. "Studying Cooperation and Conflict between Authors with history flow Visualizations." *CHI 2004*, Vienna, Austria. April 24–9, 2004.

About the Authors

Dr. Stan Ruecker
Assistant Professor
Humanities Computing
Department of English and Film Studies
3-5 Humanities Centre
University of Alberta
Edmonton AB T6G 2E5
CANADA

email: sruecker@ualberta.ca

Dr. Stan Ruecker is an Assistant Professor of Humanities Computing in the Department of English and Film Studies at the University of Alberta. He is a graduate of the University of Regina (BA Hons English 1985, BSc Computer Science 1988), the University of Toronto (MA English 1989), and the University of Alberta (MDes 1999, PhD 2003). His research interests are in the areas of computer-human interfaces, text visualization, and information design. His PhD research was on the affordances of prospect for computer interfaces to large, interpretively-tagged text collections.

Zachary Devereaux
MA Candidate
Department of Political Science
10-16 Henry Marshall Tory Building
University of Alberta
Edmonton, AB T6G 2H4
CANADA
email: zacharyo@ualberta.ca

Zachary Devereaux is an MA Candidate in Political Science at the University of Alberta. Zach holds a BA from the UofA (History / Fine Arts, 1997) and has studied media and communications in Amsterdam at the graduate level (UvA). Zach's research interests focus on the intersection of design, new media and international relations. At the MA level these themes have been pursued by the use of software based visualization techniques to study new media news coverage of the North Korean nuclear issue. This research direction is informed by Zach's experiences in film, radio, newsprint and globalization. Zach hopes to pursue PhD study of new media webtools and conflict.