Internet-Mediated Research Using Surveys and Content Analyses

By John B. Killoran
MLA Conference
December 27-30, 2005
Washington, DC

Introduction

Despite the widespread adoption of Internet-mediated research methods across the social sciences, technical communication research has been increasingly oriented to site-specific "bricks-and-mortar" ethnographic methods specifically and qualitative methods in general. Whereas such qualitative methods are sensitive to the richness and nuances of context, of authorship practices, of readership or user practices, they are not necessarily optimal for gathering data from certain understudied areas of our field:

- international populations of communicators,
- technologically-mediated documents.

Both these populations and documents are increasingly important for our field, and these populations and documents may sometimes best be accessed and studied through Internet-mediated research methods.

I'm going to talk about two methods that can be combined and carried out through the Internet:

1 surveys of Internet document producers,
2 and content analyses of their documents.

Doing both together can triangulate our data collection. However, both are quantitative methods, which presents me with a challenge, for our field has traditionally been more oriented to qualitative methods—they better answer the kinds of research questions we ask.

1 So first I'll discuss what we might lose or gain from such quantitative methods.
2 And then I'll discuss methodological issues specific to carrying out Internet-mediated surveys and content analyses.

Quantitative Methods: Drawbacks and benefits

First, drawbacks. Quantitative methods, as their name implies, tend to emphasize the quantifiable, at the expense of the non-quantifiable aspects of discourses and practices. Perhaps most seriously for technical communication researchers, quantifiable methods tend to treat such non-quantifiable

factors as context reductively, decontextualizing discourse from its natural environments, reducing its nuances to crude measures or features that can be enumerated across different contexts. So we can certainly accumulate lots of data through quantitative methods, but can that data be converted into knowledge that is appreciated by our research community?

On the other hand, such quantitative methods as surveys and content analyses are frequently employed in the social sciences, and our field of technical communication, a schizophrenic field that conceives of itself as at least partly in the social sciences, can usefully draw from their productive methods.

Internet-mediated quantitative methods in particular enable us to accomplish several things:

- To explore digital genres, and much Internet-based communication relies on either new genres or adapted versions of pre-Internet genres. Whereas researchers in other disciplines, communication researchers and social science researchers in general, have been endlessly analyzing blogs and civic-oriented discussion boards and various forms of electronic communities, they don't have the theoretical frameworks that enable them to add much insight to professional / technical / business documents that are not explicitly political or explicitly communal / social in nature. But the Web is filled with such professional / technical / business documents, and informational sites in general, and it's in analyzing such documents that our professional communication perspectives can make a contribution: with our conceptual repertoire, including genres, subject position, and so forth.
- To access international populations that tend not to be studied because technical communication tends to be such an American field, with fewer of the international colleagues and international research opportunities than are available to our colleagues in communication and the social sciences in general.
- To access demographic subpopulations and their digital documents that we would never otherwise have access to because they are so distributed geographically and socially, such as, say, reports or white papers published by PIRGs and environmental groups worldwide lobbying for greater use of sustainable energy resources.
- To legitimate our examination of discourse producers and their discourse that, in single rhetorical analyses or case studies, would be difficult to legitimate but who, in aggregate, endow a legitimacy simply by their numbers.
- To more reliably generalize our results to broader populations in ways that we can't reliably do in ethnographic studies, case studies, and other context-sensitive qualitative methods.

- To supplement other context-sensitive qualitative methods with case studies, follow-up e-mail interviews based on each participant's documents posted on-line (e.g., an on-line version of the discourse-based interview that Odell, Goswami, and Herrington described a generation ago).

Methodological Issues

Now I'd like to turn to methodological issues that distinguish Internet-based surveys and content analyses of multi-modal digital documents from their print-based counterparts:

- Sampling
- Surveys of Internet document producers
- Content analysis of Internet documents

As I present this, from time to time I'll draw on my own experiences conducting two surveys of international samples of Web authors together with analyzing their Web documents.

Sampling

First, sampling issues, and let's start with access. According to a recent study by the Pew Internet and American Life Project (2004), "44% of U.S. Internet users have contributed their thoughts and their files to the online world." Apart from the other 56% who have not contributed directly to the online world, how many among that 44% are really accessible to us? I would love to survey those who post book reviews on Amazon.com, but their e-mail addresses are hidden, and Amazon would be understandably reluctant about having some researcher spam their contributors with survey requests, and especially reluctant about having a researcher report such confidential info in public. I would also love to survey job-hunters who post their resumes on Monster.com, but apart from the cost of paying Monster for the privilege, that raises significant ethical issues (i.e., spamming Monster customers with non-employment offers); without Monster's co-operation, it would be risky, and even with Monster's cooperation, it would likely not be something they would want to see published. However, there are other populations who are accessible to us:

1 those who post their resumes not on Monster but on their own professional profile or portfolio sites,
2 activists who maintain informational sites on social or political or environmental issues,
3 professionals and consultants who maintain their own business and informational sites,
4 non-profit organizations (many of which are homey affairs and would be willing to fill out a survey in exchange for the researcher's survey results),

5 and needless to say educators who maintain sites for their courses, for their research, and for their academic departments.

In addition to the access issue, sampling can be much more problematic because defining the sample frame is so problematic. Ideally, sampling from a population requires a clear sense of the whole population. For some research questions, this is not a problem because we define our population by fairly stable parameters outside of the Internet. This has been done in several content analyses of Web sites, such as the sites of Fortune 500 companies, or the sites of main party political candidates running for election; we can precisely enumerate each of these populations. By contrast, most content analyses have traditionally been based on convenience or purposive samples rather than representative samples. And that tradition may continue with the Internet. The Internet is arguably our most extreme example of a bottom-up social structure, with little top-down governance and hence with few of the authorative bodies that record other kinds of populations, such as phone books, voters lists, and so forth. Because of this extreme democratic nature of the Internet, at least among those who do have access, we cannot easily envision much less even find and enumerate the whole Internet population of interest. Search engines help, of course, and some sample frames have been defined by whatever turned up in the search engines that "met criteria related to the purpose of the study" (McMillan, 2000, p.83). But much Internet communication is inaccessible because it has not been found by search engines, especially communication by ordinary individuals.

For example, for my dissertation research in the late 1990s, I surveyed authors of personal homepages (which were a hot thing back then), and I selected my sample fairly systematically from Yahoo's listings of 70,000 personal homepage authors, which was the largest listing I could find. But I was under no illusion that there were only 70,000 personal homepages out there in the spring of 1997. And despite the widely publicized recent claims of Yahoo and Google about the number of pages they index, search engines are still notoriously fickle.

So systematic random sampling is often not feasible. Based on the critical feedback I've received in my research, I think we build more credibility by focusing on a narrowly defined population, and acknowledging our inability to generalize confidently to other populations, rather that aiming for a demographically broad cross-section of the whole Internet population. So if you were interested in environmental communication, for instance, you might be well-served by focusing your research question and hence your sample population to one small well-defined set of

environmental groups who, say, link to each other or write about the same environmental issue, and sampling from that population very thoroughly, rather than aiming for a wide list of environmentalist who are demographically and geographically diverse, unless you intend to compare populations along those demographic and geographic lines, such as American environmentalists vs. British environmentalists.

Surveys of Internet Document Producers

Once we have a sample, we are ready for our survey and content analysis. Let's now look at surveying. A lot has already been written about surveying in the print medium, so let me focus on two issues that are particularly sensitive to Internet-based surveys:

- response rates,
- and response quality.

Reviewers easily notice response rates and they have an easy critique to levy against low response rates. There have been reports that response rates for surveys in general, not just Internet-based surveys, have been falling and that the American population, at least, is becoming oversurveyed (e.g., Sheehan, 2001). For instance, you are probably familiar with STC's annual salary surveys. Over the past decade, response rates among STC members have fallen precipitously, from 45+% in 1997 consistently downward to only 16% in 2004, though it uncharacteristically jumped up in the most recent 2005 survey to 23%.

In comparison with surveys in more traditional media, response rates for Internet-based surveys have tended to be even lower (e.g., Crawford, Couper, and Lamias, 2001; Manfreda, Batagelj, and Vehovar, 2002). Some have warned that the ease of implementing Internet-based surveys, plus the proliferation of e-mail spam and phishing schemes, may be further reducing response rates (e.g., Yun and Trumbo, 2000). Studies with print-based postal mail surveys have shown that showing a university affiliation will increase response rates (Sheehan, 2001), though the level of suspicion is such that prospective respondents still might suspect that it's a fake address.

Those who do respond are most likely to do so within the first 48 hours of receiving the solicitation (e.g., Sheehan and Hoy, 1999; Smith, 1997). In surveys conducted in traditional media, it has been proven effective to follow up non-respondents with additional requests or reminders in the weeks after the first request (e.g., Sheehan, 2001). Because of the speed of the Internet, some have advised that this interval be shortened to days so that a reminder is received

while the initial request is still fresh in one's mind (e.g., Crawford, Couper, and Lamias, 2001; Yun and Trumbo, 2000). I've been somewhat reluctant to do this because the Web sites I surveyed often made evident their defensive tactics against spam, such as by posting an e-mail address "nospam@ . . .," by posting an e-mail address only in human-readable form, and so forth.

Let's now consider response quality, especially the kinds of more elaborate better-developed responses that would be of interest to us discourse researchers. It's not for nothing that surveys are categorized as a quantitative method, for open-ended questions have traditionally not generated elaborate answers. Open-ended questions have been shown to prompt respondents to abandon on-line surveys (e.g., Bosnjak and Tuten, 2001; Crawford, Couper, and Lamias, 2001). One group of researchers point to "the extra effort, both cognitive (in formulating a response) and psychomotor (in keying the response into the computer)," that open-ended questions demand of their respondents (Crawford, Couper, and Lamias, 2001). They also observe that "There is no interviewer present to motivate respondents to continue with the survey, and new ways must be found to maintain motivation and interest" (Crawford, Couper, and Lamias, 2001).

Ironically, however, in comparison with print surveys, the *quality* of open-ended responses seems better in electronic form: "A number of researchers . . . have reported that respondents write lengthier and more self-disclosing comments on e-mail open-ended questionnaires than they do on mail survey questionnaires" (Yun and Trumbo, 2000). In my survey of Web resume authors, I was pleasantly surprised at the response to one of the final questions, a very open-ended question inviting respondents simply to offer other information, experiences, or observations that could help me understand their experience with their Web resume; about 60% wrote substantial comments, sometimes quite elaborate. In the preceding question, which asked respondents to rank their Web resume's usefulness to them and then to explain its usefulness, over 80% wrote out substantive comments, again sometimes elaborate.

Content Analysis of Internet Documents

Now let's turn to content analysis. In contrast with rhetorical analysis, which is frequently employed to analyze individual digital documents, content analysis is arguably the most frequently employed methodology to analyze multiple digital documents. Content analysis can efficiently take advantage of the Web's "granularity," in which textuality is typically constituted of discrete genres linked together technically but not necessarily rhetorically, or discrete elements within documents such as lists, graphics, and links. This makes it especially effective for broad-

scale analyses. However, because it is directed more to mining a site or page for its parts than to explaining how those parts might engage as a whole, content analysis alone is less suitable for addressing issues between a document and its context, its authors, users, and so forth. For those issues, we would have to supplement content analysis with such qualitative methods as Internet-based interviews.

Because Internet-mediated communication is so new, content analyses of Internet documents present researchers with the challenge of formulating intelligent questions for that novelty. One risk is that the research question is derivative of the medium itself. “The temptation for researchers who are examining a "new" form of communication is to simply describe the content rather than to place it in the context of theory and/or to test hypotheses” (McMillan, 2000, p.81), and indeed, early on in a medium’s development, it is easiest and arguably perhaps most appropriate simply to run descriptive studies. “However, researchers also need to move on to [other] purposes: making inferences as to antecedents of communication [such as authorship conditions and processes] and making inferences as to the effects of communication [such as effects on the user]” (McMillan, 2000, p.91).

Once we have gathered the content, one of the main challenges is to develop “useful and valid categorization schemes given the newness and novelty of the WWW” (Weare and Lin, 2000, p.284) and its multimodal interface. In the social sciences, researchers often draw their categorization schemes based on well-known genres and on the Web’s functionality. For instance, “in the analysis of political Web sites, researchers have recorded the presence of candidate information, press releases, party positions, and basic government documents. . . . In commercial sites, researchers have recorded the presence of product information, advertising, and types of promotional offers . . .” (Weare and Lin, 2000, p.285). In municipal sites, research has recorded the availability of information on “local politics, municipal service delivery, tourism, and economic development” (Weare and Lin, 2000, p.286). And of course many have recorded the presence or absence of Web functionality and multimedia, such as search functionality, e-mail links, feedback forms, graphics, and so forth. But trying to develop a categorization scheme that would fairly represent, say, the relative salience and emphasis of different kinds of info and features would be very challenging. For instance, if the e-mail link is at the bottom of the homepage, do we say that e-mail is not considered important because it’s under everything else, or do we say that the e-mail link is conscientiously placed right where people have come to expect it to be. In the words of one scholar, “researchers lack a sufficiently acute understanding of

'the syntax, semantics, and logic' . . . of multimedia, hypertexted language to permit the development of refined categorization schemes" (Weare and Lin, 2000, p.286). And even with such categorization schemes, we would still not begin to address the question of who, if anyone, answers the e-mail and the nature of their answer.

Conclusion

So quantitative methods, like any methods, have their challenges but also their benefits. . . .

References

Bosnjak, M, & Tuten, T.L. (2001). Classifying response behaviors in Web-based surveys. *Journal of Computer-Mediated-Communication, 6* (3).

Crawford, S.D., Couper, M.P., & Lamias, M.J. (2001). Web surveys: Perceptions of burden. *Social Science Computer Review, 19* (2), 146-162.

Manfreda, K.L., Batagelj, Z, & Vehovar, V. (2002). Design of Web survey questionnaires: Three basic experiments. *Journal of Computer-Mediated-Communication, 7* (3).

McMillan, S.J. (2000). The microscope and the moving target: The challenge of applying content analysis to the World Wide Web. *J&MC Quarterly, 77* (1), 80-98.

Odell, Goswami, and Herrington. "The Discourse-Based Interview: A Procedure for Exploring the Tacit Knowledge of Writers in Nonacademic Settings."

Pew Internet and American Life Project (2004, February 29). Content creation online. http://www.pewinternet.org/reports/toc.asp?Report=113

Sheehan, K. B. (2001). E-mail Survey Response Rates: A Review. *Journal of Computer-Mediated-Communication, 6* (2). <http://www.ascusc.org/jcmc/vol6/issue2/sheehan.html>.

Smith, C.B. (1997). Casting the Net: Surveying an Internet population. *Journal of Computer-Mediated-Communication, 3* (1).

Weare, C., & Lin, W. (2000) Content analysis of the World Wide Web: Opportunities and challenges. *Social Science Computer Review, 18* (3), 272-292.

Yun, G. W., & Trumbo, C. W. (2000). Comparative Response to a Survey Executed by Post, E-Mail, & Web Form. *Journal of Computer-Mediated-Communication, 6* (1).