(2004). *New Research for New Media: Innovative Research Methodologies Symposium Working Papers and Readings*, pp. 47-66.

Content Analysis for New Media: Rethinking the Paradigm

Susan C. Herring
Indiana University, Bloomington

Introduction

We live in exciting times. Researchers in the social sciences have available to them a panoply of new media as sources for research questions, data, and analytical methods to explore social, political, and cultural phenomena. New media, a term that came into prominence in the 1990s,[1] refers to any digital media production that is interactive and digitally distributed,[2] most typically via the Internet or the World Wide Web. This includes HTML-based websites and their variants (e.g., portals, news sites, weblogs, wikis), as well as computer-mediated communication (CMC)[3] (e.g., email, threaded discussion forums, chat rooms, instant messaging, text messaging via mobile phones). These two major paradigms of Internet-based media have attracted the interest of researchers in numerous disciplines in recent years, and bring with them a richness of fresh issues and phenomena for investigation.

As media of communication, Web pages and CMC lend themselves prima facie to Content Analysis, a social science methodology concerned broadly with "the objective, systematic, and quantitative description of the content of communication" (Baran, 2002). The earliest known application of Content Analysis was in the 17th century when the Church conducted a systematic examination of the content of early newspapers (Krippendorf, 1980). However, it was not until the 1940s and 1950s that Content Analysis became a well-established paradigm (Berelson, 1952; Berelson & Lazarsfeld, 1948). Its most prototypical uses have been by scholars in advertising, communication, and journalism to analyze written mass media content. With the advent of new media, however, Content Analysis is increasingly being applied by scholars in a wider range of disciplines, to a broader range of phenomena, and in a broader range of ways. These applications have benefited our understandings of new media, and expanded Content Analysis as a methodological paradigm. At the same time, the application of traditional methods to new phenomena raises challenges that must be acknowledged and met, if standards of rigor and interpretability are to be maintained.

In this overview, after contrasting traditional and alternative approaches to Content Analysis, I consider some particular challenges raised by the use of Content Analysis

[1] http://www.wordiq.com/definition/New_media

[2] http://www.sheridanc.on.ca/news/releases/newmedia/definition.html

[3] The term CMC is sometimes used to include Web-based communication, especially interactive forms such as weblogs. The two-way distinction in this paper assumes the more common definition of CMC as symmetrically interactive modes of online communication.

methods to analyze the Web and CMC, illustrating them with examples of research of each type. Some of the examples are drawn from papers accepted for the New Research for new Media symposium in Tarragona, Spain, held from September 30 to October 2, 2004, and made available to me in advance. Others are drawn from the growing body of content analysis research published by new media scholars. Taken together, these examples illustrate the range of challenges raised by analysis of new media content, including issues of definition, sampling, and research ethics, and the often times innovative solutions found to address them (cf. Mitra & Cohen, 1999; Wakeford, 2000). They further illustrate how coming to terms with these challenges can affect the conceptualizations underlying Content Analysis itself, and shape its evolution as a methodological paradigm, in ways that blur the boundaries between Content Analysis and other methods, such as social network analysis (in the case of the Web) and discourse analysis (in the case of CMC). The paper concludes by encouraging creative yet responsible extensions of Content Analysis to shed light on new media phenomena.

Content Analysis

In this section, two approaches to Content Analysis of new media are contrasted. The first seeks to apply traditional CA methods as literally as possible to new media content, striving to overcome differences between new and old media in order to hold the methods constant. The second approach construes CA more broadly, incorporating new methods to meet the requirements of analyzing digital content.

The traditional approach

The literal approach is illustrated by McMillan (2000) in her discussion of the challenges of applying content analysis to the World Wide Web. Content Analysis traditionally involves a set of procedures that can be summarized in five steps (cf. Krippendorf, 1980):

1) The researcher formulates a research question and/or hypotheses
2) The researcher selects a sample
3) Categories are defined for coding
4) Coders are trained, code the content, and the reliability of their coding is checked
5) The data collected during the coding process are analyzed and interpreted.

McMillan (2000) advocates adhering to these procedures and their traditional realizations as closely as possible when analyzing Web content. For example, research questions should be "narrowed" from the many new questions the Web raises, and a context should be found for them "either in existing or emerging communication theory" (p. 2). Following Krippendorf (1980, p. 66), McMillan states as a requirement for sampling that "within the constraints imposed by available knowledge about the phenomena, each unit has the same chance of being represented in the collection of sampling units"—that is, the sample ideally should be random. In defining coding categories, she suggests that researchers might apply traditional categories of content identified in old media studies (e.g., Bush, 1951), and that standard units of context are needed, analogous to those developed in traditional media (e.g., the column-inch for newspapers, time measured in seconds for broadcast). Multiple coders should be trained in advance on a portion of a sample, and established tools for checking intercoder reliability (such as Holsti's index and Scott's Pi) should be employed. While McMillan recognizes and discusses possible ways to overcome specific challenges the Web raises to realizing each of these goals, the goals themselves are not fundamentally questioned. Finally, although McMillan does not believe that the Web poses new challenges as regards analyzing and interpreting research findings, she cautions against the use of statistical methods that assume a random sample (given the difficulty of identifying a true random sample on the Web), concluding that "new communication tools are not an excuse for ignoring established communication research techniques" (p. 20).

Underlying these recommendations is a concern for rigor and standardization that undeniably has an important role to play in new media analysis. Rather than reinventing the methodological wheel, new media researchers can draw upon, and benefit from, well-established traditions. Moreover, the more similar the methods we apply to new media are to those used for old media, the easier it is to compare findings for the two.

At the same time, the narrowness of the above view can be problematic. First, the above procedures are not always strictly applied even in the analysis of old media. Exploratory (rather than theoretically pre-focused) studies are undertaken, non-random samples are used, and coding categories are allowed to emerge from traditional print communication; these methods are considered legitimate in many circumstances (see Bauer, 2000, for a broader view of "classical" content analysis). Such practices are even more common in the analysis of new media content. Emergent phenomena require basic description; phenomena of interest cannot always be identified in advance of establishing a coding scheme; and the dynamic nature and sheer number of units of Internet analysis makes random sampling infeasible in some cases, as McMillan also notes. Indeed, out of 19 content analyses of the Web that McMillan (2000) surveys, most fail to adhere to strict CA prescriptions. This does not necessarily render the results of such research useless or invalid, however.

A final problem is that some important types of Internet content, such as textual conversations and hyperlinks, lend themselves to analysis using paradigms that make use of different procedures than do traditional content analysis. Paradigms such as discourse analysis and social network analysis are better suited to analyzing conversations and linking patterns, respectively. It seems desirable to be able to integrate different methods into the analysis of, for example, the content of a multimodal website, rather than stopping the analysis where traditional content analysis methods leave off. For these reasons, a broader methodological perspective is needed.

Non-traditional approaches

An example of an approach to analyzing Internet content that extends the traditional notion of what CA is and how it should be applied is Computer-Mediated Discourse

Analysis (CMDA).[4] The basic methodology of CMDA has been described as "language-focused content analysis" supplemented by a toolkit of discourse analysis methods adapted from the study of spoken conversation and written text analysis (Herring, 2004). As in the more general practice of discourse analysis, methods employed can be quantitative (involving coding and counting) or qualitative; the former can resemble classical content analysis, but a broader spectrum of approaches is also included. Thus CMDA is both a variant of CA (broadly defined), and CA (narrowly-defined) is a variant of CMDA.

As regards the implementation of the "coding and counting" approach to CMDA, Herring (2004) lays out a five-step process that resembles that of classical CA:

1) Articulate research question(s)
2) Select computer-mediated data sample
3) Operationalize key concept(s) in terms of discourse features
4) Apply method(s) of analysis to data sample
5) Interpret results

However, in contrast with McMillan's exhortation that researchers closely follow established practice in order to insure rigor and interpretability, Herring (2004) recommends paradigm-independent best practices that rely on common sense (such as 'choose a research question that is in principle answerable from the available data'), and offers researchers options as regards sample types[5] (e.g., time-based, event-based, participant-based) and coding categories (e.g., pre-existing or emergent from the data), as determined by the research questions and data under consideration. The greatest challenge in CMDA, and the key to a compelling analysis, lies in operationalizing concepts of theoretical interest (Herring, 2004 gives the example of 'online community') in terms of measurable language behaviors, based on the premise that human behavior in CMC environments is carried out mostly through language and other semiotic means.

CMDA has been applied to the analysis of email, discussion forums, chat rooms, and text messaging, all of which can be considered forms of 'conversation.' It can also be applied to mediated speech (since discourse analysis is originally a spoken language paradigm), as well as textual Web content (Martinson, 2003). Finally, it can offer insight into the hypertextual nature of websites, through discourse methods associated with the analysis of 'intertextuality' (Hodsdon-Champeon, in press; Mitra, 1999). The patterns of interconnections formed by hyperlinks are also frequently addressed using methods of social network analysis.

[4] See, e.g., Herring (1997, 2001, 2004).

[5] Herring (2004) notes that "in CMDA, [sampling] is rarely done randomly, since random sampling sacrifices context, and context is important in interpreting discourse analysis results." For a similar reason, it is advocated that "textual analysis [ideally] be supplemented by ongoing participant observation" of online communication.

Social network analysis (SNA) can be considered CA in the broadest sense of the term; i.e., to the extent that hyperlinks comprise part of the content of a website (Schneider & Foot, 2004), SNA is a method well-suited to analyzing that particular form of content (Jackson, 1997). The approach, based in sociology, is quantitative and statistical, but rather than coding and counting links, it identifies patterns based on who (or what) is linking to whom (or what). Studies of linking patterns, sometimes in conjunction with other CA observations, have addressed a range of phenomena from community formation (Gibson, Kleinberg, & Rhagavan, 1998; Mitra, 1997) and authority and influence (Kleinberg, 1999) to website credibility (Fogg, Kameda, Boyd, et al., 2002) and Web genre structure (Bates & Lu, 1997; Chandler, 1998).

This section suggests that one way to make content analysis suitable for the analysis of new media is to incorporate methods from other disciplines. The justification for doing so is that new media combine modalities (e.g., text and graphics), enable new modes of communication (e.g., text-based conversation), and possess novel features (e.g., hyperlinks). However, while methods from other paradigms can help to address these characteristics, they are not enough in and of themselves. Rather, traditional content analysis must also adapt and expand in order to embrace the scope of phenomena included under new media content. This is because even in domains most central to traditional CA, new media pose special methodological challenges. These challenges include identifying units of analysis, sampling methods and frequency, and ethical issues associated with analyzing online communication. These are discussed below for the two main paradigms of online communication, the World Wide Web and CMC.

Analyzing Web Content

Units of analysis

The technical properties of any medium—combined with conventional patterns of use—tend to create natural units of analysis, and the Web is no exception. However, the spatial and temporal limits that constrain written text and television or radio broadcasting to more or less standard sizes do not apply to the Web, where seemingly limitless bandwidth and storage capacity enable some websites to be very large, while others with a similar purpose and focus may be only a single page. This variability makes 'the individual website' a problematic unit of analysis, both because large sites are time-consuming to browse through and code, and because their characteristics can end up overrepresented in the analysis due to their size alone.[6] It is also not always clear where the boundaries of a website lie—does the "site" consist of all pages on the server whose address is the URL of the home page, or should pages linked to from the site that were

[6] At least one study, by Crowston and Williams (2000), takes the individual Web page, rather than the website, as the unit of analysis, randomly sampling pages (which might be the home page or deeply embedded within a site) without regard to their context. This practice has not caught on, however.

created by the same site producer(s) on related topics be considered part of the site as well?[7]

One possible solution is to define the unit of analysis as all pages linked to from the home page at some fixed number of levels of embedding, to hold the size and nature of the sampling unit roughly constant. However, the number of pages included can still vary considerably according to the number of links at each level; nor does this address the problem of single-page websites. A more common solution is to take the homepage of each site as the basic unit of analysis (Bates & Lu, 1997). The homepage is the minimal unit that defines a website; it is the part that users are most likely to encounter, hence, arguably, the most salient and important part to analyze. However, while this practice enhances comparability across sites, it ignores potentially important content presented elsewhere in the site. Some researchers get around this limitation by applying some methods to the homepage, while taking the entire site as the context for the analysis of other types of content that do not necessarily appear on the homepage, such as personal information about the site producer (Herring, Scheidt, Bonus & Wright, 2004) or whether the site includes a discussion forum (Chung, 2004). Whatever unit of analysis is adopted, the results of analysis must be interpreted accordingly. For example, an analysis that looks only at the home pages of sites cannot claim that certain features do not occur in the sites as a whole, only that they do not occur on the pages examined.

Sampling

The number of Web pages is vast and in a dynamic flux, with new ones being created every moment, while others cease to function properly or are withdrawn. Not all extant pages can be analyzed; the researcher must extract a sample—ideally one that is representative of a larger whole. However, because the total number of sites is unknowable, randomly sampling the entire Web is impossible. Web crawlers work by taking a snowball sample, following links from one site to another, but miss the significant percentage of sites that are not linked to any others (Herring, Kouper, Paolillo, et al., 2005); hence Web crawls are not complete samples. Nor are indexes of websites, which are sometimes taken as base data for the purpose of drawing a random sample (e.g., Bates & Lu, 1997). However, thematically-related lists of links may be useful starting points for purposive sampling. The limits of any sample of websites must be recognized when generalizing from the results of analysis.

The dynamic nature of the Web also has consequences for how frequently one should sample in order to get a "representative" sense of what a site is like. Because websites vary in the frequency with which they are updated, this can be difficult to determine. Some genres of websites, by their nature, suggest a more-or-less regular update

[7] Schneider and Foot (2004) propose a larger unit of analysis, the 'Web sphere,' defined as "a hyperlinked set of dynamically defined digital resources spanning multiple Web sites deemed relevant or related to a central theme or 'object'." Such spheres of interconnected sites need not share the same base URL. An example of a Web sphere given by Schneider and Foot is all the sites associated with the 2000 presidential election in the United States.

frequency. For example, online syllabi and other course materials may be updated once a semester or once a year, while personal weblogs tend to be updated daily (Herring, et al., 2004). A challenging object of study in this regard is news websites, some of which claim to be updated "continuously." Keith, Schwalbe, and Silcock (2004) raise the methodological question of how often to sample images of the war in Iraq from news websites, in order to compare them with images from traditional media: print magazines, print newspapers, and television news. Of these four media, the Web-based news sites claim to be updated most frequently. But how frequently is "frequently?"

Relevant to this question, Kutz and Herring (2005) developed a micro-longitudinal approach to sample major news sites, including those for CNN and the BBC, every minute for a three-week period, and quantified the frequency of updates in top news stories. On average, they found that the top news story is replaced about every six to seven hours, and the photographic image associated with it changes about twice as often. These findings suggest that for top stories, at least, a sampling rate of every three or four hours would capture most images, and a sampling rate of every six or seven hours would capture at least one image from most stories. Even if the researcher does not sample exhaustively, the rate at which content is changed provides an important baseline against which to interpret the generalizability of a study's findings.

Ethical issues

The Web is for the most part a public, intentional means of communication; the contents of a site are available in principle for all Web users to see. Partly for this reason, Web content is often characterized as 'mass communication' or 'publication,' and researchers generally are not required to obtain the informed consent of site producers in order to study or quote content from their sites, provided the source is appropriately acknowledged. Nonetheless, there are grey areas as regards the ethical appropriateness of drawing attention to the content of a website through scholarly research. One example is weblogs maintained by adolescents, which often contain material of a personally revealing nature (Armstrong, 2003). Viégas (under review) surveyed weblog authors and found that many have an unrealistic perception of who their audiences are; while they may realize abstractly that anyone can read their weblogs, they tend to act as though only the people they want to read them (friends, people they want to impress, etc.) are doing so, and may respond with dismay when they learn that a teacher, parent, or employer has discovered the existence of their blog. Have such "authors" intentionally produced public documents, and should researchers thus be able to cite them without permission? This ethical grey area results from the relative novelty of the Web, on the one hand—norms of production and reception have not yet become fully established—and the inexperience of young content producers, to whom the Web provides an unprecedented opportunity to self-express on a mass scale.

Traditionally, institutional review boards require that an adult give permission on behalf of a minor for the minor to participate in a research study. However, in cases where a teenager's website may be a secret from the parents, this "solution" raises as many ethical problems as it solves. Researchers currently negotiate these issues with their institution's

IRB on a case-by-case basis. One compromise is to avoid quoting directly from, or showing images from, the websites analyzed, in order to preserve the site producer's anonymity, either with or without the producer's informed consent (Huffaker & Calvert, in press). However, other authors have operated on the assumption that weblogs are public, reproducing examples fully without asking permission (Scheidt & Wright, 2004). Of course, obtaining informed consent requires knowing how to contact the site producer, which can be difficult if no contact information is included on the site. Herring, et al. (2004), in a content analysis of a random sample of 203 weblogs, found that fewer than one-third included the author's email address on the homepage.

Analyzing CMC Content

Units of analysis

Computer-mediated communication takes many forms, including email, threaded discussion lists, chat rooms, and private messaging. Natural units of CMC are not so much defined by the specific form (mode) of CMC, however, as by synchronicity. Most asynchronous online communication—for which sender and receiver need not be logged on at the same time, such as email, discussion lists and newsgroups—takes place by means of *email messages*, sometimes grouped into topically-related *threads*, which may extend over days or months. Synchronous CMC—which involves more-or-less real time transmission and reception of messages, such as chat and IM—takes place via typically briefer *chat messages*, exchanged during a temporally-bounded *session* of chatting or IMing. Email messages, chat messages, threads and chat sessions constitute useful, technically-determined units of analysis for CMC content analysis. It is less clear, however, to what off-line units of communication these correspond. Email messages are somewhat, albeit not exactly, like memos (Yates & Orlikowski, 1992) and personal letters (Herring, 1996a). Threads resemble, albeit not perfectly, topics of conversation. Individual messages, especially chat messages, resemble turns, but with certain differences (Condon & Cech, in press). These near-analogous pairings must be understood for both their similarities and their differences when extending methods of spoken and written language analysis to CMC content, as is done in computer-mediated discourse analysis (CMDA).

Some researchers, such as Hodsdon-Champeon (in press, for asynchronous messages posted to Usenet newsgroups), have simply equated messages with turns. Condon and Cech (in press) adopt a more nuanced approach, finding through an experimental study that an email message often contains more than one functional turn, while a chat message may contain less than one functional turn. Herring (2003; Herring & Nix, 1997) also focuses on functional units in identifying topical relations among chat messages based on semantic connections, rather than on technically-defined threads. At the same time, technically-determined units of discourse in new media can be of interest in their own right, independent of whether they correspond closely to traditional discourse units. An example of this is the asymmetrical blog-entry-plus-comments structure that characterizes one form of "conversation" that can be embedded in weblogs (Herring, et al., 2005).

Sampling

Like the Web, the enormous number and dynamic nature of CMC messages render true random sampling of CMC messages on the Internet infeasible. Moreover, random sampling may be undesirable for several reasons. First, the amount of variation in CMC conditioned by technical mode and participant demographics is so great as to make the notion of a 'typical' CMC message virtually meaningless. At a minimum, synchronous messages should be separated from asynchronous messages, and HTML-based forms of asynchronous communication such as weblog entries should probably be separated from symmetrically-interactive forms of asynchronous CMC such as email messages before sampling is undertaken, so as to avoid inadvertently lumping together apples, oranges, and mangos. Second, CMC messages rarely appear in isolation; much of their meaning and pragmatic significance would be lost in the absence of the surrounding discourse context. Thus, random sampling of CMC messages makes little sense for most research purposes. The same is true for forums and other CMC environments; their technical affordances and social contexts shape their content; omitting consideration of such variables robs the researcher of potentially rich interpretative resources.

Accordingly, sampling in studies of CMC tends to be purposive, rather than random. Most researchers restrict their investigation to one or more CMC environments, sampling by thread (Herring, Johnson, & DiBenedetto, 1992), by a slice of time (Paolillo, 2001), by periodic intervals (Panyametheekul & Herring, 2003), by event (e.g., messages posted on or about September 11, 2001 in the US), or by participant demographics (e.g., messages posted by one gender). Waldman and Storey (2004) employ participant-based sampling, extracting all interactions from a larger corpus of dyadic chat interviews between troubled teens and social workers in which certain subjects of interest participated. As Herring (2004) notes, there are advantages and limitations associated with each method of sampling that must be borne in mind when interpreting the results of the analysis.

Ethical issues

CMC frequently blurs the distinction between public and private communication. As in the case of weblogs, discussion forum and chatroom participants may not intend their messages to be read by a wider audience, even though they are posted to forums that anyone can access. This observation led King (1996) to advocate that researchers should respect participants' 'perceived privacy' by not directly citing their messages, even if the messages are anonymized. A unique characteristic of text-based CMC is that it leaves a persistent record that can easily be searched, as in the case of messages posted to Usenet newsgroups, which are retained in a Google-searchable archive. Thus anonymizing messages to remove personal identifiers may not be sufficient to prevent CMC participants quoted in a research study from being identified. In contrast, others have argued that the persistent, publicly-accessible nature of CMC transcripts makes them public for all practical purposes, and that the onus is on participants to avoid making incriminating statements in open-access chatrooms and discussion forums (Herring, 1996b).

'Private' CMC environments pose other challenges, as does CMC—public or private—by minors. Email messages exchanged between two individuals may be intended as private communication, but the persistence of the messages and the ease with which they can be copied and forwarded means that such messages sometimes end up in public forums without the original author's permission. What should a researcher do with a sample that contains such embedded private messages? If permission must be sought, what if the original author's contact information is not available? Obtaining informed consent can be problematic for other reasons, as well. In the US, permission is supposed to be obtained (from parents) to study minors, as noted above, but teens participating in chatrooms may use pseudonyms that mask their identities; even if they are contactable, they might not wish their parents to know of their chatroom activities; and even if they don't mind their parents knowing, the logistics of contacting the parents to obtain their consent in the course of a synchronous chat session are daunting to the point of rendering a study of spontaneous online adolescent chat infeasible. Waldman and Storey (2004) grapple with a similar challenge in analyzing archives of private chat interviews between troubled teens and social workers on a British youth support site. Because the subjects are minors, because the context is private, and because the content of the interviews is oftentimes sensitive, informed consent should normally be sought. However, the identities of many of the teens are not recoverable because they use online pseudonyms; no contact information is available for them; and the content of their interviews should not be revealed to their parents in any case, in keeping with standards of professional confidentiality. In such a case, one may ask: if the participants cannot be identified, does researching their interviews without their knowledge pose any potential harm to them?

Fortunately for researchers, IRBs in the US and elsewhere are loosening the requirement of informed consent for publicly-accessible CMC, even for adolescents, provided that care is taken not to reveal participants' identities or otherwise expose them to harm through the research activities.[8] A similar procedure may soon become acceptable for private but anonymous CMC logs, if it has not already become so. In situations where it is possible to contact the participants and their content is potentially sensitive, even if publicly posted, various practices have been employed. Reid obtained permission in advance from community members to study the discourse of a chat community (1996). Scharf (1999) asked individual participants after the fact for permission to quote their messages to a breast cancer survivor support newsgroup in her research. Herring, Job-Sluder, Scheckler, and Barab (2000), in their study of a disruptive "troller" in a feminist discussion forum, chose not to contact the troller for permission, but anonymized all information that could identify him. Since the messages from the discussion forum were not publicly searchable, this method effectively concealed the troller's identity from all except the other participants in the feminist forum.

[8] For a synopsis of current thinking by Internet researchers about best ethical practices, see AoIR (2002).

New Media, New Paradigm?

The discussion in the previous sections shows that even the most basic aspects of content analysis of new media, such as defining and selecting content for analysis, raise challenges to the traditional CA paradigm, narrowly defined. Notions of units of analysis comparable to those in traditional media, random sampling, and informed consent must all be rethought in the context of new media research. These observations, taken together with the desirability of incorporating new methods to address characteristic features of new media such as hyperlinks and textual conversations, suggest a need for a broader construal of CA that allows new media to dictate new methods tailored to the analysis of digital content.

This broad construal assumes a more general definition of content than is typically found in traditional content analysis. 'Content' in CA tends to refer specifically to the thematic meanings present in text or images, but not to 'structures' or 'features' of the communicative medium (Schneider & Foot, 2004). In contrast, the approach to content analysis proposed here considers content to be various types of information "contained" in new media documents, including themes, structures, features, links and exchanges, all of which can communicate social, political and cultural meanings. Along with this broader definition comes a broadening of the methodological paradigm; "coding and counting" methods must be supplemented with other techniques, in order to analyze the contributions of different semiotic systems to the meaning of multimedia documents, rather than leaving out of consideration such phenomena as hyperlinks and textual exchanges, on the grounds that they do not constitute familiar forms of content. These other techniques include some not yet devised, but which will no doubt arise to address the characteristics of new media communication as it continues to evolve in new directions.

Conclusion

In this overview, I have contrasted what I believe to be an overly-narrow application of traditional content analysis methods to new media with an alternative conceptualization of what content analysis could, and I believe, should, become in response to the challenges raised by the Internet and the World Wide Web.

In opening up the paradigm in these ways, there is a risk that methodological precision and interpretability of research results may suffer in the short term. Analyses may not be comparable across researchers; some may be ad hoc. It may be difficult to appreciate initially how an analysis involving methodological innovation is representative and reproducible—the criteria for "robust" analysis (Schneider & Foot, 2004).

Over time, however, methods tend to become systematized and best practices refined. As more research on the communicative content of new media (in its myriad forms) is carried out, the knowledge created will inform future analyses. For example, the results of micro-longitudinal analysis of news site update frequency can be used to determine appropriate sampling rates in future studies. More such research needs to be done in order to determine typical update frequencies for a variety of Web genres.

Furthermore, new media themselves will stabilize. As website genres become more conventionalized over time, their sizes and formats will become increasingly standardized, facilitating the selection of units of analysis.[9] More complete indexes and archives of Web content will become available (Lyman & Kahle, 1998), facilitating sampling. Users' understandings of new media conventions will also inevitably grow more sophisticated, making it less likely that they will imagine they are communicating to a restricted audience when posting to a chatroom or a weblog. This increased awareness should make the demarcation more evident between publicly accessible content and content for which permission must be obtained by researchers.

As the expanded content analysis paradigm envisioned here advances towards these outcomes, it will not only be more systematic and rigorous. It will ultimately be more powerful for having integrated innovative responses to new media phenomena in its formative stages.

[9] This is not to suggest that units of Web communication can ever be directly comparable with units of traditional mass media communication. Different media are different. (Compare the packaging of newspaper and television messages; Keith, et al., 2004).

References

Armstrong, E. (2003). Do you blog? *The Christian Science Monitor*, May 13. http://www.csmonitor.com/2003/0513/p11s01-lecs.html

AoIR (Ethics Working Committee). (2002). *Ethical Guidelines for Internet Research.* http://www.aoir.org/reports/ethics.pdf

Baran, S. J. (2002). *Introduction to Mass Communication*, 2nd ed. New York: McGraw-Hill.

Bates, M. J., & Lu, S. (1997). An exploratory profile of personal home pages: Content, design, metaphors. *Online and CDROM Review*, 21 (6): 331-340.

Bauer, M. (2000). Classical content analysis: A review. In M. W. Bauer & G. Gaskell (Eds.), *Qualitative Researching with Text, Image, and Sound: A Practical Handbook* (pp. 131-151). London: Sage Publications.

Berelson, B. (1952). *Content Analysis in Communication Research*. New York: Free Press.

Berelson, B., & Lazarsfeld, P. F. (1948). *The Analysis of Communication Content*. Chicago and New York: University of Chicago and Columbia University.

Bush, C. R. (1951). The analysis of political campaign news. *Journalism Quarterly, 28* (2): 250-252.

Chandler, D. (1998). Personal homepages and the construction of identities on the Web. http://www.aber.ac.uk/~dgc/webident.html

Chung, D. S. (2004). *Toward Interactivity: How News Websites Use Interactive Features and Why it Matters*. Unpublished Ph.D. dissertation, School of Journalism, Indiana University, Bloomington.

Condon, S. L., & Cech, C. G. (In press). Discourse management in three modalities. In S. C. Herring (Ed.), *Computer-Mediated Conversation*. Cresskill, NJ: Hampton Press.

Crowston, K., & Williams, M. (2000). Reproduced and emergent genres of communication on the World Wide Web. *The Information Society, 16* (3): 201-215.

Fogg, B. J., Kameda, T., Boyd, J., Marshall, J., Sethi, R., Sockol, M., & Trowbridge, T. (2002). *Stanford-Makovsky Web Credibility Study 2002: Investigating what makes Web sites credible today*. http://captology.stanford.edu/pdf/Stanford-MakovskyWebCredStudy2002-prelim.pdf

Gibson, G., Kleinberg, J., & Raghavan, P. (1998). Inferring Web communities from link topology. *Proc. 9th ACM Conference on Hypertext and Hypermedia.*

Herring, S. C. (1996a). Two variants of an electronic message schema. In S. C. Herring (Ed.), *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives* (pp. 81-108). Amsterdam: John Benjamins.

Herring, S. C. (1996b). Linguistic and critical research on computer-mediated communication: Some ethical and scholarly considerations. *The Information Society*, *12* (2): 153-168.

Herring, S. C., Ed. (1997). *Computer-Mediated Discourse Analysis*. Special issue of the *Electronic Journal of Communication, 6* (3). http://www.cios.org/www/ejc/v6n396.htm

Herring, S. C. (2001). Computer-mediated discourse. In D. Schiffrin, D. Tannen, & H. Hamilton (Eds.), *The Handbook of Discourse Analysis* (pp. 612-634). Oxford: Blackwell Publishers.

Herring, S. C. (2003). Dynamic topic analysis of synchronous chat. Paper presented at the *Symposium on New Research for New Media*, University of Minnesota, Minneapolis, September 5.

Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. In S. A. Barab, R. Kling, & J. H. Gray (Eds.), *Designing for Virtual Communities in the Service of Learning*. New York: Cambridge University Press.

Herring, S. C., Job-Sluder, K., Scheckler, R., & Barab, S. (2002). Searching for safety online: Managing "trolling" in a feminist forum. *The Information Society, 18* (5): 371-383.

Herring, S. C., Johnson, D. A., & DiBenedetto, T. (1992). Participation in electronic discourse in a ‘feminist’ field. In *Locating Power: Proceedings of the 1992 Berkeley Women and Language Conference* (pp. 250-262). Berkeley: Berkeley Women and Language Group.

Herring, S. C., Kouper, I., Paolillo, J., Scheidt, L. A., Tyworth, M., Welsch, P., Wright, E., & Yu, N. (2005). Conversations in the blogosphere: An analysis "from the bottom up." *Proceedings of the Thirty-Eighth Hawai'i International Conference on System Sciences* (*HICSS 38*). Los Alamitos: IEEE Computer Society Press.

Herring, S. C., & Nix, C. G. (1997). Is 'serious chat' an oxymoron? Paper presented at the *Annual Association of Applied Linguistics*, Orlando, FL, April 11.

Herring, S. C., Scheidt, L. A., Bonus, S., & Wright, E. (2004). Bridging the gap: A genre analysis of weblogs. *Proceedings of the 37th Hawai'i International Conference on System*

Sciences (*HICSS-37*). Los Alamitos: IEEE Computer Society Press. http://www.blogninja.com/DDGDD04.doc

Hodsdon-Champeon, C. B. (In press). Conversations within conversations: Intertextuality in racially antagonistic dialogue on Usenet. In S. C. Herring (Ed.), *Computer-Mediated Conversation*. Cresskill, NJ: Hampton Press.

Huffaker, D. A., & Calvert, S. L. (In press). Gender, identity and language use in teenage blogs. *Journal of Computer-Mediated Communication, 10* (2).

Jackson, M. (1997). Assessing the structure of communication on the World Wide Web. *Journal of Computer-Mediated Communication, 3* (1). http://www.ascusc.org/jcmc/vol3/issue1/jackson.html

Keith, S., Schwalbe, C., & Silcock, W. (2004). Comparing new media with old: Equivalency challenges for content analysis. Paper presented at the New Research for New Media Symposium, Tarragona, Spain, September 30-October 2.

King, S. A. (1996). Researching Internet communities: Proposed ethical guidelines for the reporting of results. *The Information Society, 12* (2): 119-128.

Kleinberg, J. M. (1999). Hubs, authorities and communities. *ACM Computing Surveys*, 31 (4): article 5.

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology*. Newbury Park: Sage.

Kutz, D. O., & Herring, S. C. (2005). Micro-longitudinal analysis of Web news updates. *Proceedings of the Thirty-Eighth Hawai'i International Conference on System Sciences* (*HICSS 38*). Los Alamitos: IEEE Computer Society Press.

Lyman, P., & Kahle, B. (1998). Archiving digital cultural artifacts: Organizing an agenda for action. *D-Lib Magazine*. http://www.dlib.org/dlib/july98/07lyman.html

Martinson, A. (2003). Ideological discourse on the Web: The pornography debates. Paper presented at *Internet Research 4.0*, Toronto, CA, October 15-19.

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

Mitra, A. (1997). Diasporic Web sites: Ingroup and outgroup discourse. *Critical Studies in Mass Communication, 14:* 158-181.

Mitra, A. (1999). Characteristics of the WWW text: Tracing discursive strategies. *Journal of Computer-Mediated Communication*, 5 (1).

http://www.ascusc.org/jcmc/vol5/issue1/mitra.html

Mitra, A., & Cohen, E. (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.

Panyametheekul, S., & Herring, S. C. (2003). Gender and turn allocation in a Thai chat room. *Journal of Computer-Mediated Communication*, *9* (1). http://www.ascusc.org/jcmc/vol9/issue1/panya_herring.html

Paolillo, J. C. (2001). Language variation in the virtual speech community: A social network approach. *Journal of Sociolinguistics*, *5* (2): 180-213.

Reid, E. (1996). Informed consent in the study of on-line communities: A reflection on the effects of computer-mediated social research. *The Information Society*, *12* (2): 169-174.

Scharf, B. (1999). Beyond Netiquette: The ethics of doing naturalistic discourse research on the Internet. In S. Jones (Ed.), *Doing Internet Research: Critical Issues and Methods for Examining the Net* (pp. 243-256). London: Sage.

Scheidt, L. A., & Wright, E. (2004). Common visual design elements of weblogs. In L. Gurak, S. Antonijevic, L. Johnson, C. Ratliff, & J. Reyman (Eds.), *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. http://blog.lib.umn.edu/blogosphere/

Schneider, S. M., & Foot, K. A. (2004). The Web as an object of study. *New Media & Society, 6* (1): 114-122.

Viégas, F. (Under review). Bloggers' expectations of privacy and accountability.

Wakeford, N. (2000). New media, new methodologies: Studying the Web. In D. Gauntlett (Ed.), *Web.Studies: Rewiring Media Studies for the Digital Age* (pp. 31-42). London: Arnold.

Waldman, J., & Storey, A. (2004). Evaluation of There4me – Some methodological challenges. Paper presented at the New Research for New Media Symposium, Tarragona, Spain, September 30-October 2.

Yates, J., & Orlikowski, W. J. (1992). Genres of organizational communication: A structurational approach to studying communication and media. *Academy of Management Review, 17* (2): 299-326.