(Forthcoming, 2008). In J. Hunsinger, M. Allen, & L. Klastrup (Eds.), *The International Handbook of Internet Research*. Springer Verlag.

Web Content Analysis: Expanding the Paradigm

Susan C. Herring
Indiana University, Bloomington

Abstract

Are established methods of content analysis (CA) adequate to analyze web content, or should new methods be devised to address new technological developments? This chapter addresses this question by contrasting narrow and broad interpretations of the concept of web content analysis. The utility of a broad interpretation that subsumes the narrow one is then illustrated with reference to research on weblogs (blogs), a popular web format in which features of HTML documents and interactive computer-mediated communication converge. The chapter concludes by proposing an expanded Web Content Analysis (WebCA) paradigm in which insights from paradigms such as discourse analysis and social network analysis are operationalized and implemented within a general content analytic framework.

Introduction

Since the introduction of the first graphical browser in 1993, the system of interlinked, hypertext documents known as the World Wide Web (hereafter, "the web") has grown to be the primary multimodal content delivery system on the internet; indeed, today, it is one of the largest content delivery vehicles in the history of the world. Along with this increase in volume, technical web document and website types have also proliferated. From their beginnings as static HTML documents comprised mainly of text, links, and graphics, web pages have added sound, animations, and video; they have incorporated user-interface, user-content, and user-user interactivity features (including, in the latter category, email, discussion forums, chat, and Voice-over-IP); and they have generally converged with other online and offline media to produce hybrid genres such as online news sites, blogs, wikis, photo- and video-sharing sites, and social network sites.

The abundance of web pages and their diversity of form and function (as well as the unprecedented ease with which content can be collected and analyzed using automated tools) provide seemingly endless opportunities for research. At the same time, these characteristics can be daunting to researchers wishing to analyze web content. What methods should one use, and how should they be implemented? Will established methods serve, or should new methods be devised to address new technological phenomena? If new methods are coined, how can their validity and consistency of application across researchers be insured? This is important if internet research is to be taken seriously, and if the results of analysis of web content are to be comparable with previous analyses of content in other media.

Content analysis is an established social science methodology concerned broadly with "the objective, systematic, and quantitative description of the content of communication" (Baran, 2002, p. 410; see also Berelson, 1952). As media of communication, websites and web pages lend themselves prima facie to content analysis (Weare & Lin, 2000). Indeed, content analysis (henceforth, CA) was one of the first methodologies used in web analysis (e.g., Bates & Lu,

1997), and it has been employed increasingly since, albeit not always in traditional ways (McMillan, 2000).

This chapter addresses the question of how strictly internet research should embrace traditional CA methods when analyzing web content, as opposed to incorporating methodological innovation, including drawing on methods not traditionally considered CA. Narrow and broad interpretations of the concept of web content analysis are first contrasted and exemplified with relevant scholarship. The utility of a broad interpretation that subsumes the narrow one is then illustrated with reference to research on weblogs (blogs), a popular web format in which features of HTML documents and interactive computer-mediated communication converge (Herring, Scheidt, Bonus, & Wright, 2004, 2005). Examples from the literature are provided of traditional and non-traditional blog content analyses and the methodological challenges they face. It is argued that coming to terms with these challenges can affect the conceptualizations underlying content analysis as a methodological paradigm, in ways that blur the boundaries between CA and other methods, such as discourse analysis and social network analysis. The chapter concludes by proposing an expanded Web Content Analysis (WebCA) paradigm in which insights from other paradigms are operationalized and implemented within a general CA framework.

Content Analysis

Content analysis is a systematic technique for coding symbolic content (text, images, etc.) found in communication, especially structural features (e.g., message length, distribution of certain text or image components) and semantic themes (Bauer, 2000). While the primary use of CA is to identify and describe patterns in manifest content—what the audience perceives through the senses, rather than what it feels or believes as a result of that content, or what the content producer intended—the technique can also be used for making inferences about intentions and effects (Holsti, 1969; Krippendorf, 1980).

According to Krippendorf (1980), 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. 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 the analysis of written mass media content by scholars of advertising, communication, and journalism. However, in recent decades, CA techniques have also been used increasingly to analyze content on the internet. Perhaps due to its original presentation as a one-to-many broadcast medium, the web has attracted an especially large number of studies that employ content analysis methods.

Web Content Analysis

The phrase "web content analysis" is in fact ambiguous. It can be interpreted in two different senses, the second of which subsumes the first: 1) the application of traditional CA techniques, narrowly construed, to the web [web [content analysis]] and 2) the analysis of web content, broadly construed, using various (traditional and non-traditional) techniques

[[web content] analysis]. Both of these senses are represented in the web analysis literature, as discussed below.

A traditional approach

The first sense of web content analysis is explicitly argued for by McMillan (2000), who adopts a traditional approach in her discussion of the challenges of applying CA to the web. Drawing on Krippendorf (1980), she notes that CA traditionally involves a set of procedures that can be summarized in five steps:

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.

With regard to the first step*, 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.[1] In defining *coding categories*, she implies that a standard list of categories would be desirable and hints that researchers might apply established categories of content identified in old media studies (e.g., Bush, 1951). Standard units of context are also needed, analogous to those developed in traditional media (cf. the column-inch for newspapers, time measured in seconds for broadcast).

As regards the fourth step, multiple *coders* should be trained in advance on a portion of a sample, and established methods for calculating intercoder reliability (such as Scott's pi and Holsti's reliability index)[2] should be employed. Finally, although McMillan does not believe that the web poses new challenges as regards the fifth step—*analyzing and interpreting* research findings—she cautions against the “inappropriate” use of statistical tests that assume a random sample (which includes the most frequently-used statistical tests), given the difficulty of identifying/constructing a statistically random sample on the web.

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. She concludes 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, both of which are undeniably important when seeking to establish the credibility and validity of a new

research enterprise. Rather than reinventing the methodological wheel, internet and web researchers can draw upon, and benefit from, well-established traditions. Further, the more similar the methods that are applied to new media are to those used to analyze old media, the easier it is to compare findings in order to attain broader, trans-media understandings.

Problems with the traditional approach

At the same time, the narrowness of the above view can be problematic. First, Krippendorf's procedures, as interpreted by McMillan (2000), are rarely followed strictly, even in analyses of old media. Exploratory (rather than theoretically pre-focused) studies are undertaken, non-random samples are used,[3] coding categories are allowed to emerge from data—indeed, this is the cornerstone of the grounded theory approach,[4] which is followed in many content analysis studies—, and standard statistical tests are applied to non-random samples in studies of traditional modes of communication. Moreover, these methods are considered legitimate in many circumstances (see Bauer, 2000, for a broader conceptualization of "classical" content analysis).

Such practices are also common in the analysis of new media content, where they may be justified by the nature of the phenomena under investigation. Emergent phenomena require basic description, and phenomena of interest cannot always be identified in advance of establishing a coding scheme—the intermingling of channels of communication on websites may especially require novel coding categories. Moreover, the dynamic nature and sheer number of units of internet analysis makes random sampling infeasible in many cases, as McMillan and others (e.g., Schneider & Foot, 2004; cf. Weare & Lin, 2000) have also noted. Indeed, out of 19 content analysis studies of the web that McMillan (2000) surveyed, most failed to adhere to her strict CA prescriptions. This does not necessarily render the results of such research useless or invalid, however.

Similarly, recent web studies that identify their primary methodology as content analysis also vary in the degree to which they adhere to McMillan's (2000) prescriptions. For example, an informal examination of CA articles published between 2004 and 2007 in the *Journal of Computer-Mediated Communication*, a leading journal for social science research on internet and web communication, reveals studies in which research questions are indeed grounded in traditional theory, multiple coders are used to establish interrater reliability, and coding schemes are adapted from previous communication research (e.g., Singh & Baack, 2004; Waseleski, 2005). However, most of the studies analyze non-random samples (Dimitrova & Neznanski, 2006; Pfeil, Zaphiris, & Ang, 2006; Singh & Baack, 2004; Waseleski, 2005; Young & Foot, 2005), and many invent new coding schemes (e.g., Dimitrova & Neznanski, 2006; Pfeil et al., 2006; Young & Foot, 2005). This suggests the possibility that the 19 articles surveyed by McMillan (2000) do not simply represent an earlier, methodologically less rigorous, phase of web content analysis research, but rather that web content analysis may be following somewhat different norms from those traditionally prescribed for the analysis of communication content by researchers such as Krippendorf and McMillan, or even evolving new norms.

Most challenging to the traditional view, a growing number of web studies analyze types of content that differ from those usually studied in CA—such as textual conversations and hyperlinks—using methodological paradigms other than traditional CA. Although one possibility would be to exclude such studies from consideration in discussions of content analysis, it seems desirable to be able to integrate different methods into the analysis of the content of a multimodal website, rather than stopping the analysis where traditional content analysis methods leave off. For these purposes, a broader methodological perspective is needed.

Non-traditional approaches

A number of new media researchers have argued that new communication technologies call for new methods of analysis (e.g., Mitra & Cohen, 1999; Wakeford, 2000). Here it is assumed that any approach to web content analysis that aims to cover a broad range of content should include, at a minimum, methods that allow for the systematic identification of patterns in link and interactive message content, since these types of content are increasingly prevalent on the web. To fulfill this aim, some researchers draw on methodological paradigms from disciplines outside communication. Two non-traditional approaches that claim connections with CA are considered below, one grounded in linguistics and the other in sociology. Computational techniques also increasingly inform the analysis of web content, although they are not usually characterized by their practitioners as CA.

Computer-mediated discourse analysis

One 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). The basic methodology of CMDA is described by Herring (2004) as language-focused content analysis supplemented by a toolkit of discourse analysis methods adapted from the study of spoken conversation and written text analysis. As in the more general practice of discourse analysis, the 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 sub-type of CA (broadly defined), and CA (narrowly-defined) is a sub-type of CMDA.

Regarding the implementation of the "coding and counting" approach to CMDA, Herring (2004) lays out a five-step process that resembles that for 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 to McMillan's (2000) exhortation that researchers closely follow established practice in order to insure rigor and interpretability, Herring (2004) takes a pragmatic view, recommending paradigm-independent best practices, such as: choose a

research question that "is empirically answerable from the available data" [p. 346]. She also 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 the data under consideration. The greatest challenge in CMDA, and the key to a compelling analysis, lies in operationalizing concepts of theoretical interest (Herring, 2004 elaborates the example of "virtual community") in terms of measurable language behaviors, based on the premise that human behavior in CMC environments is carried out mostly through linguistic means. The importance of qualifying the interpretation of research findings in light of the characteristics of the data sampled is also emphasized.

CMDA has been applied to the analysis of email, discussion forums, chat rooms, and text messaging, all of which are forms of dialogue (or polylogue). It can also be applied to mediated speech (since discourse analysis is originally a spoken language paradigm), as well as to monologue text on web pages (Kutz & Herring, 2005). Finally, it can offer insight into the hypertextual nature of websites, through discourse methods associated with the analysis of intertextuality, or content that refers to content in other texts (Mitra, 1999). Patterns of interconnections formed by hyperlinks are also frequently addressed using methods of social network analysis.

Social network analysis

Social network analysis (SNA) could be considered CA in the broad sense of the term, in that it can be used to analyze hyperlinks, which are part of the content of websites—indeed, some would argue that links are the essence of the web (Foot, Schneider, Dougherty, Xenos, & Larsen, 2003). Classical SNA as employed by sociologists is quantitative and statistical; it is used to analyze networks of ties (e.g., as constituted by communication or transaction) between nodes (e.g., people, institutions). SNA is also well suited for analyzing patterns of linking on the web: Websites can be considered nodes, links can be considered ties, and the arrangements of links within and across sites can be represented as networks (Jackson, 1997).

While most SNA does not call itself CA, a hybrid approach known as link analysis blurs the boundaries between the two. Links are part of the manifest content of web pages, and as such are sometimes included in coding and counting studies of web features (Bates & Lu, 1997; Dimitrova & Neznanski, 2006). The nature of a link in terms of the site it connects to (sometimes called the link destination) has also been coded and analyzed in studies of website credibility (Fogg, Kameda, Boyd, et al., 2002) and political affiliation (Foot et al., 2003). Further, patterns of linking within and across websites have been analyzed as indicators of phenomena ranging from academic quality (Thelwall, 2002) to community formation (Gibson, Kleinberg, & Rhagavan, 1998).

Research by Kirsten Foot and Steven Schneider illustrates an approach to link analysis with close affinities to CA. With Park (2003), Foot et al. (2003, n.p.) assert that "hyperlinks are inscriptions of communicative and strategic choices on the part of site producers," similar to other types of web content. Their "mid-range" approach involves "systematic human coding and interpretation of linked-to producer types" in political candidate websites. Multiple, trained coders evaluated links for the presence or absence of connection to certain types of

website, and interrater reliability was calculated using Krippendorf's alpha, consistent with standard CA practice.

Schneider and Foot (2004) also analyze networks of links across websites, as in SNA, within constellations that they term "web spheres." A web sphere is "a hyperlinked set of dynamically-defined digital resources that span multiple websites and are deemed relevant, or related, to a central theme or 'object'" (p. 118). 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.

The above sections suggest that perspectives from other disciplines can be incorporated into traditional CA, while still preserving many of its essential characteristics (e.g., classification and quantification; interrater reliability assessment). The relationships among the approaches summarized above as applied to the analysis of web content are represented in Figure 1. In the figure, content analysis is listed under Communication (although it is also used in other disciplines) to simplify the presentation, and traditional CA is referred to as Theme/Feature Analysis to indicate the types of content it is typically used to address.

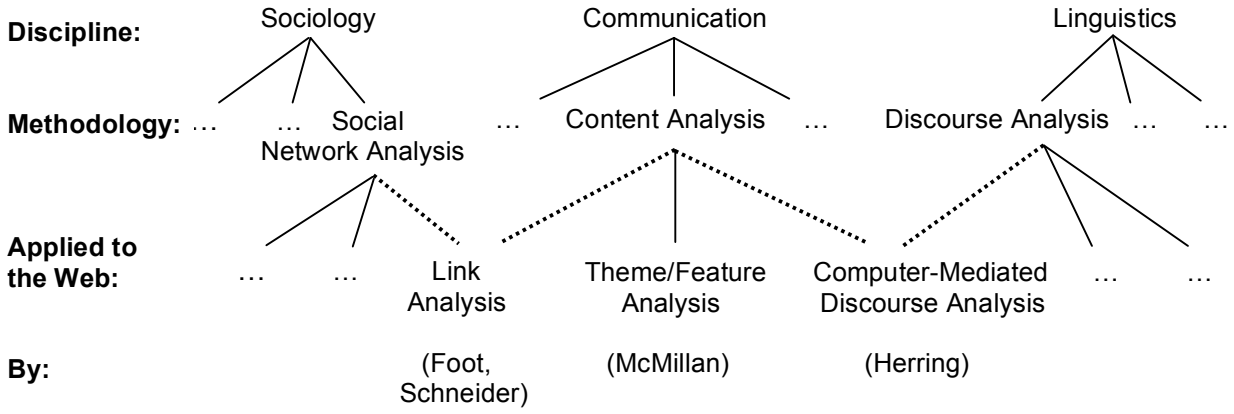

Figure 1. Some approaches to analyzing web content

Both narrow and broad CA approaches can be useful in analyzing web content, as illustrated in the next section for the content of one popular web format, the weblog.

Analyzing Blog Content

A weblog (blog, for short) is a type of web document in which dated entries appear in reverse chronological sequence. Blogs started to become popular after the introduction of the first free blogging software in 1999 and entered mainstream awareness after bloggers' commentary on the September 21, 2001 terrorist attacks and the 2003 U.S.-led war on Iraq attracted widespread media attention (Blood, 2002; Herring, Scheidt, Kouper, & Wright, 2006). As of mid-2008, blogs worldwide numbered in the hundreds of millions.[6]

Like other web documents, blogs can be multimodal or purely textual, and variants exist that feature photos, voice recordings (audio blogs), and videos (vlogs). Multimodality poses

challenges to content analysis, especially as regards the identification of units of analysis (Weare & Lin, 2000). However, traditional CA has been applied to the analysis of photographs, radio, television, and film content, so these challenges are not new, per se. Of greater interest here are aspects of blogs that enable communication phenomena not found in traditional media.

One aspect of this difference from traditional media is the option for bloggers to allow readers to *comment* on their blog entries, which can give rise to communicative exchanges between bloggers and commenters within a single blog, and which blur the boundary between static HTML web pages and interactive discussion forums (Herring, Scheidt, et al., 2004, 2005). Another is the option to incorporate *links* into blog sidebars and entries; this is part of the definition of a blog, according to Blood (2002). Links make intertextual connections among blogs and between blogs and other kinds of online media technologically explicit; linking to someone's blog can also function as a turn in a "conversation" between bloggers (Herring, Kouper, et al., 2005). Further, linking from text and images creates integrated, multimodal units in which the contributions of different modalities cannot easily be separated (cf. Weare & Lin, 2000). These features are not unique to blogs, but blogs were among the first types of web document to display them.

In early 2003, when blogs were attracting increasing media coverage in conjunction with the impending U.S. attack on Iraq, rigorous scholarship on blogs was virtually non-existent. At that time, the author, together with several others, formed a research group to study blogs.[7] The original goal of the group was to apply CA methods to randomly-selected blogs in order to characterize "typical" blogs, as opposed to the political blogs that were attracting most of the media attention, and thereby to shed light on the blog as an emergent internet genre. However, in order to gain a full picture of the blog genre, the researchers soon realized that it was necessary to extend traditional CA methods, in particular to include methods for the analysis of links and comments.

Traditional content analysis of blogs

Traditional content analysis methods are well suited for analyzing structural features of blog interfaces. Contemporary blogs typically have sidebars containing information about the author(s) and/or the blog, links to other blogs, and sometimes a calendar, photos, advertisements, and icons with links to organizations or products (such as blogging software) with which the blogger is associated. The frequency of these features, and how distinguishing each is of the blog genre, were analyzed by Herring, Scheidt, et al. (2004, 2005, 2006) using a classical CA approach. Similar general feature analyses have been conducted by Scheidt and Wright (2004), focusing on visual design elements in randomly-selected blogs, and by Trammell et al. (2006), focusing on blogs on a popular Polish blog-hosting site.

Traditional CA also works well for analyzing themes represented in blog entries and comments. In an analysis of posts in Bush and Kerry's blogs during the 2004 United States presidential campaign, Trammell (2006) coded items "for mention of the opponent; attacks, target of the attack (person or record), and the presence of emotional, logical, or source credibility appeals" (p. 403); she found that most posts contained attacks and that Kerry's blog

attacked Bush, the incumbent, more than the inverse. Thematic CA was also employed by Tremayne, Zheng, Lee, and Jeong (2006) in analyzing the political views (e.g., liberal or conservative) and content type (e.g., surveillance, opinion, reporting, personal) expressed in entries posted to "war blogs" during the recent U.S.-Iraq conflict.

However, there are limits to how revealing this approach can be. For example, while the *presence* of links and their manifestation on a blog page (e.g., as text or graphics) can be coded relatively straightforwardly as structural features,[8] this surface approach is unable to capture the *trajectories* of hyperlinks from blogs to other websites or even the nature of link destinations, which is important to understanding the function and meaning of links. Moreover, while the distribution of comments can be analyzed with interesting results,[9] the discrete coding and counting approach of traditional CA is ill suited for analyzing patterns of interaction via comments, which are inherently relational. Finally, traditional CA is unrevealing about the stylistic or linguistic strategies used to construct entries and comments. These limitations have led blog researchers who favor CA to expand the methodological paradigm in various ways.

Expanded methods of blog content analysis

A number of blog studies have combined thematic analysis of blog content with link analysis. Williams et al. (2006) coded the presence of thematic content such as war, economy, and health care in Bush and Kerry's campaign blog entries, but also analyzed the number of hyperlinks, the internal to external hyperlink ratio, and hyperlink destination, to determine if links "led users to media outlets, advocacy groups, within the candidate's own site, political party site, or other external websites" (p. 182). They found that candidate blogs were more likely to provide directives to external links than to direct viewers to content within the blogs, in contrast to the candidates' official websites. Tremayne et al.'s (2006) CA of war blogs also analyzed the distribution of incoming links in relation to other characteristics of blog content and included a social network analysis, which revealed that liberal and conservative war bloggers comprised distinct spheres of interaction with limited connections between them.

In the author's blog research group, the original impetus for moving beyond classical CA was a desire to address empirically a popular perception that the blogosphere, or universe of blogs, was actively "conversational." Sidebar links were analyzed as a manifestation of interaction between blogs: Links from randomly-selected blogs were followed from blog to blog to create a snowball sample that was then plotted as a social network diagram, from which topically-focused cliques emerged. Within these cliques, however, even reciprocally-linked bloggers blogging on the same topic only rarely left comments in each other's blogs or referred to each other in blog entries (Herring, Kouper, et al., 2005). Ali-Hasan and Adamic (2007) found a similar lack of correspondence between comments or citations in blog entries and contacts linked in the blogrolls of Kuwaiti bloggers. In contrast, Efimova and De Moor (2005) followed links in their study of an extended cross-blog conversation, which they found to be highly interactive, although the conversation itself was their pre-defined unit of analysis, rather than the individual blog.

In all of the above studies of cross-blog "conversation," link analysis was supplemented by analysis of interaction through blog entries and comments. Relatively few studies have as yet focused on interaction in entries and comments, however, in part due to the difficulty of capturing for analysis all parts of conversations that extend across multiple blogs. One computational solution was proposed by Nakajima et al. (2005), who automatically extracted cross-blog "threads," as defined by links in entries to other blogs, in an effort to identify bloggers who take on important conversational roles, such as "agitator" and "summarizer."[10]

Finally, several studies have focused on the language used in expressing blog content. Most of these make use of corpus linguistic methods involving automated counts of word frequencies. For example, Herring and Paolillo (2006) analyzed the frequency of grammatical function words (such as noun determiners and personal pronouns) hypothesized to correspond to male and female writing styles, finding that the two styles better predicted whether the blog was a filter or personal journal than the gender of its author. Huffaker and Calvert (2005) analyzed language style in teenage blogs, using the DICTION analysis software to classify lexical items in relation to assertiveness and emotion. Similarly, Balog, Mishe, and de Rijke (2006) analyzed the occurrence of emotion words (such as "excited," "worried," and "sad") in a corpus of blog entries over time, relating spikes in emotional language use to world events.

The last two sections have shown that narrow applications of CA can be revealing about certain types of blog content, but that a broader conception of CA is required in order to capture important features of blogs that the narrow approach does not, including patterns associated with linking, commenting, and language style. Moreover, since the broad conception encompasses the narrow conception (traditional CA can be included in the methodological repertoire of [[web content] analysis]), it is not necessary to adopt both approaches; the broad approach alone is sufficient.

Challenges in blog content analysis

Even when expanded analytical methods are available, challenges to analyzing blog content remain. Data sampling and defining units of analysis still pose challenges similar to those identified by Schneider and Foot (2004) and Weare and Lin (2000) for web analysis in general.

The full extent of the blogosphere is nearly as unmeasurable as that for the web as a whole, given the high rate of churn in blog creation and abandonment, the existence of private blogs, the presumed high number of blogs in other languages hosted by services that are not indexed by English-language search engines, and so forth; this makes random sampling of the blogosphere a practical impossibility. Studies that have aimed at broad representation have for the most part had to be satisfied with random sampling from a subset of blogs, accessed from blog hosting services or blog tracking services. Blogs have one advantage over traditional websites, however, in that many preserve archives of earlier content. Still, blog researchers are well advised to download and save versions of the pages they intend to analyze as data, as blogs can and do disappear (Wikipedia, 2008).

With regard to units of analysis, blogs provide a number of natural structural options: Units that have been analyzed in blog content studies include the individual blog, its front page (which presents the most recent entries or posts), the entry + comments, or either the entry or the comment alone. As with websites more generally, however, the interlinked nature of blogs poses problems for delimiting natural groupings of blogs, leading researchers such as Herring, Kouper, et al. (2005) to set arbitrary limits to their snowball samples in terms of degrees of separation and to allow clusters of blogs to emerge from patterns of reciprocal linking.

Finally, the identification and capture of cross-blog exchanges remains a persistent challenge for researchers interested in interactive content, in that "conversations" take place not only through links but also through citations in entries (with or without links), comments left on other blogs, and, in many cases, communication via other media, such as email or instant messaging. The bottom-up and top-down approaches to identifying blog conversations used by Herring, Kouper, et al. (2005) and Efimova and De Moor (2005), respectively, illustrate the types of methodological innovation that blog researchers have made in order to address certain questions about blog content. While traditional CA, CA-related paradigms, and earlier web content analyses all provide useful precedents, most blog researchers have found it necessary to innovate methodologically in some respects.

Toward an Expanded Paradigm

The previous sections have demonstrated the need for a broader construal of web content analysis, one that draws on methods from other disciplines to address characteristic features of the web such as hyperlinks and textual exchanges, and that recasts traditional CA notions such as comparable units of analysis, fixed coding schemes, and random sampling to fit the requirements of web research.

This broad construal assumes a more general definition of content than is typically found in traditional CA. In the narrowest sense, "content" in CA refers to the thematic meanings present in text or images and sometimes to the "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, features, links, and exchanges, all of which can communicate meaning. Along with this broader definition comes a broadening of the methodological paradigm; theme and feature analysis methods need to be supplemented with other techniques, in order to capture the contributions of different semiotic systems to the meaning of multimodal, multifunctional web content.

The solution proposed here is a methodologically plural paradigm under the general umbrella of Web Content Analysis (WebCA), which includes the methods discussed in this chapter, along with other techniques that can address the characteristics of web content (and internet content more generally) as it continues to evolve in new directions. One conceptualization of the proposed paradigm is represented schematically in Figure 2. Image analysis is included in Figure 2 as a separate component, because even though image content can be analyzed for its themes and features, the interpretation of visual content can benefit from methods drawn from iconography and semiotics, which are not included in any other component. The ellipses on

the right of the figure represent other components not discussed in this chapter, but that could potentially emerge as important in future research.

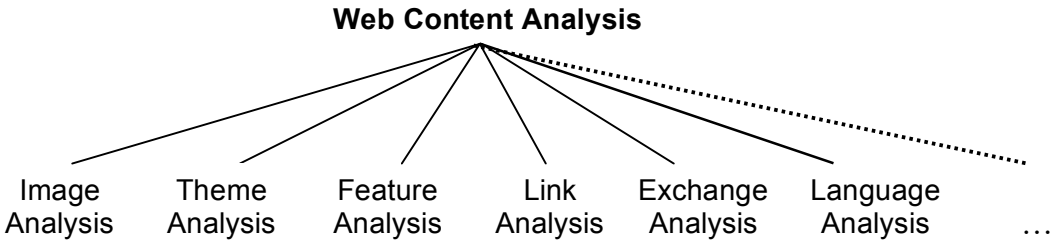

Figure 2. WebCA: An expanded paradigm

The coherence of this approach, and the reason for labeling it "content analysis," is that the methods are informed by the general principles of CA; that is, they must enable "objective, systematic, and quantitative description of the content of [web] communication" (Baran, 2002, p. 410). Thus "exchange analysis," for instance, is not simply a set of methods borrowed from discourse analysis; rather, discourse analysis insights about exchanges are operationalized and implemented as content analysis, following, in as much as possible, a general coding and counting procedure. While the particulars of each WebCA methodological component would need to be specified in future work, one proposed characteristic of the general approach is that classifying and counting phenomena of interest could either be done manually or automated. Although traditional CA has generally relied on manual coding, automated means of identifying phenomena of interest are proving increasingly useful in generating "objective, systematic, and quantitative descriptions" of web content (e.g., Balag et al., 2006; Nakajima et al., 2005).

At the highest level, the WebCA umbrella could serve to stimulate articulation of much-needed general recommendations regarding data collection and analysis based on the realities of the present-day web and the norms emerging from the growing body of web content research. Moreover, such a pluralistic paradigm could facilitate the generation of principled accounts of, and guidelines for, analyzing content in multiple modes with integrated function (such as links and images or text combined) and single-mode content with multiple functions (such as links that both function as conversational moves and define networks), in order to address the trend toward increasing media convergence on the web (cf. Weare & Lin, 2000).

Conclusion

This chapter has contrasted a narrow application of traditional content analysis methods to the web with an alternative conceptualization of what content analysis could (and, it has been argued, should) become in response to the challenges raised by new online media. As the review of weblog research illustrated, non-traditional content analyses can benefit scholarly understandings of the web and expand CA as a methodological paradigm. At the same time, any significant expansion of an established paradigm is likely to generate some resistance.

Some might object, for instance, that in opening up the paradigm as suggested above, methodological rigor and interpretability of research results could suffer. Analyses may not be

comparable across researchers; some may be ad hoc (cf. McMillan, 2000). It could be difficult to appreciate initially how an analysis involving methodological innovation is representative and reproducible—the criteria for "robust" analysis (Schneider & Foot, 2004). If researchers are permitted to innovate freely, web content analysis could be taken less seriously than other branches of social science.

In response to these concerns, it should be recalled that innovation is a vital process in the evolution of any research paradigm; without it, the paradigm would stagnate. Innovation is especially needed when new phenomena present themselves. This does not mean that web researchers should be allowed to have lax standards; they do, however, need to hold *themselves* to high standards of conceptual clarity, systematicity of sampling and data analysis, and awareness of limitations in interpreting their results, since they cannot depend entirely on traditional CA prescriptions to guide them.

In the meantime, research paradigms tend naturally to become more systematized and formalized over time, as best practices are distilled and refined. As more research on the communicative content of digital media (in its myriad forms) is carried out, the knowledge created will inform future analyses. Coding schemes designed and validated for web content will become available, facilitating comparison of findings across studies. 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. More complete indexes and archives of web content will also become available, and better search tools will be developed (e.g., for blog content), facilitating sampling.

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

Notes

[1] While McMillan (2000) acknowledges that "the *size* of the sample depends on factors such as the goals of the study" (p. 2, emphasis added), she does not mention that different research goals/questions might call for different *types* of samples. Rather, she asserts that random samples are required for "rigor" in all CA studies—a claim that many researchers would dispute (see, e.g., note 5).

[2] For descriptions of these and other classic interrater reliability measures, see Scott (1955), Holsti (1969), and Krippendorf (1980, in press).

[3] In a review of 25 years of content analyses, Riffe and Freitag (1997; cited in Weare & Lin, 2000) found that most studies were based on convenience or purposive samples; only 22.2% of the studies attempted to be representative of the population of interest.

[4] On grounded theory, see Glaser and Strauss (1967).

[5] Herring (2004, p. 350) notes that "in CMDA, [sampling] is rarely done randomly, since random sampling sacrifices context, and context is important in interpreting discourse analysis results."

[6] This estimate is based on a report that the number of blogs created at major hosts was 134-144 million in October 2005 (http://www.blogherald.com/2005/10/10/the-blog-herald-blog-count-october-2005/, accessed December 7, 2007). Blog creation, especially in countries outside the U.S., has increased since then, although many blogs have also been abandoned (Wikipedia, June 28, 2008).

[7] The (We)blog Research on Genre (BROG) project. See http://en.wikipedia.org/wiki/BROG, accessed June 28, 2008.

[8] For example, Herring, Scheidt, et al. (2004, 2005) found that contrary to popular claims that blog entries typically contain links and link often to other blogs, the average number of links in entries in randomly-selected blogs was .65, and most entries contained 0 links. Moreover, the majority of links were to websites created by others, with links to other blogs coming in a distant third.

[9] See, e.g., Herring, Scheidt, et al. (2004, 2005); Mishne and Glance (2006).

[10] This study is an exception to the generalization that most computational web studies do not orient toward content analysis. The stated goal of Nakajima et al. (2005, p. 1) is to capture and analyze "conversational web content" in blogs.

References

Ali-Hasan, N., & Adamic, L. (2007, March). *Expressing social relationships on the blog through links and comments*. Paper presented at the International Conference for Weblogs and Social Media, Boulder, CO.

Balog, K., Mishne, G., & Rijke, M. (2006, April). *Why are they excited? Identifying and explaining spikes in blog mood levels*. Paper presented at the 11th Meeting of the European Chapter of the Association for Computational Linguistics, Trento, Italy.

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.

Berelson, B. (1952). *Content analysis in communication research*. New York: Free Press.

Berelson, B., & Lazarsfeld, P. F. (1948). *The analysis of communication content*. Chicago/New York: University of Chicago and Columbia University.

Blood, R. (2002). Introduction. In J. Rodzvilla (Ed.), *We've got blog: How weblogs are changing our culture* (pp. ix-xiii). Cambridge, MA: Perseus.

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

Dimitrova, D. V., & Neznanski, M. (2006). Online journalism and the war in cyberspace: A comparison between U.S. and international newspapers. *Journal of Computer-Mediated Communication, 12*(1), article 13. http://jcmc.indiana.edu/vol12/issue1/dimitrova.html

Efimova, L., & de Moor, A. (2005). Beyond personal web publishing: An exploratory study of conversational blogging practices. *Proceedings of the Thirty-Eighth Hawaii International Conference on System Sciences*. Los Alamitos, CA: IEEE.

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

Foot, K. A., Schneider, S. M., Dougherty, M., Xenos, M., & Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 U.S. electoral Web sphere. *Journal of Computer-Mediated Communication, 8*(4). http://jcmc.indiana.edu/vol8/issue4/foot.html

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

Glaser, B., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago: Aldine.

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* (pp. 338-376). New York: Cambridge University Press.

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*. Los Alamitos: IEEE.

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics, 10*(4), 439-459.

Herring, S. C., Scheidt, L. A., Bonus, S., & Wright, E. (2004). Bridging the gap: A genre analysis of weblogs. *Proceedings of the Thirty-Seventh Hawai'i International Conference on System Sciences*. Los Alamitos, CA: IEEE.

Herring, S. C., Scheidt, L. A., Bonus, S., & Wright, E. (2005). Weblogs as a bridging genre. *Information, Technology & People, 18*(2), 142-171.

Herring, S. C., Scheidt, L. A., Kouper, I., & Wright, E. (2006). Longitudinal content analysis of weblogs: 2003-2004. In M. Tremayne (Ed.), *Blogging, citizenship, and the future of media* (pp. 3-20). London: Routledge.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison Wesley.

Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity and language use in teenage blogs. *Journal of Computer-Mediated Communication, 10*(2). http://jcmc.indiana.edu/vol10/issue2/huffaker.html

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

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Newbury Park: Sage.

Krippendorf, K. (In press). Testing the reliability of content analysis data: What is involved and why. In K. Krippendorff & M. A. Bock (Eds.), *The content analysis reader*. http://www.asc.upenn.edu/usr/krippendorff/dogs.html

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*. Los Alamitos, CA: IEEE.

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.

Mishne, G., & Glance, N. (2006). *Leave a reply: An analysis of weblog comments*. Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem, 15th World Wide Web Conference, Edinburgh.

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.

Nakajima, S., Tatemura, J, Hino, Y., Hara, Y. & Tanaka, K. (2005, May). *Discovering important bloggers based on analyzing blog threads.* Paper presented at WWW2005, Chiba, Japan.

Park, H. W. (2003). What is hyperlink network analysis?: New method for the study of social structure on the Web. *Connections, 25*(1), 49-61.

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.

Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *17*, 321-325.

Thelwall, M. (2002). The top 100 linked pages on UK university web sites: High inlink counts are not usually directly associated with quality scholarly content. *Journal of Information Science, 28*(6), 485-493.

Trammell, K. D. (2006). Blog offensive: An exploratory analysis of attacks published on campaign blog posts from a political public relations perspective. *Public Relations Review, 32*(4), 402-406.

Trammell, K. D., Tarkowski, A., Hofmokl, J., & Sapp, A. M. (2006). Rzeczpospolita blogów [Republic of Blog]: Examining Polish bloggers through content analysis. *Journal of Computer-Mediated Communication, 11*(3), article 2. http://jcmc.indiana.edu/vol11/issue3/trammell.html

Tremayne, M., Zheng, N., Lee, J. K., & Jeong, J. (2006). Issue publics on the web: Applying network theory to the war blogosphere. *Journal of Computer-Mediated Communication*, *12*(1), article 15. http://jcmc.indiana.edu/vol12/issue1/tremayne.html

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.

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

Wikipedia. (2008, June 28). *Blog*. http://en.wikipedia.org/wiki/Blog

Williams, P., Tramell, K., Postelnicu, M., Landreville, K., & Martin, J. (2005). Blogging and hyperlinking: Use of the web to enhance visibility during the 2004 U.S. campaign. *Journalism Studies*, *6*(2), 177-186.