Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 8, Number 3, May, 2002 ISSN=1531-7714

Analyzing Online Discussions: Ethics, Data, and Interpretation

Sarah K. Brem
Division of Psychology in Education
Arizona State University

Online discussions are attractive sources of information for many reasons. Discussion forums frequently offer automated tracking services, such as a transcript or an archive, so that you can engage in animated conversation and analyze it at leisure, or locate conversations that took place months or years ago. Online tools provide an opportunity to observe a group without introducing your own agenda, to follow the development of an issue, or to review a public exchange that took place in the past, or outside the influence of researchers and policymakers. You can test additions and revisions to tools for communication, building more effective online classrooms, research groups, and professional organizations. Whether you are looking for ways to improve interactions within a working group (Ahuja & Carley, 1998), studying the interactions of a community that interests you (Klinger, 2000), or assessing student learning (Brem, Russell, Weems, 2001), online discussions can be a valuable tool.

An online discussion is identified by the use of a computer-mediated conversational environment. It may be synchronous, such as real-time chat, or instant messaging, or asynchronous, such as a listserver, or bulletin board. It may be text-only, or provide facilities for displaying images, animations, hyperlinks, and other multimedia. It may require a Web browser, a Unix connection, or special software that supports such features as instant messaging. Tools for online conversation are becoming increasingly sophisticated, popular, and available, and this increases the appeal of using online discourse as a source of data.

Online discussions present new opportunities to teachers, policymakers, and researchers, but they also present new concerns and considerations. This article is about access to, and management and interpretation of, online data. Online research is similar, but not identical to, face-to-face (f2f) research. There are new ethical considerations that arise when it is not clear whether the participants in a conversation know they are being monitored, or when that monitoring is so unobtrusive that it can easily be forgotten. Instead of collecting data using audio and video recording as in f2f conversations, preserving online conversations requires ways to download or track the electronic files in which the information is stored. Finally, in f2f interactions we examine body language and intonation as well as the words spoken, and in an online interaction, we have to look beyond the words written to the electronic equivalents of gestures and social conventions. This article will address these issues of ethics, data collection, and data interpretation.

This article is *not* about recommending any particular method of analysis. Whether you use grounded theory, quantifying techniques, experimental manipulations, ethnography, or any other method, you will have to deal with issues of collecting and managing data, as well as the structure of online communication. (For information about analyzing discourse, see Stemler, 2001; techniques and considerations that are specific to online discourse can be found in the 1997 special issue of *The Journal of Computer Mediated Communication*, "Studying the Net."). Information about tools for theory-based data manipulation is available at http://kerlins.net/bobbi/research/qualresearch/researchware.html and http://directory.google.com/Top/Science/Social_Sciences/Methodology/Qualitative/Tools/.

Ethical Considerations

Before we consider how to analyze an online conversation, we need to first consider what precautions should be taken to protect participants in the conversation. Because online conversation is relatively new and unfamiliar, and takes place at a distance, it is relatively easy to overlook possible ethical violations. People may not realize that their conversations could be made public, may not realize that they are being monitored, or may forget that they are being monitored because the observer's presence is virtual and unobtrusive. Some participants may feel relatively invulnerable because of the distance and relative anonymity of online exchanges, and may use these protections to harass other participants. Online exchanges of information require the same levels of protection as f2f exchanges, but it can be more complicated to achieve this.

If you belong to a university or similar institution, you will need the approval of an Institutional Review Board, created for the protection of human beings who participate in studies. Teacher-researchers and others who do not have an IRB

and are not associated with any such institution should nevertheless follow the ethical principles and guidelines laid out in *The Belmont Report*, available at http://ohsr.od.nih.gov/mpa/belmont.php3. Other useful resources include Sales and Folkman (2000), and NIH ethics resources at http://www.nih.gov/sigs/bioethics/researchethics.html.

The least problematic conversations are those that take place entirely in the public domain; people know they are publishing to a public area with unrestricted viewing, as if they were writing a letter to the editor. Newsgroups are an example of such exchanges—anyone with access to http://groups.google.com/ can access any conversation in the past twenty years. In many cases, this sort of research is considered "exempt" under Federal guidelines for the protection of human subjects; for researchers at institutions with an IRB, the board must confirm this status. Still, even public areas may contain sensitive information that the user inadvertently provided; novices are especially prone to accidentally giving out personal information, or including personal information without considering possible misuse. In addition to the usual procedures for anonymizing data (e.g., removing names, addresses, etc.), there are some additional concerns to address. Every post must be scoured for both intentional and unintentional indicators of identity. Here are some common ways that anonymity is compromised:

- Usernames like "tiger1000" do not provide anonymity; people who are active online are as well known by their usernames as their traditional names. Usernames must be replaced with identifiers that provide no link to the actual participant.
- You must also be vigilant in removing a participant's .sig (the signature file that is appended to a post) and any other quotes, graphics, and other idiosyncratic inclusions that are readily identifiable as belonging to a particular individual.
- Identifying information is often embedded in a post through quoting; for example, if I were quoted by another participant, my email address might be embedded in the middle of his or her message as "tiger1000 (sarah.brem@asu.edu) posted on 1 February 2002, 11:15."

If a domain establishes any degree of privacy through membership, registration, passwords, etc., or if you wish to contact participants directly, then the communications should be considered privileged to some degree. In addition to the safeguards required for public domain data, using these conversations in research requires at very least the informed consent of all participants whose work will be included in the analysis, with explicit description of how confidentiality and/or anonymity will be ensured. The procedures for informed consent, recruitment, and data collection will require "expedited" or "full" review by an Institutional Review Board. Once approval has been given, consent forms will have to be distributed to every participant, and only the contributions of consenting members can be stored and analyzed.

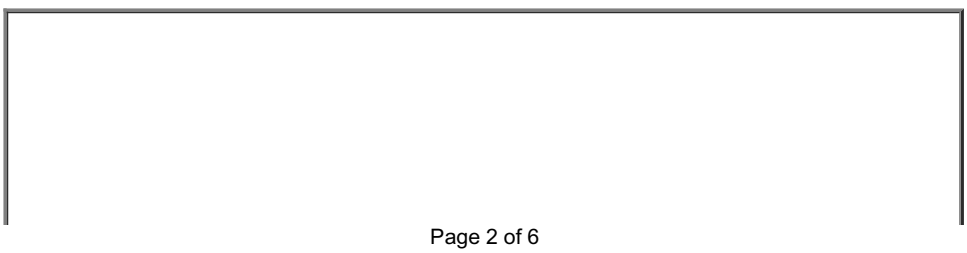
If you set up a site for collecting data, regardless of how much privacy and anonymity you promise, you are ethically bound to inform all potential participants that their contributions will be used as data in research. One example of how to provide this information has been implemented by the Public Knowledge Project. To see how they obtained consent, visit http://www.pkp.ubc.ca/bctf/terms.html. Likewise, if you contact participants directly, you need to make their rights clear and obtain their permission to use the information they provide for research purposes before engaging in any conversation with them.

In addition to preserving the safety and comfort of participants, you must also consider their intellectual property rights. All postings are automatically copyrighted under U.S. and international laws. Extended quotes could violate copyright laws, so quoting should be limited, or permission should be obtained from the author prior to publication. For more about U.S. and international laws, visit http://www.law.cornell.edu/topics/copyright.html.

Data Collection and Management

Once you have received the necessary permissions and taken the necessary precautions, the next concern is the best way to collect and organize the data for analysis. An online exchange often evolves over days or months, and may require handling tens of thousands of lines of text, along with graphics, hyperlinks, video, and other multimedia. Consider what media will be present before choosing tools for management and manipulation.

For text-only exchanges, a flatfile spreadsheet is often sufficient. The text can be downloaded as plaintext, or cut and pasted in sections. Paragraphs or lines of text become entries in the spreadsheet, and can be parsed into smaller units if desired. Once the data is placed in a spreadsheet, additional rows and columns can be used to hold codes and comments, and the spreadsheet can be sorted on these to reveal and examine patterns. An example of how this can be done is presented in Figure 1.



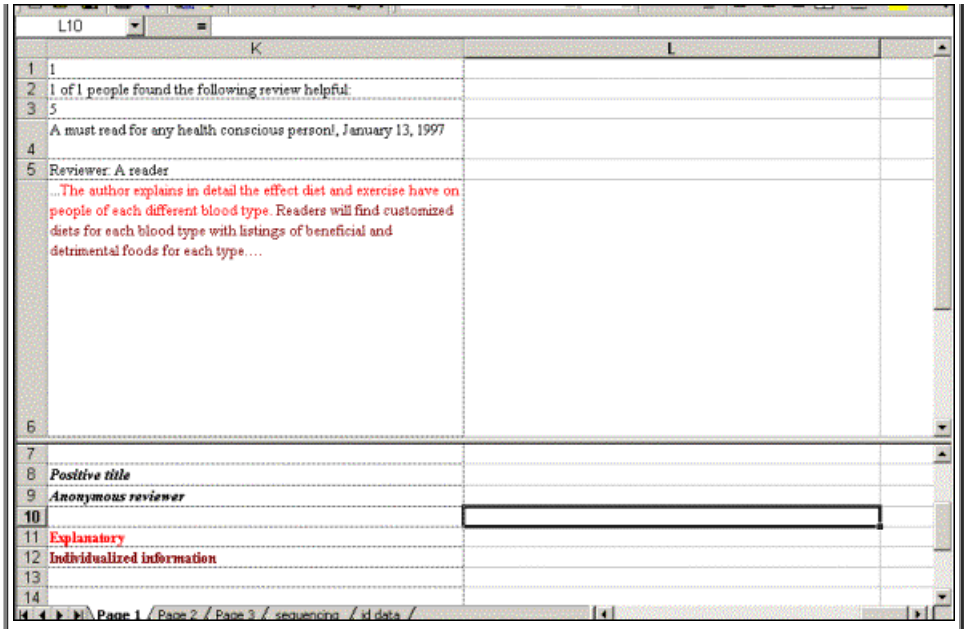


Figure 1. This sample is taken from an analysis of an online discussion of a book. At the top of the screen shot is a participant's entry, along with the information regarding its position in the thread, date of posting, and so on. Codes were added below each entry, shown at the bottom of the screen shot. The color of the codes corresponds to the color of the relevant text. Each participant's contribution can be added as an additional column.

There are many cases, however, when this technique will be ineffective. Because they can last for years, online conversations differ from f2f conversations in that they can be extremely long, often exceeding spreadsheet limits. Furthermore, they often contain hyperlinks, graphics, video, and other multimedia; these are often essential to the conversation, and most spreadsheets will not display them. When it is desirable to maintain these elements, there are two straightforward ways to do this. The first is to simply download all the relevant files and create a mirror archive on your own hard drive. This assures you constant, reliable access to the data, but may take up large amounts of space, and not all files can be downloaded (e.g. there may be security restrictions, additional software requirements, or intellectual property considerations). An alternative approach is to create a flatfile spreadsheet that contains hyperlinks to the original exchanges rather than the exchanges themselves. The disadvantage is that you cannot be sure the original files will always be available, but the spreadsheets containing these pointers take up very little space, are less problematic legally and technologically, and provide the full functionality of a spreadsheet (e.g., sorting and manipulation).

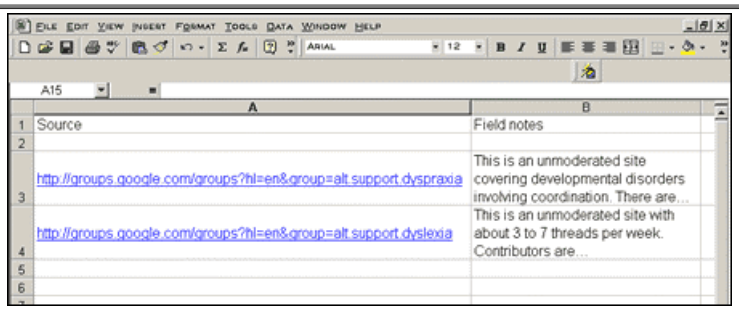


Figure 2. This example includes links to discussions on developmental disabilities that affect schoolchildren. The links take you to the conversation described in the field notes.

The advantage to using a flatfile database is that it allows for flexible coding. The disadvantage is that it does not support any particular theoretical perspective. For this reason, you may want to begin by using a flatfile, then transfer data to a theory-based format after you have done some initial processing and can narrow down what you want to focus on. Such tools are described at the sites mentioned above.

Data Preparation, Manipulation, and Preservation

Online data creates extremely large files, not only because of their potential length but also because online conversations tend to be highly repetitive. Replies often contain portions of previous messages, if not the complete original; even if each individual contribution is relatively short, quoting can quickly create messages containing hundreds of lines. In addition, multimedia elements tend to take up considerable space. It is not unusual for a datafile to grow to 30 megabytes or more. Files of this size are very difficult to manipulate and can be prohibitively slow to load and

save. Therefore, it may become necessary to decide what information should be kept verbatim, what should be deleted altogether, and what can be replaced with a smaller reference code (e.g., if many participants quote message 112, you might replace each reposting of this message with the code "msg112"; advertisements might be indicated by the code "banner ad" or a hyperlink to the ad on the original site). These methods of abridging the record can be implemented before engaging in extensive analysis, so that the file that you work with most often is the smallest one.

In deciding on these matters, you should be guided by your research questions and you should preserve all information that is relevant to your questions; thus, advertising may be a central issue, or it may play a relatively small role. In any case, it is best to err on the side of preserving too much information. Once removed, a hyperlink, graphic, or reposted message can be difficult to recover. Start by keeping as much information as possible, and pare it down as you find elements that seriously interfere with speed, or that are adding nothing to your analysis. You may want to keep multiple versions, using the most streamlined for your analysis, and archiving richer versions in case they are needed later on.

Coding, Analysis, and Interpretation

The structure of an online exchange can be difficult to reconstruct, and its boundaries can be difficult to locate. Capturing the perspective of participants, challenging in any context or medium, is further complicated by new ambiguities created by the way in which conversations are created, stored, and accessed. While it may not be possible to resolve all inconsistencies and ambiguities, being aware of them and their implications for any particular interpretation is essential.

Reconstructing the Conversation

One significant difference between online and f2f conversations is that participants often view online conversations differently. Online discussions do not necessarily develop sequentially, nor can we be sure that all participants are seeing the same exchange. We can see this by comparing how listservs and bulletin boards are visited and revisited. A listserv sends messages to the subscriber's email account. Listservs send all messages in chronological order, regardless of the conversational thread to which they belong, so multiple conversations are interleaved. It is easy to miss a post, and each person may read a different set of messages. If you join a listserv after a conversation has begun, you will not see the beginning of the exchange. In contrast, bulletin boards keep message separate by thread, and all messages are available for the life of the bulletin board, or until they are archived.

A participant may follow conversations thread by thread, read everything written by a single author, skip from one topic to the next, or otherwise deviate from the original presentation. You should consider reviewing the conversation in a variety of ways in order to understand better how participants receive and work with the information.

For example, Usenet groups often attract users who only wish to ask a single question, get an answer, and never return. In addition, while some servers provide a complete, searchable Usenet archive (http://groups.google.com/), others regularly delete files to save space, or may not provide much in the way of searchability. For these reasons, it is common for several participants to ask the same question, sometimes word for word, over and over. Understanding why this happens and how the conversation develops requires looking at the records both as if you are a user with access to the full record and, as if you are a user with access to a very limited record. It is virtually impossible to capture all possible viewings, but you will probably want to capture several.

Tracking a conversation, regardless of the perspective you choose, can be challenging, rather like assembling a rough-cut jigsaw puzzle. The threads of conversation are easily broken; if a participant or server changes a subject line, archiving tools cannot follow the conversation and the line of thought becomes disconnected. People use multiple accounts and identities, either because they are deliberately trying to hide their identity, or for innocent reasons, such as logging in differently from work and home. There are, however, ways to reconstruct a conversation. To track a thread, examine subject lines to see if they correspond except for a reply indicator, look at dates of posting, or examine the text for quotes from previous messages in the thread or other references to previous postings in the thread. In the case of users, even if participants' usernames change, they may be identifiable through their email addresses, their signatures, hyperlinks to their home pages, or their writing styles and preferred themes. For example, in analyzing one Usenet group in which the topic of speed reading frequently arose, I noted that there were several usernames employed by one company; these users would respond as if they were "ordinary" individuals, rather than identifying themselves as company representatives. However, all used the same prefabricated plug for the company's product. Thus, I could use this to mark the posts as coming from related users, or perhaps the same user.

Where, What, and Who is the Conversation?

In addition, consider the context. F2f conversations consist of a relatively well-bounded exchange; the participants, location, and duration are easier to determine than they are in online discourse. Online, participants can possess multiple identities, steal another's identity, or carry on a conversation with themselves. The conversation not only crosses geographical boundaries, but may send participants to archives of prior exchanges, websites, FAQs, and other resources. As a result, the conversation may not be neatly contained by a single listserv, chat room, or other discourse

environment. Even within a single environment, the conversation you are interested in may be no more than a few lines or posts tucked in among other threads, spam (mass mailings), flames (inflammatory posts), and announcements. Finally, regarding duration, online conversations may last minutes or years, and may proceed at the rate of several exchanges per minute or one exchange every few weeks or months.

Given these complexities, the best approach is to be aware that you will have to draw somewhat arbitrary boundaries around a group of participants and exchanges, letting your choice be led by your questions. If identifying participants is crucial (perhaps you suspect that warring factions are trying to discredit one another by posing as members of the other camp), then you will have to look for clues that reveal identity and consider how your interpretations are affected by the possibility of imposters. If the conversation takes place amongst a small, tightly knit group with a strong foundation of common knowledge, then shared spaces like FAQs and group websites becomes crucial, and should be included. If there have been significant changes in the political or educational climate during the course of the conversation, duration will become important, and the timeline of the exchange may need careful examination.

You will always have to draw boundaries, and there will never be one right set of boundaries to draw. The important thing is to draw them in such a way that you can explain your reasoning to others, and in a way that allows you to get rich, useful, and dependable answers to the questions that interest you.

Knowing How to Talk Online

We do not analyze f2f conversations without having some experience with f2f conversation, both at an everyday level and at the more finely honed level of a discourse expert. You should also become a participant in online communities before trying to research them, gaining both everyday and scholarly familiarity. Rather than just knowing the basics of navigation and communication, it is important to be fluent in everyday conventions and the online analogs of body language and nonverbal f2f communication. These include "emoticons" (e.g., symbols for smiling `8^)`, disgust `8-P`, and so on), as well as conventions such as SCREAMING BY TYPING IN ALL CAPS, or including actions, such as **::**hugs newbie**::** or <<grins at newbie>>. (For more information about communication conventions on the Internet, visit http://www.udel.edu/interlit/chapter5.html.)

In addition, learn how to relate to participants as individuals; it is easy to fall into the trap of treating them as disembodied voices, or automatons, rather than as complete people. What are their interests online and off? Is their style of conversation friendly, combative, joking, pedantic? What topics will get an emotional reaction from them? What sorts of conversational moves will get a reaction from them (e.g., some people are offended by posts IN ALL CAPS, and will tell the poster to stop shouting)? In an extended conversation with a group, you should get to a point that you can recognize participants without relying solely on usernames.

Final Thoughts

The study of online discourse is still quite new, and there is much about the treatment and analysis of these data that has not yet been addressed. When faced with a situation for which there is no standard procedure, the best course of action is to begin with established techniques and then adapt these to the online environment. Have a rationale for any adaptations or deviations you decide to make because these will help you to establish credibility with editors and peers and will allow others to adopt, recycle, and refine your approach.

Notes:

This work was supported by an NSF Early Career Award (REC-0133446). Investigating Critical Thinking In Multimedia Environments To Improve Public Utilization Of Science.

References

Ahuja, M. K. & Carley, K. M. (1998). Network structure in virtual organizations. *Journal of Computer-Mediated Communication,* 3 (4). Available online: http://www.ascusc.org/jcmc/vol3/issue4/ahuja.html.

Brem, S. K., Russell, J. & Weems, L. (2001). Science on the Web: Student evaluations of scientific arguments. *Discourse Processes,* 32, 191-213. Available online: http://www.public.asu.edu/~sbrem/docs/BremRussellWeems.webversion.htm.

Klinger, S. (2000). "Are they talking yet?": Online discourse as political action. *Paper presented at the Participatory Design Conference*, CUNY, New York. Available online: http://www.pkp.ubc.ca/publications/talkingyet.doc.

Sales, B. D. & Folkman, S. (2000). *Ethics in research with human participants*. Washington, DC: American Psychological Association

Stemler, Steve (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, 7(17). Available online: http://pareonline.net/getvn.asp?v=7&n=17.

Descriptors: Content Analysis; Research Methods; World Wide Web

Citation: Brem, Sarah (2002). Analyzing online discussions: ethics, data, and interpretation. *Practical Assessment, Research & Evaluation*, 8(3). Available online: http://PAREonline.net/getvn.asp?v=8&n=3.