Problems in the Use-Centered Development of a Taxonomy of Web Genres *

Kevin Crowston, Barbara Kwaśnik and Joseph Rubleske
Syracuse University School of Information Studies
bkwasnik@syr.edu, crowston@syr.edu, jrublesk@gmail.com

October 24, 2008

1 Introduction

Much of the recent research on Web genres has been driven by the assumption that information seeking on the Web can be made easier and more effective by information retrieval (IR) tools that differentiate among indexed Web documents in terms of their genre (or genres). Currently, Web search engines such as Google.com yield results according (in large part) to the frequency and location of keywords in Web documents indexed by the engine (additional information can be used to then rank these results). To the extent that keywords signify the subjects of Web documents, these results can be described as topical. Unfortunately, topical searches are not always sufficient to solve information needs, as task-driven searchers often must distinguish between documents that share a topic but assume a different form, purpose and/or function. For example, before purchasing a digital camera, an individual may want to read reviews from online magazines and see the blogs in which people who have used this camera express their opinions and personal stories. Using a query term such as "Canon Powershot G6" yields the bulk of the results referring to digital camera sellers, not magazines, discussion forums or blogs. A renewed search with a more refined query term such as "Canon Powershot G6 review" might prove effective, but the query term "Canon Powershot G6 opinion" might yield mixed results. Efforts to locate a current, trustworthy and pertinent discussion forum might require considerable manual search.

One way to improve the precision of a search is to utilize additional meta-data, such as genre, to distinguish relevant and irrelevant documents. Document genre can be defined—at least for traditional hard-copy documents—as "essentially a document type based on purpose, form and content" (Rosso, 2008, p. 2). For example, a *flier* has a certain form—a single page, usually smaller than North American letter size—and a certain purpose—to advertise an event, activity or service. Its form and purpose constrain the type and amount of content that can be printed on the *flier*. For example, a *flier*'s size (it is often placed on windshields or stapled to a utility pole) prevents a restaurateur from including a full menu

*This research was partially supported by NSF IIS Grant 04-14482

as part of a *flier* advertising a restaurant. Conversely, because a *flier*'s recipients expect *fliers* to serve a marketing function, many documents that assume the form of a *flier* but do not in any way advertise or promote may not be construed as *fliers*.

Genre is useful in information tasks because it makes documents more easily recognizable and understandable to recipients, thus reducing the cognitive load of processing them (Bartlett, 1967). As well, knowledge of the genre can be exploited in a number of tasks because genre provides some fixity to otherwise infinitely variable texts (Yates and Sumner, 1997). Genre acts as a template of attributes that are regular and can be systematically identified. Therefore, if a Web search could use genre meta-data, it might be possible to specify the desired information more precisely (e.g., finding a document whose purpose matches the user's). Following this idea, researchers from the fields of information science, communications and linguistics have tried during the past decade to demonstrate the efficacy and viability of tools that group Web documents—as search results or contained in hierarchical directories—in terms of Web genre (see, e.g., Dewe et al., 1998; Crowston and Williams, 2000; Haas and Grams, 2000; Nilan et al., 2001; zu Eissen and Stein, 2004; Freund et al., 2006).

Central to these efforts is the development of some sort of classification of genres, e.g., taxonomies that organize a controlled vocabulary for Web-genre terms. Unfortunately, our review of the literature reveals a lack of consensus about the Web genre taxonomy on which to base such systems. Furthermore, our review of reported efforts to develop such taxonomies suggests that consensus is unlikely. Rather, we argue that these issues actually resist resolution because the acceptance of potential answers depends on a researcher's epistemological and ontological orientation. For example, a researcher may view a Web genre as an innate property of a Web document, with coding activity assumed to be a process of genre identification. On the other hand, a researcher deriving warrant for terms from the perspective of professional practice may assume that a genre term and set of genre attributes can be assigned by experts based on consensus from long-standing tradition. Finally, a researcher may believe that a Web document could conceivably instantiate a number of Web genres, depending on attributes specific to the document coder and his or her context. It is not impossible to imagine a situation in which all three orientations coincide: some terms come from literary warrant, some from expert warrant, and some bubble up from the particular and unique context in which they are used. More often though, these various approaches will suggest different ways to describe the same documents. Moreover, coders and taxonomy developers may view their efforts through different conceptual lenses, and as a result arrive at positions that are tricky to reconcile.

The purpose of this chapter is to support this claim by first briefly reviewing prior work on developing taxonomies of Web genres and second describe the problems we encountered in a study aimed at developing a genre taxonomy from a user study.

2 Why is it hard to develop a Web genre taxonomy?

As noted above, document genre can be defined as "a document type based on purpose, form and content" (Rosso, 2008, p. 2). A fundamental question that must be addressed to develop a satisfactory taxonomy concerns the origin of genre terms in the taxonomy. Simply

put, where should genre terms come from? (A second question to address is the organization of terms, an issue we addressed in prior work (Kwaśnik and Crowston, 2004).) For example, in many taxonomy development efforts, an individual coder assigns each Web document contained in a corpus to one or more genres. Multiple individuals may perform this task with the eventual aim of measuring inter-coder reliability, or a small group of coders perform it simultaneously. We note two problems that arise in generating such terms: the difficulty of defining genres precisely and the difficulties in generating a collection of genre terms that cover a collection of documents.

2.1 Difficulties in defining genres

A first challenge in studying genre is that there never has been, nor is there presently, a consensus on what a genre is, what qualifies for genre status, how genres "work," how we work with genres, how genres work with each other, or how best to identify, construe, or study genres. Genres are a way people refer to communicative acts that is understood by them, more or less, but which is often difficult to describe in its particulars. Thus, genres are recognized and used, but not so readily described and defined.

The confusion and lack of a one-size-fits-all solution when it comes to Web-genre taxonomies is, in our opinion, a result of the fact that genres are frequently not construed the same way across all communities of users. In addition, even if they are more or less "universally" understood (such as a home page), there is still some debate about boundaries, granularity, and definition. In other words, genres may not be as generic as we would like in terms of implementing them in applications. This is not surprising, since the very essence of what makes a genre powerful is its intimate connection to the circumstances in which it is enacted. A genre only exists in use.

Given a definition of document genre as including both socially recognized form and purpose, in studying document genres it is necessary to look at the context of use as well as the formal technical details of the documents. Nevertheless, it is possible to make a logical division between intrinsic genre attributes and the extrinsic function that genre fulfills in human activities. The relative emphasis on form versus function also depends on the domain from which the genre emerges. For instance, musical genres emphasize the form and structure first (as in a sonata) while the way the sonata functions aesthetically "in the world" is generally given secondary attention in a discussion of the sonata genre. This does not mean that the two aspects are mutually exclusive, but that one aspect can take precedence. In studying emails in organizational settings, on the other hand, we might focus on how email messages function in the discourse between sender and receiver, and indeed emerge from it, thereby giving that aspect the greater emphasis.

The focus on the nature of the document genres themselves, or conversely the focus on what the document genres reveal about something else, such as human social activity, are two directions present in much of contemporary genre studies. Many researchers investigate both the intrinsic attributes and the situated implementation of genres, following from the assumption that these are inseparable. Other researchers assume that intrinsic attributes are relative, in the sense that genres are created from communicative action, and not the other way around. Genres, from this perspective, assume consensual attributes over time

as a result of well-trodden paths of use. Yet other researchers do not concern themselves with a comprehensive understanding of either intrinsic or extrinsic genre attributes but focus rather on any attributes that will allow them to exploit genre for knowledge-representation functions. We wanted to bridge all of these approaches: that is to identify the discernible attributes of genres as they are used in real situations so that we could incorporate this information as an enhanced representation of the document in an information-access system.

From studying non-digital genres we know that the role of content and form inform each other. For example, if we are presented with only the empty framework of the format of a letter (heading, salutation, body, and closing) most people can identify the genre. Similarly if we are presented with the content without the form—just the text—we can still recognize it as a letter (Toms et al., 1999). For some genres, the content is more important, but for some the form is equally so. In studying digital genres we rely not only on traditional indicators of a genre, such as specific content and form, but also new and different cues for both identifying and then analyzing and making sense of them. Above all, we recognize that any approach to attribute analysis must deal with the problem of a genre's intrinsic multifaceted nature, that is, the cues that not only identify the genre as an artefact, but also as a medium for participation in a communicative act (Kwaśnik and Crowston, 2004).

What has changed from formal Aristotelean models, though, is that today we recognize that an exhaustive identification of attributes, even if that were possible, may not be sufficient for a full understanding of a document's genre. This is because we have come to understand the power and primacy of the document's actual implementation in a life situation in addition to its content and technical attributes. In the realm of print documents, genres have evolved over the centuries, often slowly and gradually, occasionally suddenly, and while there may be lively discussion about when, say, a novella becomes a novel, genres in general have been relatively stable. A play remains an essentially recognizable genre despite genre-bending endeavors at various points in the history of drama. We can still easily identify the prototypical limerick, the tempo of a rousing march, or an office memo.

As documents have migrated to the Web though, their identity as genres has also evolved. New document genres have emerged (Crowston and Williams, 2000; Dillon and Gushrowski, 2000), while older ones have blended, changed, and been incorporated into different social endeavors. Print-document genres adapted to the Web, and new electronic genres emerging frequently, appear to be shuffled, disassembled and then put together again, in a seemingly chaotic manner. Many researchers, and indeed the public at large, assume that there are significant and fundamental differences in how these adapted and new genres will now function and be used. As with many new technologies, there are fond hopes that these genres will be socially transformative, enabling better communication, as well as more flexibility and expressiveness.

Emerging from these discussions is the broader question of whether technology leads human activities or follows it. In terms of genres of digital documents, the questions that arise are whether digital genres emerge from what people do on the Web, or whether the technology itself affords ways of doing things that people can then discover and exploit. This is by no means an easy question to answer, since people have always found ways to repurpose technologies, and digital technologies are no different. What is even more difficult

in the electronic environment is that many technologies are converging—voice, image, text, databases, computing-creating opportunities for combining and recombining genres of many different kinds in inventive ways and for unexpected purposes.

So, a discussion of genre is challenging for a number of reasons— among them the differences in the concept's role in various domains and the contextual nature of genre in action. Still, we find genre a useful concept because in identifying and labeling genres we try to capture the gestalt of the various components of the communicative act. This is all the more important for digital genres on the Web, since so many socially agreed-upon cues present in traditional print documents and oral communication are no longer available to us.

Despite these difficulties, we have a continuing and, indeed, growing need for understanding a document's genre. This is because genres provide an efficient way of dealing with documents at all stages of a document's lifecycle-from creation to dissemination to storage and retrieval and to utilization for new and creative purposes. In a vast landscape of communicative choices and strategies, genres provide a shortcut by which people can identify and participate in social endeavors. Knowing a document's genre, and therefore its communicative utility, helps a person formulate a precise query, for instance, or recognize the relevance of a document that is presented as the result of processing that query.

2.2 Difficulties in developing the scope and expressiveness of the taxonomy

Beyond the issues involved in defining the boundaries of a single genre are the problems involved in developing a collection of genres to comprise a genre taxonomy that is sufficient to describe a collection of documents. There are several benchmarks of a robust taxonomy: first and foremost is the attribute of reflecting the structure of the domain, but also very important is the ability of a taxonomy to be sufficeintly expressive. This means that the taxonomy comprises genres that are able to adequately represent the documents to which it will be applied. As Kwaśnik and Crowston (2004) have noted, there are two basic approaches to this task of genre term production: top-down and bottom-up.

Top-down Many attempts to develop a categorization of genres have been top-down, that is, they analyzed a set of documents based on theoretical principles or according to *a priori* classifications. In a top-down approach, the researcher draws from an existing set of genres and also from knowledge and understanding of Web genres of that domain. In one study, for example, each of two researchers "add[ed] new genres to the list" where "none of the already defined genres were appropriate... [The] two raters agreed completely on the coding for 68%" of the documents (Crowston and Williams, 2000, p. 205).

A key difference in these efforts is the number of genre categories distinguished. Many studies of Web pages have used fewer broader categories: for example, zu Eissen and Stein (2004) used only eight genres (*help*; *article*; *discussion*; *shop*; *portrayal, non-private*; *portrayal, private*; *link collection*; and *download*). At the other extreme, Görlach (2004) offered a catalog of some 2000 genre (or text type) terms intended to be an exhaustive list of the terms used in English. Somewhere in between, Lee (2004) categorized documents in the British National Corpus (BNC) into 70 genres or subgenres (with some document assigned more than one genre). He notes, however, that the genre terms used were "meant to provide starting points, not a definitive taxonomy", for example grouping *textbooks* and *journal*

articles as *academic texts* that can be further distinguished by medium.

In studies where taxonomy developers start with (but ultimately modify) a palette of Web genres proposed in a prior study, there is the question of which starter palette to use. At least two studies (Lim et al., 2005; Stubbe et al., 2007) made initial use of Dewe et al. (1998)'s genre taxonomy, for example, while Crowston and Williams (2000)'s taxonomy of 'document genres' was used by Roussinov and Chen (2001). This question is important methodologically because the use of any starter palette frames how Web documents in a corpus will be viewed. One may end up with a new taxonomy that does not much resemble the one she started with, but that was almost certainly influenced the form and shape of the taxonomy. In other words, a researcher might have have created a completely different taxonomy had she used a different starter palette or no starter palette at all.

Very few of these top-down studies include a discussion of the role that personal attributes (e.g., experience or expertise) play in this process, or precisely how multiple researchers reach agreement on Web genre terms. In another study, for instance, the authors tell us only that "page descriptions evolved through the course of the analysis into a system of page types" (Haas and Grams, 2000, p. 183).

Bottom-up In a bottom-up approach, Web users who have volunteered to participate in a study do the same thing—draw to the extent possible (and sometimes aided by tutorials) from their understanding of Web genres—to produce Web genre terms for the taxonomy. Such an approach seems desirable because it avoids imposing an *a priori* vocabulary with which users may lack familiarity. As Rosso (2008, p. 4) put it, "a good genre candidate for document descriptor should be recognizable to searchers". However, this approach relies on the ability of the users surveyed to adequately recognize and label documents by genre, which is problematic for the reasons surveyed above.

As zu Eissen and Stein (2004) note, "An inherent problem of Web genre classification is that even humans are not able to consistently specify the genre of a given page." Web documents are often ambiguous, and may not resemble the exemplar of a certain genre closely enough. Crowston and Williams (2000) point out that some Web documents did not have a "recognizable genre;" others seemed to instantiate an emerging genre that does not yet have a name. Indeed, the intended purpose of many Web documents is unclear, in part because of the "increasingly wide range of uses to which the Web can be put" (Haas and Grams, 2000). Alternately, multiple genre terms may seem appropriate to describe a particular document. Web documents may instantiate multiple genres (Crowston and Williams, 2000; Haas and Grams, 2000). As Santini (2008, p. 6) puts it, "genres are not mutually exclusive and different genres can be merged into a single document, generating hybrid forms." As well, more or less specific terms may be available. For example "...scholaly material can be seen as a super-genre that covers help, article and discussion pages" (zu Eissen and Stein, 2004). Which do we choose and how do we decide on the granularity? Finally, many lay users are unfamiliar with the formal genre concept and, as a result, some tend to conflate genre with topic, perceived document quality (e.g., "boring pages") or intended audience (e.g., "internal documents") (Dewe et al., 1998).

In the face of the difficulties noted above, researchers may intervene by explaining the genre concepts to participants (e.g., zu Eissen and Stein, 2004) and/or modifying the genre

terms supplied by participants (e.g., Dewe et al., 1998). As a result, most ostensibly bottom-up taxonomy development efforts may actually incorporate elements of both top-down and bottom-up approaches. In one such study, for instance, researchers "proposed ten genre classes" then asked interviewees to "specify up to three additional genre classes" (zu Eissen and Stein, 2004, p. 4).

Other recent attempts at developing a genre classification aim at discovering relevant attributes automatically, rather than identifying them first and then utilizing them in various tasks (Bagdanov and Worring, 2001; Karjalainen et al., 2000; Karlgren and Cutting, 1994; Kessler et al., 1997). This line of research assumes that genre attributes may be too unwieldy and slippery to identify "from the top," and that there may be too many genres in a rapidly growing and expanding field of digital documents and their implementations (Dillon and Gushrowski, 2000; Kennedy and Shepherd, 2005; Kwaśnik and Crowston, 2004; Watters and Shepherd, 1999).

3 A use-centered development of a taxonomy of Web genres

We turn now to describing our own efforts at building a taxonomy of genres based on a user study of Web document use. We first describe the research design and data elicitation and analysis methods we adopted before briefly discussing the results of our study. We then present the main challenges we faced in the study and its resulting limitations as the basis for a genre taxonomy.

3.1 Research design: Naturalistic field study

Our goal was to develop a better understanding of the use of genre in information-access tasks and then to develop a human-centered taxonomy of genres for use in subsequent phases of the overall research plan. Because genres are situated in a community's language and work processes, we felt it was important to learn about genres from people engaged in real tasks, and in their own words. While the concept of genre generates lively debate in terms of what it is and how best to study it, there is one aspect that is agreed upon by virtually every researcher, and this is genre's fundamentally social role—more specifically its communicative role. We can say that genre emerges from social activity (Miller and Friesen, 1984), and it, in turn shapes social activity by providing templates, frameworks, and socially agreed-upon constraints for communicating. How this knowledge of genre is then implemented or exploited is, in fact, as varied as the situations in which it is embedded.

We considered a top-down approach using a researcher-generated or standard list of genres as problematic for two reasons: First, genres are socially constructed, so different social groups using documents with similar structural features may think about them and describe them differently. A document may be unfamiliar and difficult to understand for someone outside of the community in which the genre is used, therefore, it is important to capture the users' own language and understanding of these genres. Second, it is imperative to extend any investigation to genres that are not necessarily vetted by traditional schemes, such as those that come out of domain-specific work (e.g., block-scheduled curriculum plans). As pointed out by Dillon and Gushrowski (2000, p. 202), genres are no longer necessarily "slow-forming, often emerging only over generations of production and consumption". Thus,

Respondents	No.	Typical Tasks	Typical Genres	Comments
Teachers	15	Preparing and revising lesson plans	Lesson plan Story page Resource page	Teachers from four public and private schools; most grades from K-12 are represented
Journalists	20	Developing a story or article: generating ideas; searching for other stories on the same topic; collecting new information; fact-checking	News story Directory Press release	18 print journalists, 2 television journalists
Engineers	20	Searches for tutorials, detailed information about products and tools, new or updated "knowledge" about a topic	Manual page Commercial page Product page	Includes 20 aeronautical and software engineers from one multinational firm

Table 1: Our Source of Genre Information: Three Groups of Respondents

we assumed that a traditional typology of genre or document forms would not be sufficient to describe the emerging and dynamic genres identifiable by users in general and our study community in particular.

3.2 Research informants

Knowing that we could not study the universe of Web genres or searchers, our first task was to identify respondents who would, in the course of their daily work, need to search on the Web, and who most likely would want to distinguish between one type of Web page and another. That is, we tried to identify people for whom genre information might be useful—indeed necessary—for determining whether a given Web page might be relevant to their needs.

Our study solicited information about genre from three groups of respondents: K-12 teachers, journalists and engineers as summarized in Table 1. We chose these three groups because the members of each share a discourse community in which a set of identifiable tasks and genres may play a role, and in which the identification of the genre of a document is likely to be important for their tasks yielding a wide range of tasks, genres and genre attributes.

3.3 Data elicitation

In general, our data-elicitation goal was to identify, for a collection of Web pages, the genre of the page, the clues each respondent used to recognize the genre, and the usefulness of the page for a task, all in the words of the respondents. We used think-aloud technique to understand the search goals and general strategy, but then followed it with a debriefing. During the interview, for every page visited we asked four questions:

1. What is your search goal?
2. What type of Web page would you call this?
3. What is it about the page that makes you call it that? (If they did not understand the question, we would ask, "Which features/clues on the page make you call it that?")
4. Was this page useful to you? How so (or why not)?

At the conclusion of the debriefing, and with permission from the respondent, we copied the URLs of the Web pages and the sequence in which it was visited into a database. This data was used to later re-create the search. From this re-creation, screenshots were taken of each Web page visited by the respondent, and a Web-based slide show (with accompanying URLs) of the entire sequence was created for each session. We are able to use this for coding and analysis, and intend to draw from these slide shows to develop a corpus of Web pages that a subsequent set of respondents can view and evaluate. We have nearly 1,000 screenshots of Web pages visited by respondents, each accompanied by its original URL and audio recordings of the sessions with transcripts, or detailed field notes for those interviews where recording was not permitted.

3.4 Data analysis

Content analysis was employed for identifying genre terms. We analyzed:

- Transcripts of audio files from the debriefing for the 32 respondents—19 journalists and 13 teachers (3 of the original transcripts were corrupted and could not be used).
- We also content analyzed the detailed field notes for 20 engineer respondents where audio recording had not been permitted.

First, we collected the terms used in answer to the question: "What type of Web page would you call this?" We transcribed the terms as given to us, without making a judgment about whether it was a legitimate "genre" or not. In other words, we allowed the respondent to identify the candidate genre terms for the analysis.

Before calculating the frequency, we made a few changes to some genre terms which we call "trimming." This included merging terms with inflectional differences or derivational forms of a word. For example, class note was merged to class notes, and governmental page with government page. As well, we considered both list of stories and list of articles as simply a list for frequency analysis.

Using the following rules, we further reduced the list of terms, bearing in mind that our goal was not so much to compile an exhaustive list, but rather to build a taxonomy to use in subsequent stages of the research. We eliminated:

	Engineers	Journalists	Teachers
Respondents	20	19	13
Genre term tokens	226 (11.3)	404 (21.26)	137 (10.53)
Genre term types	167 (8.35)	262 (13.78)	93 (7.15)

Table 2: Raw numbers and averages per respondent of candidate genre terms
The numbers in parentheses indicate average genre terms per respondent.

	Original Genre Terms		Trimmed Genre Terms		Selected Genre Terms	
	Token	Type	Token	Type	Token	Type
Engineers (20)	226	167	226	131	127	104
Journalists (19)	404	226	404	209	191	150
Teachers (15)	137	93	137	70	62	44
Total	767	522	767	410	380	298

Table 3: Results of trimming and selection.

- Terms that had only a personal meaning to the respondent, e.g, "good page."
- Terms that were so domain-specific that they would not be understood in any other context.

4 Results

We collected 226 genre terms from 20 engineers, 404 from 19 journalists, and 137 from 13 teachers for a total of 767 genre term tokens from the 52 subjects. The total of genre types (unique terms ignoring repetitions) was 522 (167 from engineers, 262 journalists, and 93 teachers). The count of genre terms is shown in Table 2. Table 3 shows the final number of genre terms following the trimming of variants and the elimination of terms we deemed not useful for the purposes of our study. Common genre terms across the populations studied are shown in Table 4, while Table 5 lists terms that were unique to particular groups.

5 Discussion

Even though we learned a great deal about studying genres in the field and about the differences in genre use by our three respondent groups, in the end, we were disappointed with the results of our study with respect to its usefulness in building a taxonomy of genre terms for further application. We discuss these challenges briefly here and in more detail in (Kwaśnik et al., 2006):

1. Difficulties with Identifying the Genre Unit. For practical purposes we had decided to arbitrarily limit the identification of genre to the Web page as a whole, operationalized

as the URL of that page. In practice, this decision did not work out quite as well as we had envisioned because it is sometimes difficult to ascertain from the interviews which part of the page is the genre that is being described. For example, homepages are often described as both a homepage and an index page, because homepages usually have a list or an index of links embedded in the Web page. One Web page which consists of a search box, search directory and other related links was described as both a search engine and search directory, these labels being dependent on the emphasis of a different element of the page.

2. Difficulty of Eliciting Unambiguous Genre Labels. We learned that the genres of some types of Web pages are more difficult than others for respondents to articulate. For example:

   - Multiple genre terms were applied to one document. Several genre terms (both conceptually similar and different), might be suggested for one Web page. For example, one page was described as a first search step page, navigation page, and menu with the comment "I don't know if I have the vocabulary to describe it."
   - Different types of pages were labeled with same genre term. In the flow of the iterative process of asking for genre terms, respondents had a tendency to use some words repeatedly. One respondent described a page as a highlights page since she saw the word "highlights" on it. Later, she used the same term to describe what to us seemed to be a memo, a news release, a calendar page, and so on.
   - The respondent lacked a term for a given genre. When respondents could not easily name a genre, it was either because they could not think of the term or because they didn't know if a term exists. In the first case, a respondent may just describe the page based on a personal feeling, such as calling it a frustrating page, or admit to not having a word for the page.
   - Nested genres. A Web page can be composed of one or more elements, each of which can be construed as a stand-alone genre by itself. For example, a Web page was described as both an article and a newspaper.
   - Terms were too general or unspecific. When a genre term does not come readily to mind, respondents often provide a general or vague term such as, a page with information.

3. Difficulties with Identifying Genre Attributes. We wanted the respondents to identify the criteria by which an entity (in our case a Webpage genre) is aggregated with like entities or differentiated from unlike ones. The lack of clear and precise labels pointing to a given Web page, as described above, was not our only problem. We have attempted to distinguish genre attributes along a number of criteria: form, content, and purpose. Participants were often vague about clues to these attributes. For instance, they might refer to a page as having a "look and feel" but not specifying in what way. Since journalists are very familiar with the format of a news story page, for instance, they are good at identifying that genre; however, they may have difficulty specifying the clues that helped them identify it because such clues have become implicit and they barely pay attention to them.

4. Challenges in Distinguishing Form and Content. In coding we first flagged the genre term applied to a Web page, and then tried to mark the clues the respondents identified in establishing their concept of that genre. Marking clues in a consistent manner according to the tripartite definition of form, expected content, and purpose was not easy, however. The first two aspects are often convolved in the participants' utterances where it is difficult to ferret out both what they mean or what is in their minds when they invoke a genre term. This convolution of form and content has three manifestations:

    - Identifying aspects of key page elements that signify a page belongs to a genre. For example, one participant invoked a municipality genre, and using the municipality's seal as a clue. How much of a simplified seal "form" would have been enough to qualify it as a municipality page? Or, was she looking at the particular "content" of the seal that made it specific to a municipality of interest?
    - The mixture of form and content in total that establish a page as part of a genre. For example, a participant readily assigned a genre term based on the presence of tabs that allowed for presentation of categories and subcategories. Was it the form of the page, with spatial separation of categories and less visual emphasis given to the subcategories that mattered to him? Or, was it the contextual relationships among the written material on the Web page to which he was referring?
    - Our own preconceived notions of what these "form" and "content" concepts mean. Achieving consistent coding for clues has been difficult when coders bring different conceptions to the task. For example, in deciding on whether an image represented form or content, one coder interprets the meaning of the image and calls it "content," while the other coder, interprets an image as pure "form."

5. Challenges in Identifying Purpose. One of the key ways in which genre provides context is by incorporating an understanding of the genre's purpose or function. While most of the respondents can identify the purpose of the Web page for their own work it is not always clear whether the task requires a particular genre or whether the genre identified happens to be useful (but another one could have been just as useful).
6. Borrowed Purpose. Another situation that causes some confusion is the difficulty in assessing whether the purpose of a genre is generated by the respondents' situation, or whether they recognize the purpose others have for that genre. A homepage of a university that is described as an institutional page has several purposes depending on the discourse community. The purpose of the page from the institution's perspective is to "get its message out," while from the perspective of students and their parents, its purpose is to find out information about the university.
7. Granularity of Tasks. We are finding that people's tasks, as well as the genres that are useful for them are at various levels of specificity. Some are expressed broadly, such as "double-checking facts," while some are narrowly defined, such as "finding the phone number of Joe Smith."

6 Conclusions

In summary, in our study we discovered how difficult it is to study genres "naturalistically." At the same time, we also learned that this is an area of great promise. Rather than trying to study the genres themselves, researchers can instead study human activity through genres, especially those activities that focus on communication (Swales, 1990). This is, obviously, not new. We have studied diaries and letters for many hundreds of years for what they reveal about their writers and the times they lived in. Others have looked at epitaphs, songs, and political slogans. These texts are useful because they can be studied not only at the level of what they say, literally, but what they convey at many other levels. Genres are consensually created and thus they capture not only the meanings of the individual, but also the meanings of the community in which that text is used.

As a result, genre provides an excellent lens for discourse analysis—that is the analysis of language in use in a given community. This type of analysis strives to understand not only the words, per se, but the contexts in which those words acquire meaning. So, for instance, a discourse-based study of rap-music lyrics reveals the culture in which they are created, as well as the values held by the artists and fans. The rap-music genre captures this culture and reveals it simultaneously.

In this vein, we have noticed that several factors that may determine the identification and use of Web genres as well as their place in an overall conceptual map of genres, which our taxonomies try, but fail, to capture. Among these are such factors as the professional affiliation of the person identifying the genre as well as their familiarity with the function for which the genre was created. Most interestingly, though, we have picked up hints—no proof—that perhaps a strong correlation can be made between tasks and genre. That is, perhaps we could structure our Web-genre taxonomies in part by the types of tasks for which a given genre might be useful.

There are many unanswered questions, of course. At the top of the list is the big question of whether a searcher can identify the type of task he or she is contemplating, and second, is the question of whether there is a way of mapping the genres onto the task types in such a way that there is some flexibility and room for individual search strategies. Nonetheless, even a small improvement in the effective use of genre information would be welcome.

References

Bagdanov, A. and Worring, M. (2001). Fine-grained document genre classification using first order random graphs. In *Document analysis and recognition*, Seattle, WA. IEEE Computer Society; 2001.

Bartlett, F. (1932/1967). *Remembering: A Study in Experimental and Social Psychology.* University Press, Cambridge, England.

Crowston, K. and Williams, M. (2000). Reproduced and emergent genres of communication on the world wide web. *Information Society*, 16(3):201–215.

Dewe, J., Karlgren, J., and Bretan, I. (1998). Assembling a balanced corpus from the internet. In *11th Nordic Conference of Computational Linguistics*, Copenhagen, Denmark.

Dillon, A. and Gushrowski, B. (2000). Genres and the web: Is the personal home page the

first uniquely digital genre? *Journal of the American Society for Information Science*, 51(2):202–205.

Freund, L., Clarke, C. L. A., and Toms, E. G. (2006). Towards genre classification for IR in the workplace. In *Proceedings of the 1st International Conference on Information Interaction in Context*, pages 30–36, Copenhagen, Denmark.

Görlach, M. (2004). *Text Types and the History of English*. Trends in Linguistics. Studies and Monographs 139. Mouton de Gruyter, New York.

Haas, S. W. and Grams, E. S. (2000). Readers, authors and page structure: A discussion of four questions arising from a content analysis of web pages. *Journal of the American Society for Information Science*, 51(2):181–192.

Karjalainen, A., Päivärinta, T., Tyrväinen, P., and Rajala, J. (2000). Genre-based metadata for enterprise document management. In *Proceedings of the 33rd Hawai'i International Conference on System Sciences*.

Karlgren, J. and Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *the 15th International Conference on Computational Linguistics*, Kyoto, Japan.

Kennedy, A. and Shepherd, M. (2005). Automatic identification of home pages on the web. In *Proceedings of the 38th Hawaii International Conference on System Sciences*.

Kessler, B., Nunberg, G., and Schuetze, H. (1997). Automatic detection of text genre. In *the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics*, pages 32–38, Madrid. Morgan Kaufmann Publishers.

Kwaśnik, B. H., Chun, Y.-L., Crowston, K., D'Ignazio, J., and Rubleske, J. (2006). Challenges in creating a taxonomy for genres of digital documents. In *2006 ISKO Conference*, Vienna, Austria.

Kwaśnik, B. H. and Crowston, K. (2004). A framework for creating a facetted classification for genres: Addressing issues of multidimensionality. In *Proceedings of the 37th Hawai'i International Conference on System Science (HICSS)*, Big Island, Hawai'i.

Lee, K. J. (2004). Document genre classification for user interface of web search engine. *Ieice Transactions on Information and Systems*, E87D(7):1982–1986.

Lim, C. S., Lee, K. J., and Kim, G. C. (2005). Multiple sets of features for automatic genre classification of web documents. *Information Processing & Management*, 41(5):1263–1276.

Miller, D. and Friesen, P. H. (1984). *Organizations: A Quantum View*. Prentice-Hall, Englewood Cliffs, NJ.

Nilan, M. S., Pomerantz, J., and Paling, S. (2001). Genres from the bottom up: What has the web brought us? In *Proceedings of the American Society for Information Science and Technology Conference*, pages 330–339.

Rosso, M. A. (2008). User-based identification of web genres. *Journal of the American Society for Information Science & Technology*, 59(7):1053–1072.

Roussinov, D. G. and Chen, H. (2001). Information navigation on the web by clustering and summarizing query results. *Information Processing and Management*, 37(6):789–816.

Santini, M. (2008). Zero, single, or multi? Genre of web pages through the users' perspective. *Information Processing and Management*, 44(2):702–737.

Stubbe, A., Ringlstetter, C., and Schulz, K. U. (2007). Genre as noise—noise in genre. In *Proceedings of the IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data*, Hyderabad, India.

Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, New York.

Toms, E. G., Campbell, D. G., and Blades, R. (1999). Does genre define the shape of information? the role of form and function in user interaction with digital documents. In *American Society for Information Science; ASIS '99*, Washington, DC. Information Today; 1999.

Watters, C. and Shepherd, M. (1999). Cybergenre and web functionality. In *the Thirty-second annual Hawaii International Conference on Systems Sciences*, Maui, Hawaii. IEEE Press.

Yates, S. J. and Sumner, T. (1997). Digital genres and the new burden of fixity. In *Hawaiian International Conference on System Sciences (HICCS 30)*, Wailea, HA. IEEE Computer Press.

zu Eissen, S. M. and Stein, B. (2004). Genre classification of web pages: User study and feasibility analysis. In Biundo, S., Frühwirth, T., and Palm, G., editors, *Proceedings of the 27th Annual German Conference on Artificial Intelligence (KI 04)*, pages 256–269, Ulm, Germany. Springer.

Common to E J T	Common to E J	Common to J T	Common to E T
article	about us page	education page	book
government page	advertising page	front page	commercial page
home page	blog	Gateway	journal article
index	company home page	how-to page	magazine
information page	corporate page	link page	resource page
list	definition page	Newspaper	organization page / organization home page
main page	entry page	organization page	
search engine	FAQ	full story list / list of stories	
search page	letter	magazine / magazine article	
search results	list of links		
site map	navigation page		
summary	organization home page		
table of contents	PDF		
Magazine/ magazine article	press release		
	question and answer		
	terms and conditions		
	archive of abstracts /archives		
	executive overview / overview		
	magazine /magazine article		
	meeting notes / minutes		
13 Genre Terms	16 Genre terms	7 Genre Terms	5 Genre Terms
1 similar terms	4 similar terms	1 similar term	1 similar term
14	20	8	6

Table 4: Examples of common genres

Engineers	Journalists	Teachers
change summary page	editorial (2)	activity
coding manual	fact box	lesson plan (3)
compiler listing page	gray page	lesson resource
compilers home page	index of news coverage	list of course offerings
data (3)	index to the news stories	list of lesson plan
datasheet	interview	outline of a textbook
directory to white papers	list of headlines	
explanation of the code	news blog	
library (2)	news entry	
license	news page (2)	
man page (3)	news portal	
manual	news release	
online manual	news story	
software description page	news summary page	
software test document	press release	
standards	press resources page	
technical committee report	story (2)	
technical paper	story list	
test plan (2)	transcript of an interview	
White paper		

Table 5: Examples of unique genres