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
crowston@syr.edu, bkwasnik@syr.edu, jrublesk@gmail.com

Abstract. A document's genre reflects the purpose of a document and as such is potentially useful meta-data to improve search effectiveness. Using genre in an information retrieval system seems to require a taxonomy of genres to provide a controlled vocabulary and to show relations among genres. In this chapter, we report on a study to develop a 'bottom-up' genre taxonomy, that is, from the genre terms identified by informants. We collected a total of 767 genre terms from 52 respondents (teachers, journalists and engineers) engaged in natural use of the Web, and reduced this list to a set of 298 genres.
We report on various difficulties we encountered in the study. Respondents frequently had difficulty coming up with an unambiguous genre label for a page, offering several possibilities, or applied the same label to many pages. In many cases, respondents could not think of a term, or applied an overly general term, such as an "information page". Furthermore, even when respondents did offer a clear genre term, they often were unable to say what about the page led to that choice. These difficulties seem to reflect underlying problems in the definition of genres as social constructions, that have meaning only in use.

1 Introduction

Web search engines such as Google or Yahoo determine relevance of Web pages according to the occurrence of words in the pages indexed by the engine (additional information is then used to rank these results). Unfortunately, such searches are not always sufficient to solve information needs since task-driven searchers often must distinguish between documents that share a set of keywords (i.e., a topic) but assume a different form to serve a different purpose 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" could yield the bulk of results referring to digital-camera sellers, not magazines, discussion forums or blogs. A renewed search with a more refined query might prove incrementally more effective, but might just as easily yield mixed results. Efforts to locate a current,

* This research was partially supported by NSF IIS Grant 04-14482. We thank John D'Ignazio and You-Lee Chun for their contributions to this research project.

trustworthy and pertinent discussion forum might require considerable manual searching through search results.

One way to improve the precision of a search and to ensure a better match of the results to a user's task is to utilize additional metadata to distinguish or group relevant and irrelevant documents. We focus in particular on the role of document genre. Document genre can be defined as "essentially a document type based on purpose, form and content" [1, p. 2]: e.g., a digital-camera *advertisement*, a digital-camera *review*, or a *schematic drawing* of a digital camera.

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 [2]. As well, knowledge of the genre can be exploited in a number of tasks because genre provides some fixity to otherwise infinitely variable texts [3]. Genre acts as a template of attributes that are regular and can be systematically identified. Most important, genre reflects the purpose of documents. Therefore, if a Web search could use genre metadata, it might be possible to use it to specify the desired information more precisely and find a document whose purpose matches the user's. Indeed, [4] reports that searches on America Online that included a genre term such as *recipe* or *lyrics* seemed to yield more precise search results. Towards this end, 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., [5,6,7,8,9,10]).

1.1 What is the purpose of a genre taxonomy?

At the core of building applications that apply the notion of genre to information-provision tasks is the fundamental problem of identifying, defining, labelling and organizing the genres in a useful structure—that is, a genre taxonomy [11]. Such a structure enables several functions:

- First, it provides a controlled vocabulary that resolves the issue of variation in labelling and meaning: synonyms, acronyms, variant spellings, grammatical variants such as plurals, and so on.
- Second, a taxonomy can arrange the entities in a meaningful structure—typically a hierarchy or a facetted scheme—where the scope and definition of each entity is further described by its relationship to other entities. Thus, we can say a *digital-camera review* is a kind of *product review*, the *product review* being a more inclusive term. Other structures are possible as well, such as part-whole arrangements in which entities can be described by their componential parts. For example, an *abstract* is part of a *scholarly article*. In this case an *abstract* is a genre that is typically part of another genre, each sharing part of the functional properties that make knowing the genre of something so useful.
- A well-designed taxonomy is useful at both ends of the information-provision process. From the user's perspective it allows for a more cognitively efficient

way of choosing terms for a query. Rather than "thinking of a genre off the top of your head", a user can choose from an organized array. The organization further allows expansion of the search to more general terms, or conversely a narrowing of the search for more specificity. For retrieval, a taxonomy allows for gathering terms with similar meaning together under one label, allowing for adjustments in the granularity of the results.

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 efforts to develop such taxonomies, reported below, suggests that consensus is unlikely. As many researchers have found, reaching consensus on genre terms, their attributes, or their relationships to each other is not easy. This difficulty applies both to genre information gleaned from "genre use in the wild" and to reaching intercoder consistency for manually marked-up genre pallettes in research studies. As [12] comments, there seems to be "a gap between genre theory and the practice of average users". The purpose of this chapter is to support this claim by first briefly reviewing prior work on developing taxonomies of Web genres and second to 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" [1, 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 [13].) 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 definition of document genre we quoted above includes both socially recognized form and purpose, and it is possible to make a logical division between intrinsic genre attributes (i.e., form and content) and the extrinsic function that genre fulfills in human activities. Many studies focus on the first aspect, that is, the nature of the document genres themselves or on the attributes of the

documents that will allow them to exploit genre for knowledge-representation functions. From studying non-digital genres we know that the roles 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 [14]. 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 [13].

What has changed from formal genre 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 (as also argued by [4]). This recognition 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, their identity as examples of genres has also evolved. New document genres have emerged [6,15], 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.

The 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 varied communities of users. In addition, even if some 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.

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.

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 sufficiently expressive. This means that the taxonomy comprises genres that are able to adequately represent the documents to which it will be applied. As [13] 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 [6, 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, [9] used only eight genres (*help*; *article*; *discussion*; *shop*; *portrayal, non-private*; *portrayal, private*; *link collection*; and *download*). At the other extreme, [16] 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, [17] 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 [18,19] made initial use of [5]'s genre taxonomy, for example, while [6]'s taxonomy of document genres was based on the Art and Architecture Thesaurus [20] and used by [21]. This question is important methodologically because the use of any starter palette frames how Web documents in a corpus will be viewed. A researcher may end up with a new taxonomy that does not much resemble the one she started with, but that was almost certainly influenced by 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" [7, 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 (see for example [4,11] for examples). Such an approach seems desirable because it avoids imposing an *a priori* vocabulary with which users may lack familiarity. As [1, 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 [9] notes, "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. [6] 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" [7]. Alternately, multiple genre terms may seem appropriate to describe a particular document. Web documents may instantiate multiple genres [6,7]. As [22, 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 "...scholarly material can be seen as a super-genre that covers help, article and discussion pages" [9]. 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") [5].

In the face of the difficulties noted above, researchers may intervene by explaining the genre concepts to participants (e.g., [9]) and/or modifying the genre terms supplied by participants (e.g., [5]). 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" [9, p. 4].

Other recent attempts at developing a genre classification aim at discovering relevant attributes automatically, rather than identifying them *a priori* [23,24,25,26]. These attributes are then used to cluster documents into genres. 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 [15,27,13,28].

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 (a full description of the projects is beyond the scope of this chapter). 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. 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, so 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 [15, p. 202], genres are no longer necessarily "slow-forming, often emerging only over generations of production and consumption". Thus, we assumed that

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

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. (Because we recognized that the genre terms elicited would likely be somewhat specific to the groups studies, we planned to use the same communities in later phases of the research plan.) Our study solicited information about genre from three groups of respondents: K-12 (kindergarten through grade 12, i.e., primary school) 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 was thought likely to be important for their tasks.

Respondents were recruited via a snowball-sampling approach, chosen to fit our goal of collecting a wide range of tasks, genres and genre attributes. (A more systematic sample would have be required for making inferences to a population,

e.g., for documenting the relative frequency of use of terms, but that was not our purpose in this study.) All respondents were working full-time in one of these three professions and had the required educational background to do so, making them qualified to identify genres relevant to their work. Ages ranged from early twenties to late fifties; 40% were female and 60% male.

3.3 Data elicitation

In general, our data-elicitation goal was to identify, for a collection of Web pages, the genre (or genres) of the page, the clues each respondent used to recognize the genre (or genres), 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. These interviews were carried out in the respondents' offices, using their own computers. Respondents were asked to carry out a Web search for a real task of their own choice (e.g., a journalist searching for background information on an interview subject; an engineer looking for software documentation). 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 visited and the sequence in which they were visited. These data were used to re-create the search process. 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 digital 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:

- The captured Web pages.
- Transcripts of audio files from the debriefing for the 32 respondents—19 journalists and 13 teachers (3 of the original transcripts were corrupted by problems with the digital recorder 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. Respondents had the option of offering multiple terms for the same page.

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 or a taxonomy that represented a particular domain, but rather to build a taxonomy to use in subsequent stages of the research with these groups. We also wanted the taxonomy to be used eventually with a general audience. Thus we needed genre terms that we believed would be understood by our future study participants, who might not be from the same exact discourse communities as the participants in this study. Thus, we eliminated:

– Terms that had only a personal meaning to the respondent, e.g, "good page."
– Terms that were so situation- or domain-specific that they would not be understood in any other context, e.g., an "uncontrolled resource page" from an engineer.

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 [29]:

	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.

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		

Table 4. Examples of common genres (E = Engineers, J = Journalists and T = Teachers)

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
data (3)	index to the news stories	offerings
datasheet	interview	list of lesson plan
directory to white papers	list of headlines	outline of a
explanation of the code	news blog	textbook
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

1. Difficulties with identifying the genre unit. 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*. In these cases, it was sometimes difficult to ascertain from the interviews which part of the page had the genre that was being described. For example, homepages were often described as both a *homepage* and an *index page*, presumably because homepages often have a list or an index of links embedded in the Web page. One Web page that consisted 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 with 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 as respondents struggled to find an appropriate term. 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 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.
   - 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. We expected respondents to identify genre based on document attributes of form, content and purpose. However, 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 has not been 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. The *homepage* of a university that is described as an *institutional* page has several purposes depending on the perspective of the user. 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 provide different kinds of 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 [30]. 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

1. Rosso, M.A.: User-based identification of web genres. Journal of the American Society for Information Science & Technology **59**(7) (2008) 1053–1072
2. Bartlett, F.: Remembering: A Study in Experimental and Social Psychology. University Press, Cambridge, England (1932/1967)
3. Yates, S.J., Sumner, T.: Digital genres and the new burden of fixity. In: Hawaiian International Conference on System Sciences (HICCS 30), Wailea, HA, IEEE Computer Press (1997)
4. Karlgren, J.: Conventions and mutual expectations: Understanding sources for web genres. [31]
5. Dewe, J., Karlgren, J., Bretan, I.: Assembling a balanced corpus from the internet. In: 11th Nordic Conference of Computational Linguistics, Copenhagen, Denmark (28–29 January 1998)
6. Crowston, K., Williams, M.: Reproduced and emergent genres of communication on the world wide web. Information Society **16**(3) (2000) 201–215
7. Haas, S.W., Grams, E.S.: 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) (2000) 181–192
8. Nilan, M.S., Pomerantz, J., Paling, S.: Genres from the bottom up: What has the Web brought us? In: Proceedings of the American Society for Information Science and Technology Conference, Washington, DC (2001) 330–339
9. Meyer zu Eissen, S., Stein, B.: Genre classification of web pages: User study and feasibility analysis. In Biundo, S., Frühwirth, T., Palm, G., eds.: Proceedings of the 27th Annual German Conference on Artificial Intelligence (KI 04), Ulm, Germany, Springer (2004) 256–269
10. Freund, L., Clarke, C.L.A., Toms, E.G.: Towards genre classification for IR in the workplace. In: Proceedings of the 1st International Conference on Information Interaction in Context, Copenhagen, Denmark (2006) 30–36
11. Rosso, M.A., Haas, S.W.: Identification of web genres by user warrant. [31]
12. Sharoff, S.: In the garden and in the jungle: Comparing genres in the BNC and Internet. [31]

13. Kwaśnik, B.H., Crowston, K.: 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 (2004)
14. Toms, E.G., Campbell, D.G., Blades, R.: 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 (1999)
15. Dillon, A., Gushrowski, B.: Genres and the web: Is the personal home page the first uniquely digital genre? Journal of the American Society for Information Science **51**(2) (2000) 202–205
16. Görlach, M.: Text Types and the History of English. Trends in Linguistics. Studies and Monographs 139. Mouton de Gruyter, New York (2004)
17. Lee, D.Y.W.: Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. Language Learning & Technology **5**(3) (2001) 37–72
18. Lim, C.S., Lee, K.J., Kim, G.C.: Multiple sets of features for automatic genre classification of web documents. Information Processing & Management **41**(5) (2005) 1263–1276
19. Stubbe, A., Ringlstetter, C., Schulz, K.U.: Genre as noise—noise in genre. In: Proceedings of the IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data, Hyderabad, India (2007)
20. Petersen, T.: Art and Architecture Thesaurus. Oxford, New York (1994)
21. Roussinov, D.G., Chen, H.: Information navigation on the web by clustering and summarizing query results. Information Processing and Management **37**(6) (2001) 789–816
22. Santini, M.: Zero, single, or multi? Genre of web pages through the users' perspective. Information Processing and Management **44**(2) (2008) 702–737
23. Bagdanov, A., Worring, M.: Fine-grained document genre classification using first order random graphs. In: Document analysis and recognition, Seattle, WA, IEEE Computer Society; 2001 (2001)
24. Karjalainen, A., Päivärinta, T., Tyrväinen, P., Rajala, J.: Genre-based metadata for enterprise document management. In: Proceedings of the 33rd Hawai'i International Conference on System Sciences. (2000)
25. Karlgren, J., Cutting, D.: Recognizing text genres with simple metrics using discriminant analysis. In: the 15th International Conference on Computational Linguistics, Kyoto, Japan (1994)
26. Kessler, B., Nunberg, G., Schuetze, H.: 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, Madrid, Morgan Kaufmann Publishers (1997) 32–38
27. Kennedy, A., Shepherd, M.: Automatic identification of home pages on the web. In: Proceedings of the 38th Hawaii International Conference on System Sciences. (2005)
28. Watters, C., Shepherd, M.: Cybergenre and web functionality. In: the Thirty-second annual Hawaii International Conference on Systems Sciences, Maui, Hawaii, IEEE Press (1999)
29. Kwaśnik, B.H., Chun, Y.L., Crowston, K., D'Ignazio, J., Rubleske, J.: Challenges in creating a taxonomy for genres of digital documents. In: 2006 ISKO Conference, Vienna, Austria (2006)

30. Swales, J.M.: Genre Analysis: English in Academic and Research Settings. Cambridge University Press, New York (1990)
31. Mehler, A., Sharoff, S., Rehm, G., Santini, M., eds.: Genres on the web: Computational Models and Empirical Studies. Springer (In Press)