Challenges in Creating a Taxonomy of Genres of Digital Documents*

Abstract: We report on the process and difficulties of building a taxonomy of genres of digital documents. The taxonomy is being created to be used in the experimental phase of an ongoing study to learn about the usefulness of providing genre information to support information-seeking tasks. To build the taxonomy, we conducted field studies to collect webpage-genre information from 55 respondents: K-12 teachers, journalists, and engineers, who routinely use the web for information seeking. Challenges described in this paper include the difficulties respondents experienced in identifying and naming genres and that the researchers faced in unambiguously linking the genre identifications with clues to genre attributes and purposes.

1.0 Introduction

We report on one phase of a project whose aim is to discover whether and how identifying the genres of digital documents helps in a variety of information-seeking tasks (Crowston & Kwaśnik, 2005-07). Our project was motivated by the recognition that many bottlenecks in successful access to information occur because users are not able to adequately specify their needs, and systems are not able to adequately disambiguate the results of searches, resulting in undifferentiated and overwhelming amounts of information. The process of information seeking is often detached from the purpose for which the information is sought, and the context in which the information will ultimately be used. Incorporating that context is of great importance in tailoring an information-seeking experience, but it is not easy to do.

We propose that the inclusion of genre metadata might be useful in remedying this situation. Documents are communicative acts produced by members of a discourse community for a mutually understood purpose as reflected in the document’s genre. Every genre embodies not only a particular form and content, but also a recognized function – context in a capsule, as it were. We hypothesize that recognizing and using genre metadata for searching will allow the results to be contextualized for the searcher and therefore more precise.

One of the challenges of studying genre in general is that there does not seem to be a consensus on what a genre is or how best to identify, construe, or study genres. In general though, we note that most definitions include some consideration of the form of the document and often the expected content, as well as the notion of intended communicative purpose. The definition of genre we have adopted for our study: “a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form” (Orlikowski & Yates, 1994, p. 543), appeals to us because of its recognition of genre as a fusion of form, function and content that is situated in a context of human endeavor.
1.1 A Summary of Our Study

Our overall project has three phases:

- Phase I. Harvesting and identifying a test-set of web pages that have been coded for the motivating task, genre terms and the clues people use to identify each genre’s form, content and function.
- Phase II. In the second phase, presently underway, we attempt to build a faceted taxonomy of the genres identified in Phase I. This is the phase on which we focus in this paper.
- Phase III. In the final phase we will test the utility of including genre information in an information-access environment that will be simulated using the results of the analysis in Phase II.

1.1.1 Information about Genres from the Searchers’ Perspective

In this section we briefly summarize the data-collection process in Phase 1 because it affects the conceptualization and building of the taxonomy described in subsequent sections of this paper. 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. 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.

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

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

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.
1.1.2 The Process of Eliciting Genre Information from Respondents

In general, our data-collection 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.

2.0 Creating a Working Taxonomy of Genres

2.1 Why We Need a Taxonomy of Genres

Phase III of our study will consist of a series of experimental simulations, based on our collection of web pages, in which the subjects will be able to formulate queries or navigate results using genre metadata to enhance the process. We intend to simulate such enhancements by providing aids for query construction, clustering of identical or related genres, ranking of documents by a match on genre, and enabling genre-specific search strategies.

Before we proceed to testing these ideas, however, we are faced with the formidable task of describing and organizing the genres and their attributes into a working taxonomy, which we consider an important formative step. Such a taxonomy will enable us to manipulate the experimental conditions and the interface to find out:

- Which specific facets of genre improve performance most?
- To what extent does using genre metadata to cluster and/or rank documents improve performance and utility?
- Are there differences in utility between interfaces that explicitly name genres and those that use other methods of labeling clusters of documents (e.g., by providing example documents);
- What are the effects of granularity, that is, the specificity of genre identification? Is there, perhaps, a basic level of genre that is neither too general and abstract, nor too specific?
2.2 Why a Facetted Taxonomy?

We recognize that because genres embody attributes of form, function, and content, they are complex and thus do not lend themselves to classification using a simple set of criteria. Thus, we are attempting to create a facetted classification in which all the important aspects of the genres will be taken into consideration. Genre is a subtle and difficult-to-define notion, and our field studies, described in this paper confirm this. Moreover, the possible set of genres is very large, and we have no way of knowing when and if we have a complete set. Without a strong theory of genre to guide us, it is problematic to set up a classification structure that will accommodate all known and future genres. For these reasons, we believe that a facetted scheme will serve us well for the purposes of this study.

A facetted classification is a useful tool because:

- It does not require complete knowledge of all genres;
- It is relatively hospitable to new genres, so long as they can be described by the fundamental facets of form, purpose, and content and any subfacets yet to be identified;
- Facetted schemes have flexibility in that the elements can be invoked in an almost endlessly flexible way;
- It allows for requisite expressiveness because each facet can be developed independently to the degree of specificity needed. For example we can have a more general taxonomy of purposes, but a much more specific taxonomy of genre types;
- It can accommodate a variety of theoretical structures and models – again, a different structure for forms, for tasks, for genres – thereby allowing a number of perspectives that can be invoked in the future phase of our study.

There are, of course, some obstacles: among these are the difficulty of establishing a robust set of basic facets, a lack of a “natural” relationship among the facets (e.g., the relationship of a given genre to a given task), and the difficulty of visualizing or representing a multidimensional scheme in any implementation of it. We feel that on balance, a facetted scheme is our best option. We need a taxonomy that will do justice to the richness and complexity of the concept of genres, and which can be developed and expanded as needed for future use (Kwaśnik & Crowston, 2004).

3.0 Challenges to Creating the Taxonomy

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 challenges fall into several categories: difficulties in identifying the genres themselves and difficulties in identifying genre attributes such as form and content, and finally difficulties in unambiguously linking the genres with their purpose.

3.1 Challenges to Identifying Genres

The first step in building a taxonomy is to identify the entities. In this case we were looking for genres and asked our respondents to label the various web-page “types.” Several difficulties emerged, which we describe below.

3.1.1 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
has not worked 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 search box, search directory and other related links is described as 
both a **search engine** and **search directory**, these labels being dependent on the emphasis of 
a different element of the page.

### 3.1.2 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. We describe a few examples of such instances:

- **Multiple genre terms are 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.” Under 
  such circumstances, respondents might come up with genre terms based on either the 
  purpose or the content of the page, personal feelings, or just based on words on the 
  page that catch their attention.

- **Different types of pages are labeled with same genre term.** In the flow of the 
  iterative process of asking for genre terms, respondents have 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 a memo, a news 
  release, a calendar page, and so on.

- **The respondent lacks a term for a given genre.** When respondents can’t easily 
  name a genre, it is either because they can’t think of the term or because they don’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. If the page is identifiable, but the respondent doesn’t know if there 
  is an appropriate word for it, he or she may struggle to find a term, but in the end, not 
  be sure of it.

- **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**. We coded **article** as a Genre 
  Component of **newspaper**.

- **Terms are 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**.

### 3.2 Difficulties with Identifying Genre Attributes.

Another step in building a taxonomy is 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.

3.2.1 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. Should a passage be a clue for form, or content, or both? 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.”

3.2.2 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).

- **Borrowed Purpose.** Another situation that causes some confusion is the difficulty in assessing whether the purpose of a genre is generated by the respondent’s 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.
- **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.” This range and variety presents a problem in creating a taxonomy that has as one of its facets the tasks associated with the genre.
We are, therefore, working on a more general formulation of tasks that would capture the experience of our respondents but be applicable to a wide range of genres.

4.0 The Building Blocks of a Taxonomy: A Discussion

We describe a work in progress. Our intent is to build a taxonomy in which the entities are genres of webpages as they are used and recognized by people in the course of their usual professional endeavors. Because genres are a fusion of form, content, and purpose, it is important for us to identify these aspects of the genres as well and incorporate them into the taxonomy as facets. Thus far, we have been able to elicit a range of genre terms, and at least some of the clues to these genres’ attributes and function. Along the way, we have encountered some dilemmas in the sense that it seems obvious that people do in fact recognize and use genre information, but that they are not always explicitly aware of this process, and thus are not able to offer us clear and unambiguous terms for genres or their attributes. As well, we have noticed that our three respondent groups have identified different sets of genres (in addition to many that overlap), and a different set of tasks and goals associated with those genres. This is both encouraging, in that we are confident that we have identified a range of genre terms and uses, but it has also presented us with the added burden of analysis.

5.0 Conclusion

For many of the reasons outlined above, we are building the taxonomy relying on some compromises and some interpretation on our part. Nevertheless, we believe that having the taxonomy emerge for the most part from the respondents’ own reports is worth the effort in that it will add validity to the experiments in our next phase where we will attempt to describe and assess the value of the genre information in information seeking. Besides being critical to our own study, the digital-document genre taxonomy can be augmented in the future, and we hope it will be useful in other related studies, as well as in other efforts to build multidimensional representations of complex phenomena.

6.0 Notes

* This research was partially supported by NSF Grant 04-14482. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the National Science Foundation.

7.0 References

