Challenges in Creating a Taxonomy for Genres of Digital Documents

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 & Kwa□nik, 2005-07). The project has three phases:

I. Harvesting and identifying a test-set of webpages from journalists, teachers, and engineers, three groups that 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. For each webpage we ask the respondent to identify the task, the type of Webpage (genre), the clues the respondent used to identify the genre, and the usefulness of that document to their task.

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.

III. In the final phase we will test the utility of including genre information. Using the taxonomy created in Phase II, we will manipulate a simulated search environment to test the effect of genre identification on such tasks as query formulation, searching, and processing of search output.

Challenges in Studying Genres.

One of the challenges of studying genre in general is that there does not seem to be 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. In general though, we note that most definitions include some consideration of the form of the document and often the expected content. Most definitions also include the notion of intended communicative purpose. As well, a document is of a particular genre to the extent that it is recognized as such within a given discourse community. In fact, successful membership in any number of social contexts requires a fluency in the genres in use in that context. The definition we have adopted for our study (Orlikowski & Yates, 1994), 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.

Before we proceed to the testing of genre as metadata for information-seeking activities, we are faced with the formidable task of describing and organizing the genres and their attributes into a working taxonomy. We consider this an important formative step, but also, in itself, a useful tool for use by future researchers, including ourselves.

Creating a Faceted 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 will attempt to create a facetted classification in which all important aspects of the genres will be taken into consideration (Kwa□nik & Crowston, 2004). The core facets, as well as the particular attributes, will be derived from a content analysis of the think-aloud protocols gathered for Phase I – that is the respondents' reports on the genres and their form, content and function.

Challenges to Creating the Taxonomy

Difficulty of eliciting specific genre labels. Based on think-aloud sessions and interviews with more than 30 journalists and teachers so far, we have learned that some types of webpages are more difficult to articulate than others. While all (or nearly all) of these participants share the use of terms such as *homepages*, *search pages*, and *list pages*, there is much less agreement on pages that

(a) comprise chiefly content,
(b) do not include a large proportion of links, and
(c) are not homepages.

Participants have referred to these pages as "*resource*," "*content"* and "*information"* pages – terms that are too general and unspecific for use in a taxonomy. The ostensible lack of a shared vocabulary for these and

other types of webpages -- even among members of the same discourse community -- poses a significant challenge in the development of a taxonomy.

Difficulties with creating coding categories. It is sometimes difficult to express one webpage using only one genre term for it, especially if the participant is unsure about the terms in the first place. For example, several genre terms (both conceptually similar and different), might be suggested for one webpage. For example, the same webpage might be described as a *definition*, a *textbook*, and *news story*. The lack of clear and precise labels pointing to a given webpage is not the only problem, however. Participants are also often vague about clues, for instance, referring to a page as having a "look and feel" but not specifying in what way. This indicates that they perceive the differences in "look and feel" of different genres, but they have difficulty choosing a particular element of the webpage as a clue.

Conclusion

In our paper we will discuss how we are dealing with these difficulties of identifying and describing the genres of digital documents with sufficient richness and precision to enable us to create a working taxonomy for the future phases of our work.

References

Crowston, K. and Kwa□nik, B.H. (2005-07). How can document-genre metadata improve information-access for large digital collections? NSF IIS Grant 04–14482.

Kwa□nik, B. H., & Crowston, K. (2004). A framework for creating a facetted classification for genres: Addressing issues of multidimensionality. In *Proceedings of the Thirty-Seventh Hawai'i International Conference on System Science (HICSS-37)*. Big Island, Hawai'i.

Orlikowski, W. J., & Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative Sciences Quarterly, 33*, 541–574.