Fall 2003

Can document-genre metadata improve information access to large digital collections?

Kevin Crowston, *Syracuse University*
Barbara H. Kwasnik, *Syracuse University*

Can document-genre metadata improve information access to large digital collections?

Kevin Crowston and Barbara H. Kwasnik
Syracuse University, School of Information Studies

4-206 Centre for Science and Technology
Syracuse, NY 13244-4100 USA
Telephone: +1 (315) 443-1676
Fax: +1 (315) 443-5806
Email: {bkwasnik, crowston}@syr.edu

Draft of June 7, 2003
Submitted to *Library Trends*

Can document-genre metadata improve information access to large digital collections?

Introduction

Computerized information-access systems face a fundamental limitation: they know what documents say, but not what they mean or for what purposes they might be useful. Extracting and representing the meaning of documents is difficult and time-consuming to do, and automatic systems still have significant limitations. We note though that humans rarely have to read every word of a document to understand its purpose. Instead, people take a shortcut: they start by identifying the kinds of documents they are faced with (i.e., the document's genre), and then use different types of documents inappropriate ways. For example, a grant proposal is used differently from a syllabus, a product brochure or a bank statement. Accordingly, differences in an information situation are often reflected in the *kind* of document that is considered helpful (e.g., a problem set, a lesson plan and a tutorial about mathematics are all about math but useful in different situations). Information-access systems would be more useful for many tasks if they could similarly distinguish the purpose of documents and handle them in appropriate ways.

In this paper we discuss the possibility of improving information access in large digital collections through the identification and use of document genre as a facet of document and query representation. First, we provide some historical background on the concept of genre and the approach it provides to the problem of incorporating context into information retrieval. We outline the framework of the information-retrieval problem with respect to genre, and some traditional resolutions that have been attempted. Finally we outline a research agenda that addresses some of the questions and issues that investigating genre entails.

Theory: Document genre

Rhetoricians since Aristotle have attempted to classify communications with similar form or purpose into types or "genres." Numerous definitions of genre, or discourse type, have been suggested (e.g., Longacre, 1983; Miller, 1984; Swales, 1990). In our discussion, we draw on the definition of genre proposed by Yates and Orlikowski (1992), who describe genre as "a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form" (Orlikowski & Yates, p. 543). For instance, this document is an example of the journalarticle genre. It has a form familiar to most researchers and practitioners, and is monitored by the journal's editorial policies as well as the profession's communication practices. There are many document genres: some common, such as a report or a newsletter, and others restricted to specific domains, such as the course syllabus or a problem set in higher education. Genre is applicable to electronic as well as physical documents. For example, in a study of Web documents, Crowston and Williams (2000) were able to identify documents of many familiar genres and of a few genres that seemed to be new to the Web, such as the home page (Dillon and Gushrowski, 2000) or the hotlist.

Genre is useful because it makes documents more easily recognizable and understandable to recipients, thus reducing the cognitive load of processing them (Bartlett, 1932/1967). Yates and Sumner (1997) argue that on the Web, genres help in both the production and consumption of documents because genre adds "fixity" in a medium that does not otherwise distinguish very well between text types (say, a book and a post-it). For example, since this article conforms to the journal-article genre, a reader can quickly determine the purpose of our communication, locate relevant sections, evaluate the document's contribution and possibly to use it to prepare build further queries. In our preliminary studies of people searching the Web (Roussinov, Crowston, Nilan, Kwasnik, Liu & Cai, 2001), we observed that the genre of the document was one of the clues used in assessing document relevance, value, quality and usefulness.

The problems of information access.

To explain how genre can be useful, we will first briefly review the problem faced by an information-access system. An information-access system has three components: 1) the users, who approach the system with contextually-based information needs, 2) a store of information (e.g., the documents or databases), and 3) an intermediating mechanism to connect needs and information. The intermediary may be a person, a search algorithm, a browsing environment, or a summarizer, among others.

The basic process of matching users' needs to potentially useful information in the system is complicated by many factors. First, problems may occur due to improper or incomplete representations of the information itself. When the information-access system is created, the documents or texts must be represented in such a way that they can be retrieved again as needed. Librarianship has occupied itself for over a century with systematic approaches to organizing and representing information in systems. In creating bibliographic records we call this process cataloging; in organizing actual documents or topics for meaningful retrieval, we call it classification; in providing access to bibliographic databases we call the representation process indexing.

There are similar processes of information representation on the Web and in many other applications in which large stores of information are prepared for eventual use in the future. Many of the schemes are adaptations of traditional schemes, such as that used on Yahoo.com, the Dublin Core Project, or the GEM Metadata Project for educational materials (http://www.thegateway.org/). Others comprise grass-roots, emergent sorts of organization and representations, such as the evolving classification on eBay.com or on amazon.com) (Kwasnik & Liu, 2000; Kwasnik, 2002). An increasingly popular approach relies solely on the full-text of the documents.

Another problem that may arise is that the process itself of matching users' queries to the document representations may be inadequate or faulty. Much effort on the part of information scientists has been spent in developing and perfecting search strategies, including various matching algorithms, probabilistic techniques, citation mapping, and natural language processing. These efforts struggle with many obstacles—among them the difficulties of evaluating search results in real environments, as well as problems of scaling, reliability and the representation bottleneck.

On the user side, we have people in need of information. Often, though, users are unable to precisely specify what it is they need, and even if they do, the way in which humans articulate their needs produces a great variety of expression. The problem of appropriate representation of users' queries is not just a question of finding the correct representation according to some absolute criteria. Because information use is situated in specific contexts, there is also the need to be able to represent the information in such a way that a match can be made not only on the level of physical description and topic, say, but also in terms of matching the information with a potential use. For example, consider a person approaching a system with the query "I want to prepare a Passover dinner." At a certain level we can see that there is a need for concrete information in the form of actual recipes. We might even interpret this as a "known item search." However, recipes may satisfy the need only partially, since the person may want to know much more about the rituals and meanings of a Passover dinner and not just the food itself. The information need may be either broader than what is asked for, or much narrower and specific. We know that people ask for what they expect they can get that will most closely match what they *really* want, and thus their requests are often presented in a compromised form.

Thus, we can see that topic alone is not enough to define an information problem because different users may require different solutions to seemingly similar information problems. Indeed, even the same user may require different information at different times. These different needs arise because the situation (or context) of a user determines not only what topics are requested and what strategies are invoked in searching and evaluating output, but also what types of resources are considered relevant and useful. For example, methods for learning mathematics (a topic) may be construed differently by a student, by a parent and by a classroom teacher because of their different information use situations. While we know that it is important to understand the situation of the user, the representation of the situation and then its implementation in a system is a difficult problem. Our efforts to create user profiles, universal situation grammars, and so on suffer from limitations of scope to specific domains and lack of extensibility and flexibility.

Why we think identification of genre would be useful

We suggest that enhancing document representations by incorporating non-topical characteristics of the documents that signal their purpose—that is, their genre—will enrich document (and query) representations in such a way that they resonate more truly with the information need of a user as situated in a particular context.

Because most genres are characterized by both form and purpose, identifying the genre of a document provides information as to the document's purpose and its fit to the user's situation, which can be otherwise difficult to assess. For instance, a university professor looking for information about computer database systems for the class that she teaches would most likely be interested in documents of educational genres (e.g., syllabi, assignments, class

notes). On the other hand, when working on a research paper in the database area, the same professor would more likely appreciate scholarly work (e.g., papers, annotated biographies, calls for papers). The relevant documents for these two searches would be quite different, even though the topic and query keywords might be nearly the same.

Explicit identification of genre seems particularly important for large digital collections because any search of these collections usually retrieves documents with a diversity of genres, and, what is worse, these genres are undifferentiated by obvious clues to their identity. This is in contrast to non-digital information-seeking situations in which the searcher generally has an idea of what sorts of documents exist in the collection. Even if he or she does not, clues of physical form and location increase the chances that a document's genre is recognized. For example, a user searching in a library can visually distinguish CDs from books, from encyclopedias, or from newspapers. Similarly, a user searching a database containing only journal articles has already implicitly restricted the search to that genre of documents. On the Web, however, a search of a large and diverse document collection will usually retrieve some documents of relevant genres along with many documents of irrelevant genres—a low precision result—even if all retrieved documents conform to search specifications regarding the topical content of the document.

Recognition of genre also has implications for automated methods of representing documents, such as automated summarization and indexing. A one-size-fits-all approach to summarizing or evaluating Web documents without regard for their form and function is likely to misrepresent many of them. For example, a newspaper article can be summarized by the first few sentences of the document, but such an approach will not work for a home page or a frequently-asked-questions document (FAQ) (Marcu, 1997). When medical information is sought, it makes a difference to the evaluation whether the document retrieved is a newsletter, a personal homepage, or a hospital's patient information site.

How librarians have addressed the notion of genre in library information systems

We do not mean to imply that information science has never addressed the notion of genre, or that genre has not been incorporated into any information-representation schemes. Indeed, several classification systems allow some articulation of genre, and many metadata standards, including the Dublin Core, include a field for genre. The treatment of genre is limited or not very well defined, however. Our understanding of the nature and role of document genre is still nascent, and so the use of this kind of information is underdeveloped in information-retrieval systems. Furthermore, it is not clear whether the extension of genre designations originally designed for physical collections will export well to digital ones.

Historically, most library information systems took genre for granted since most collections contained only a limited array of document types. The exceptions are literary genres (such as poetry) and publication types (such as almanacs or newspapers), which have had a lively existence in explicit document representation for several centuries. Aside from these, the primary facets of access to documents in traditional systems are descriptive components and subject, while genre is relatively rare. The descriptive access points derive from traditional ways of talking about books and book-like documents, and include: title, author, place and name of publisher, edition, date, series, physical description in terms of pages, size, volumes, and sometimes information about components, which are called *analytics*. The subject analysis of a document captures what a document is about—that is, its topic.

Librarians and information scientists have recognized that the topical approach is extremely important, but insufficient in some situations, and completely inappropriate in others. Not every document is necessarily about something. Sometimes the document's nature as a document represents the most important or useful aspect of it. For instance, on the one hand we can say that a book may be *about* symphonies—their history or structure—but what is Beethoven's Fifth *about*? It simply is. A symphony has a form and identifiable characteristics but it does not have a readily identifiable topic, *per se*, except that which can be attributed to it through subtle and non-consensual processes of interpretation. As the notion of *document* becomes broader and more diverse, as it does in the environment of the Web, we can see how the concept of subject does not stretch verywell to cover all types of information.

In response to the need to identify a document's form or genre in addition to its subject, librarians have created auxiliary tools in the form of tables and subdivisions to be used with existing topically based classification and subject-heading schedules. Here are a few examples:

The *Dewey Decimal Classification (DDC)* (Dewey, Mitchell, Beall, Matthews & New, 1996) provides several ways to denote a document's form or genre. The first is to incorporate a designation in the number itself. This is used in the 800s, which cover *belles lettres*. The first part of the number designates country/language, and the final digits represent the genre—1 for poetry, 2 for drama, 3 for fiction, 5 for speeches…7 for humor and satire, and so on:

English poetry 82**1**
English drama 82**2**
English fiction 82**3**
English speeches 82**5**
English humor and satire 82**7**
Bulgarian poetry 891.8**1**
Bulgarian drama 891.8**2**
Bulgarian fiction 891.8**3**
Bulgarian speeches 891.8**5**
Bulgarian humor and satire 891.8**7**

These genre designations are limited to the genres generally accepted by Western literary scholars, and do not necessarily do a good job of describing emerging, culturally diverse, or hybrid genres. Still, it is a way of privileging genre in the organization of literary works. It is interesting to note, however, that most public libraries do not make use of this formal system for fiction and arrange such works by author, with the *ad hoc* tradition of separating out popular genres into separate sections for easy access and browseability: Mysteries, Romances, Science Fiction, and so on.

Another technique in the *DDC* is to use suffixes from the Tables. The number for the topic is established, and then suffixes from Table 1 are added to denote the form or genre. For example:

Middle Eastern Cooking 641.5956
Middle Eastern Cooking Encyclopedia 641.5956+03
(dictionaries & encyclopedias)
Middle Eastern Cooking Magazine 641.5956+05
(serial publication)

In physical collections, the suffixes serve to distinguish materials on the same topic but in different publication formats one from the other. This notion of form/genre evolved from the physical distinctions of publication and document types, and thus is grounded in publishing practices and realities. The further interpretation of how such documents will be used remains implicit in the nature of the forms themselves, but has practical implications for collections. For instance, many dictionaries and encyclopedias comprise the non-circulating reference collection; magazines are indexed and stored differently than are books, and so on. In terms of digital collections, however, where the physical clues of publication format are largely absent, these suffixes might provide useful indicators for sorting and filtering search results.

Another way in which subject is indicated on the bibliographic record is through the use of subject terms from a thesaurus or list, such as the *Library of Congress Subject Headings (LCSH).* The *LCSH* comprises an evolving list of terms used by catalogers to assign subject designations to a work. Terms can denote topics, such as "sonnets," in which case this would be a work *about* sonnets, not the sonnets themselves. Proper names may also be subjects. For example, a document *about* William Shakespeare will be assigned Shakespeare's name as a subject, while a work *by* William Shakespeare would not. Modern cataloging practices abound in confusions about topic, creative responsibility and genre/form, since in many documents these three are inextricably fused. This confusion extends to searchers as well, who do not realize that searching for a genre using *LCSH* is problematic at best.

This distinction of reserving subject headings for topics/subjects only, is somewhat moderated by the addition of a subdivision. There are several kinds of subdivisions that can be used to "subdivide" a subject by time, geographical location, and further topical aspects. For example:

Witchcraft—Sweden
Witchcraft—15th Century
Witchcraft—Biblical teaching

The subdivisions of interest here, though, are the ones from the Form Subdivisions list. This type of subdivision allows the cataloger to further describe a work by its form or literary genre. This list is limited to several hundred well-established types. The genres included have literary warrant, since every subject heading and division in the *LCSH* was developed for an existing, rather than a hypothetical, work.

Witchcraft—Bibliography
Witchcraft—Case studies
Witchcraft—Dictionaries
Witchcraft—Handbooks, manuals, etc.
Witchcraft—Periodicals
Witchcraft—Poetry

The fact remains, however, that form and genre are not, as a rule, an important finding aid in traditional systems. For instance the work: *Final environmental impact statement for the Green Mountain and Finger Lakes National Forests land resource management plan* is assigned the following subject headings from the *LCSH*.

Forest reserves—Vermont Green Mountain National Forest
Forest reserves—New York (State)—Finger Lakes National Forest
Forest management—Vermont Green Mountain National Forest
Forest management—New York (State)—Finger Lakes National Forest
Green Mountain National Forest (Vt.)
Finger Lakes National Forest (N.Y.)

Thus, this work can be retrieved by either of the two national forests covered in the report and by two topics: *forest reserves* and *forest management*. It is not possible to retrieve this work as an *environmental impact statement* except for the coincidence that the terms appear in the title and would come up on a keyword search. There are many genres, such as this one, that serve a useful purpose as templates and are of interest in their own right, aside from the specific topic, but since this work is an example of an environmental impact statement, rather than about one, there is no subject heading assigned for this important aspect of the document.

Some libraries recognize that genre and form are often perceived as "topical" and have made some additional access points to accommodate this. For instance, the Rare Book, Manuscript, and Special Collections Library at Duke University (http://scriptorium.lib.duke.edu.genre-headings.html) has an interesting set of auxiliary tools for searching its collection. One of these is a Genre/Form list from which genre terms can be used for searching as if they were topics. Here is a sample of terms from that list:

Accounts
Business letters
Manuscripts
Official reports
Pattern books
Petitions
Recipes
Seals
Subliterary papyri
Tax returns
Vouchers

It is immediately obvious how very helpful such a list might be in studying the communicative forms of the cultures represented in the collection.

How to study genre

Having presented our case for understanding more about document genres in order to enhance retrieval of information from large digital collections, we turn now to the issues of precisely how we might study this phenomenon. Because genre is often an implicit and subtle notion, studying it in a systematic way presents many problems. Our overarching question is: would identification of document genres improve information access technologies in large digital collections such as digital libraries and the Web? This question cannot be answered directly, given the current understanding of genre or of genre's role in information retrieval. Thus, we envision a research agenda for investigating genre that proceeds through a series of componential studies, each of which we see as necessary for a full understanding. Thus in answering the central question with respect to genre it is necessary to investigate the following:

• The identification of Web document genres from the users' perspective and articulated in the users' own terms;
• The creation of a facetted (i.e., multidimensional) classification of these genres that can be used for controlled investigations in later stages of study;
• An investigation of how users integrate genre metadata into their own searching, evaluation, and use of documents;
• An evaluation of the degree to which incorporation of genre metadata in information-access systems makes a difference to the effectiveness of searching, sorting, ranking, and eventual use of documents; and
• An evaluation of various interfaces for visualizing and presenting genre metadata once it has been identified.

We also recognize that studying genre cannot be a once-and-for-all endeavor, since new genres are emerging all the time and old ones are being used in ways that are different than originally conceived. Thus, we propose that any study of genre must also establish a conceptual framework from within which to design continuing investigations. That is, we need a set of working hypotheses based on what we know about genre as a social construct. How are genres recognized? How do they evolve and change? How are they used and understood?

Studying genre from the users' perspective

We take it as a given that studies of genres must be based in real situations with real users. Since traditional designations of document genres will probably not adequately or accurately describe all the document types present and emerging on the Web and in digital libraries, it would make no sense to use such designations as a checklist against which digital genres are compared. Such a comparison would inevitably miss genres new or unique to the Web, or even more confusing, mistake traditional genres that have been adapted to new uses on the Web.

Thus, we see the first step as a descriptive phase of inductively extracting from what people say their terminology and sense of genres. At the same time we recognize that studying the *entire* range of possible document genres and the tasks for which identification might be useful is not realistic. Furthermore, genres can be specific to a particular discourse community, so too broad a scope may make it difficult to identify a useful set of genres in a manageable time with limited resources. Thus, as a first step we suggest that document genres be studied for a particular set of users, such as lawyers, educators, city planners, real-estate agents, and so on.

If the right population is chosen, limiting the scope does not necessarily mean that only a small subset of genres will emerge. For instance, teachers can potentially search for a wide range of topics and utilize documents of many genres, including a number that are particular to education, e.g., "lesson plans," and "academic standards." This diversity means there would be a wide range of potential document types to study. On the other hand, in the domain of education there is a core set of tasks that provides a base on which to study the impact of genre identification. Such tasks and situations for teachers include, for instance, writing a lesson plan, creating a reading list, adapting an existing class or developing tests and assignments. For administrators it might include conforming to educational standards, managing human resources, writing reports, and communicating with students and parents. Limiting the domain of inquiry will help focus a study, establish a reasonable scope, and provide a manageable set of situations with which to work, and on which to test the impact of genre identification.

We realize that limiting a study in this way also limits the ability to generalize, but for starters the aim is to show that genre identification is of value for certain tasks. Having demonstrated the basic concept for educators, for instance, we expect that it would be possible to then extend the principles beyond the domain of education.

Identification of genres

In order to design effective empirical studies to investigate people's use of genre, it is necessary to identify, describe and categorize the range of document genres used by the target population and the tasks associated with these documents. There is a substantial body of work on analyzing genre in printed documents and some work studying them on the Web (e.g., Bretan, Dewe, Hallberg & Wolkert, 1998; Crowston & Williams, 2000; Dillon & Gushrowski, 2000; Furuta & Marshall, 1996; Karlgren, Bretan, Dewe, Hallberg & Wolkert, 1998; Stamatatos, Fakotakis & Kokkinakis, 2000). However, these studies have typically been top-down that is; they analyzed a set of documents based on theoretical principles or according to *a priori* classifications. For example, Crowston and Williams (2000) based their classification on the *Art and Architecture Thesaurus* (Petersen, 1994) and a number of studies used the categories of the Brown Corpus.

A top-down approach to genre is 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, we assume that a traditional typology of genre or document forms will not be sufficient to describe the emerging and dynamic genres identifiable by users in general and our community in particular.

Some researchers have attempted to identify genres bottom-up through relatively small-scale user studies (e.g., Dewe, Karlgren & Bretan, 1998; Nilan, Pomerantz & Paling, 2001). However, we do not as yet have a fully articulated set of data that reveals what genres people recognizes nor for what tasks they find documents of specific genres useful.

In investigating the range of genres identified by users, we suggest that the following questions should be addressed:

- How do people talk about the genre of documents?
- Does the naming and identification of digital-document genres correspond to the naming of traditional non-digital genres?
- How do people understand and make use of new, unnamed, emerging, and "colonized" genres (Beghtol, 2000) in digital collections?
- What clues do people use to identify genre when engaged in information-access activities?
- What facets (basic attributes) of genre do people perceive?

Creation of a facetted (i.e., multidimensional) classification of genres

If genre is to be used as another facet of description for documents and queries, we are still left with the issue of how to describe genre itself in such a way that it can be implemented in a system. Genre itself is a multidimensional phenomenon, incorporating form, function, and the numerous clues and components that allow us to discriminate one genre from another. Towards this end we see the need to create a rich and flexible description of document genres that will do justice to their complexity while at the same time providing a structured tool for systematic inquiry. One way to achieve this is through a facetted classification.

A classification will help determine the level of granularity that can be achieved in genre identification. Genre complexity can be managed by organizing the genres in a classification from more general to more specific. By picking appropriate levels of specificity it might be possible to avoid having to identify hundreds of detailed genres, while still providing a basic level of distinction in areas of particular interest.

Most organized lists of genres are structured as hierarchies. The criticism of traditional hierarchies is that they rely on a single organizing principle, which may not be useful for all cases. To overcome this problem we suggest using the facetted-analysis approach. Facetted analysis identifies *multiple* fundamental dimensions along which objects, such as genres, can be described and then clustered. For example a genre such as a "lesson plan" can be identified by

its source, its purpose, its structural features, and so on. Each facet, or basic dimension, can be articulated following its own logic and subsequently can be used for its own type of clues for classification. In suggesting the use of facetted analysis we follow the example of previous genre-identification studies such as Päivärinta (1999), Tyrväinen and Päivärinta (1999) and Karjalainen, Päivärinta, Tyrväinen and Rajala (2000) who looked at the management of enterprise documents, and Kessler, Nunberg and Schuetze (1997) who sought to identify a limited set of facets for communicative purposes.

A facetted approach to classification is pragmatic and not dependent on any one conceptual perspective. It allows for the development of description and clustering using a number of fundamental dimensions, rather than just one. The results of this process yield a classification that is flexible, expressive and hospitable to new genres and genre combinations. It also allows a view of genres at a variety of conceptual levels, from the general and inclusive to the very specific, which will be useful in simulations in later phases of inquiry.

How would a facetted approach to genres work?

In principle, this approach requires several passes. The first pass identifies and labels facets that seem to be important. These might include form, content, style, implied use, and the relationship of that document to others. These facets serve as starting points, and new facets may emerge. After identifying the basic facets, one must again review the entire corpus repeatedly to see the range of categories on which these facets are revealed—for instance, what do people use to describe "source"? The process continues until saturation is reached (i.e., no new categories emerge). If necessary, more data is collected.

Once the Web genres are identified, it might be interesting to compare them to the more traditional sources of genres for overlaps of structure and coverage. We expect that there will be a significant amount of redundancy among the genres identified in this way. The aim is to generate a classification that reflects not only currently identified genres, but will also flexibly accommodate identification of emerging and future genres (Beghtol, 2000; Kwasnik, 1999), thus providing a basis for future work in this area.

How users integrate genre metadata into their own searching, evaluation, and use of documents

Besides identifying genres and their attributes, any study of genre effectiveness must also establish how people in fact utilize genre in searching, in generating queries, and in their evaluation of documents. To accomplish this, further observations are necessary in order to answer the following questions:

• In what contexts and for what tasks is the identification of genre useful?
• To what extent are documents of various genres specific to certain tasks as opposed to being generally useful?
• To what extent are people interested in documents of genres specific to their domain and environment *vs.* those used more widely?

In summary, the results of the necessary initial phases of studying genres on the Web would be a better understanding of how users in the community of interest describe the genre of documents and how they use genre when they work with documents to solve information problems. This phase should also provide a database of documents categorized according to their user-specified genre, genre features, and user evaluations and related information tasks. Such a database can be used as the baseline for simulations and evaluation studies in subsequent phases. It would also provide an inductively derived facetted classification scheme for usefully clustering genres and features and for determining granularity.

Evaluating the effectiveness of genre identification

Using the basic information discovered in the preliminary phases, it should then be possible to carry out controlled user studies whereby various aspects of genre use can be manipulated to see the differences in retrieval effectiveness associated with each manipulation. Thus, in this phase one can study how genre metadata can best be utilized in information-access tasks. By "best" we mean initially improving users' performance (e.g., time, accuracy or perceived usability) in information searching, filtering and evaluation tasks. Ultimately, of course, best means improving the performance in the kinds of information tasks people face in their day-to-day work lives.

This phase in the research plan must address several questions:

• How best to use genre metadata in information-access systems?
• To what extent does providing genre metadata improve performance and utility?
• 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?
• Can genre metadata be used to inform other aspects of the process of searching, summarizing or evaluating documents?

To answer these questions, one approach would be to develop simulated information-access system interfaces that incorporate genre metadata and then to conduct evaluations of the efficacy of these interfaces in controlled laboratory experiments. Depending on resources and available design expertise, one could implement these prototypes in various ways, starting with storyboards and paper mockups, scripted interfaces and if possible by implementing them on a real search engine. The following are some suggestions for scenarios in which genre information could be implemented in order to test its efficacy:

• Provide aids for query construction using genre metadata;
• Cluster documents based on genre metadata (explicitly labelled, vs. labelled using other techniques vs. unlabelled); or
• Show genre information in combination with other metadata.
• Raise or lower a document's rank base on genre metadata;
• Incorporate genre-specific information-access processes.

The method of presentation of genre information will inevitably influence its efficacy. For this reason, before any retrieval evaluation can take place, this aspect of genre must also be investigated. Specifically studies should investigate:

• How best to represent and display genre metadata to the user and receive feedback?
• What level of granularity of genre metadata improves performance most (i.e., how specific should the description of genre be)? Is there, perhaps, a basic level of genre that is neither too general and abstract, nor too specific?
• To what extent does combining genre metadata with other kinds of metadata (e.g., subject) improve performance? For example, is it more useful to identify documents as a "5th-grade science lesson plans" or just a "lesson plans"?
• To what extent is the user's performance degraded by miscategorization of documents based on genre?
In answering these questions there are some general interface design issues that must also be addressed. For example, it is important to decide on:
• The choice between *opaque* and *transparent* modes of presentation (Roussinov et al., 2001). In *transparent mode*, the system will expose its identification of genre to the user by labelling or otherwise identifying it. In opaque mode, results will make use of genre metadata, e.g., for ranking, but the user will not be made explicitly aware of it;
• If the transparent mode is used, then one must make a choice between interfaces that explicitly name genres and those that use other methods of labelling (e.g., by providing example documents).

A typical experiment might for instance contrast two interfaces. Participants could use one interface in the first half and a different interface in the second half of the experiment, with the order of presentation counterbalanced across subjects. An example task might be to find a relevant Web page for a given problem scenario.

Answering the big question: Does genre help?

Identifying Web document genres, observing how genre is used in searching and evaluation of search results, the multifacetted representation of genres, and finally the design of presentation and visualization techniques allows a systematic exploration of the overall effectiveness and utility of genre information for Web documents.

We know that people use genre information and that for many applications it is perhaps the one most important piece of information that can be provided. Nevertheless, the extent to which the general inclusion of genre information will enhance information access on the Web is an open question. There is much to understand before such information could be practically implemented, and so it is perhaps wise to answer our question before embarking on time-consuming and expensive applications. On the other hand, if, as we suspect, genre information is

helpful, then studying it in a systematic way, as we propose above, can provide the initial baseline understanding that must precede any automatic implementation.

References

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

Beghtol, C. (2000). The concept of genre and its characteristics. *Bulletin of the American Society for Information Science & Technology, 26*(5).

Bretan, I., Dewe, J., Hallberg, A., & Wolkert, N. (1998). *Web-Specific genre visualization.* Paper presented at the WebNet '98, Orlando.

Crowston, K., & Williams, M. (2000). Reproduced and emergent genres of communication on the World-Wide Web. *The Information Society, 16*(3), 201- 216.

Dewe, J., Karlgren, J., & Bretan, I. (1998). *Assembling a Balanced Corpus from the Internet.* Paper presented at the The 11th Nordic Computational Linguistics Conference, Copenhagen, Denmark.

Dewey, M. A., Mitchell, J. S. E., Beall, J. E., Matthews, W. E. J. E., & New, G. R. E. (Eds.). (1996). *Dewey Decimal Classification and Relative Index Set* (21th ed.). Dublin: OCLC Forest Press.

Dillon, A., & 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, 5*(12), 202-205.

Furuta, R., & Marshall, C. C. (1996). *Genre as Reflection of Technology in the World- Wide Web* (Technical Report): Hypermedia Research Lab, Texas A&M.

Karjalainen, A., Päivärinta, T., Tyrväinen, P., & Rajala, J. (2000). Genre-based metadata for enterprise document management. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*. Los Alamos, CA: IEEE Computer Society Press.

Karlgren, J., Bretan, I., Dewe, J., Hallberg, A., & Wolkert, N. (1998). *Iterative information retrieval using fast clustering and usage-specific genres.* Paper presented at the Eighth DELOS Workshop: User Interface in Digital Libraries, Stockholm, Sweden.

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

Kwasnik, B. H. and Liu, X. "Classification Structures in the Changing Environment of Active Commercial Websites: The Case of eBay.com". In: Beghtol, Clare,

Howarth Lynne C., and Williamson, Nancy J. (Eds.) *Dynamism and Stability in Knowledge Organization. Proceedings of the Sixth International ISKO Conference*. 10-13 July, 2000. Toronto, Canada. (*Advances in Knowledge Organization*, Vol. 7, 2000: 372-377).

Kwasnik, B. H. (1999). The role of classification in knowledge representation and discovery. *Library Trends, 48*(1), 22-47.

Kwasnik, B. H. (2002). Commercial Websites and the Use of Classification Schemes: The Case of amazon.com. In M. J. Lopez-Huertas (Ed.), *Challenges in Knowledge Representation and Organization for the 21st Century: Integration of Knowledge across Boundaries. Proceedings of the Seventh International ISKO Conference* (pp. 279-285): Ergon Verlag.

Kwasnik, B. H., Crowston, K., Nilan, M., & Roussinov, D. (2000). Identifying document genre to improve web search effectiveness. *The Bulletin of the American Society for Information Science and Technology, 27*(2), 23-26.

Longacre, R. (1983). *The Grammar of Discourse*. New York: Plenum Press.

Marcu, D. (1997). *From discourse structures to text summaries.* Paper presented at the 14th National Conference on Artificial Intelligence (AAAI-97).

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech, 70*, 151-167.

Nilan, M. S., Pomerantz, J., & Paling, S. (2001). Genres from the Bottom Up: What Has the Web Brought Us? In T. B. Hahn (Ed.), *Information in a Networked World: Harnessing the Flow. Proceedings of the ASIST 2001 Annual Meeting*. Washington, DC.

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

Päivärinta, T. (1999). A genre approach to applying critical social theory to information systems development. In C. H. J. Gilson, I. Grugulis & H. Willmott (Eds.), *Proceedings of the 1st Critical Management Studies Conference, Information Technology and Critical Theory stream*. Manchester, England.

Petersen, T. (1994). *Art and Architecture Thesaurus*. New York: Oxford.

Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Liu, X., & Cai, J. (2001). *Genrebased navigation on the Web.* Paper presented at the Thirty-Four Hawaii

International Conference on Systems Science (HICSS-34), Maui, HI..Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics, 26*(4), 471-498.
Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings*. New York: Cambridge University Press.
Tyrväinen, P., & Päivärinta, T. (1999). On rethinking organizational document genres for electronic document management. In *Proceedings of the 32nd Annual Hawaii*
International Conference on System Sciences. Los Alamitos, CA: IEEE Computer Society Press.
Yates, J., & Orlikowski, W. J. (1992). Genres of organizational communication: A structurational approach to studying communications and media. *Academy of Management Review, 17*(2), 299-326.
Yates, S. J., & 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.