A Framework for Creating a Facetted Classification for Genres: Addressing Issues of Multidimensionality

Kevin Crowston and Barbara H. Kwasnik
Syracuse University School of Information Studies
crowston@syr.edu, bkwasnik@syr.edu

Abstract

People recognize and use document genres as a way of identifying useful information and of participating in mutually understood communicative acts. Crowston and Kwasnik [1] discuss the possibility of improving information access in large digital collections through the identification and use of document genre metadata. They draw on the definition of genre proposed by Orlikowski and Yates [3], who describe genre as "a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form" (p. 543). Scholars in fields such as rhetoric and library science have attempted to describe and systematize the notion of genre, and have offered many different definitions of genre. We like Orlikowski and Yates's definition because it takes into account all three aspects of genre that we recognize as fundamental: content, form, and purpose.

A document's genre is a subtle and complex concept in which the content and form of a document are fused with its purpose or function. As such, a document's genre cannot be separated from the context in which it is used; the same document may be construed as being of a different genre depending on how it is invoked in a given situation. Starting from the document, a letter may be a personal communication, or a piece of evidence in a court of law, or an agreement, or even a work of art. Starting from the situation, we note that differences in an information situation are often reflected in the kind of document that is considered helpful (e.g., a problem set vs. a lesson plan vs. a tutorial about mathematics, for instance). Thus, we see genre as a multidimensional phenomenon, which takes into account not only the attributes of the document itself, but also of its role in human endeavor. In this paper, we discuss some considerations in developing a facetted classification for genres to address the problem of multi-dimensionality.

1. Genre within an information retrieval framework

We begin by considering the role of genre identification as part of the larger process of information retrieval (IR). Access to information has been the subject of a very extensive body of research for many decades, but the advent of the Web has intensified the necessity of better methods for searching the vast stores of information that have become more easily accessible. Progress in this field is difficult because human information seeking is a complex and variable process. Nevertheless, the framework within which such research has taken place is useful in our study because it succinctly identifies the various components of information access and allows us to pinpoint where the identification of genre might be most useful.

In its simplest articulation, we can view the information-retrieval process as follows:

> A user represents an information need by submitting a query to the system *via* an intermediating mechanism. The system searches through the document representations in its store and uses some form of matching to "retrieve" either the documents themselves, parts of the documents, or representations of the documents. These search results are then presented to the user for evaluation.

The aim of this process is to retrieve all the relevant and useful documents, to avoid retrieving those that are not relevant or useful, and to present the results in such a way that the searcher can make use of them.

There are countless variations on this basic process, but we know that even under the best of circumstances it is rarely, if ever, one-hundred percent effective or efficient. Matching users' needs to potential information in the system is complicated by many factors, but the following are the most pertinent to the present discussion:

- Users may be unable to formulate a query that represents the information need well, or in a way that the system can recognize. Even if they can articulate a query, the way in which humans express information needs produces a great deal of linguistic variety. Furthermore, we know that people often 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.

- The system's representations of documents may be incomplete, inappropriate to the search, or inaccurate, resulting in poor results. 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.
- The results may be very noisy and imprecise—that is, the system returns correct/useful results, but also many incorrect ones as well.
- Or, conversely, the results may be misleadingly sparse—implying that the system is not able to satisfy the information need, even if in fact documents matching the need do exist. Put another way, many relevant or useful documents may never be retrieved.
- The results, while accurate, may be presented in such as way that the task of processing them by the user is too difficult or time-consuming. This is especially true of systems that do not rank results or when results are imprecisely represented and the user must wade through a great deal of undifferentiated information. For example, a system may present a large list of possibly relevant documents but without indicating where in the documents the relevant information can be found.
- Finally, the system may be able to perform simple matches, but be unable to provide the capability of expanding, exploring, or otherwise interacting with the system further.

These problems all fall under the rubric of **representation.** The query must be appropriately represented; the system must have adequate internal representations of its information in order to retrieve it precisely and thoroughly; results must be represented in such a way that actually making use of them is manageable and satisfying; and representations must provide fruitful connections and navigational cues to enable users to discover or explore information via browsing.

Traditionally, information scientists and librarians have relied on "topic" (or more simply "keyword") to provide the representation of both the query and the data store. We know, however, 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.

1.1. Why we think identification of genre would be useful

We hypothesize that enhancing document representations by incorporating non-topical characteristics of the documents that signal their purpose—specifically, their genre—would enrich document (and query) representations. By incorporating genre we believe we can ameliorate several of the information-access problems described above and thereby improve all stages of the IR process: the articulation of a query, the matching or intermediation process, and the filtering or ranking of results to present documents that better represent not only the topic but also the intended purpose.

A query might be enriched by including information about expected genres of the results (either initially or as part of the relevance feedback). 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.

Knowledge of the form of genres can help in the matching process. For example, FAQs documents are divided into question and answer pairs. If we require search terms to be found in the same question-answer pair, we may reduce spurious matches.

Knowledge of document genre may improve accuracy of relevance judgments that modern search engines make in order to rank order the search results. It has been noted that some genres are less likely to be relevant for the majority of search tasks. This implies that certain Web pages might be promoted or demoted in the ranked order if their genre were known. For example, it has been noted that most searchers are not interested in getting personal home pages [4], so the latter could be moved down the list by request.

Finally, 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) [5].

2. Representing genres

We have suggested the advantage of incorporating genre information into query and document representations, but how do we represent the genres themselves? We propose that a possible representation might be a facetted classification, but before describing what this is or how it might be done, we first consider why we need a classification of genre at all rather than a simple list of genre terms.

- First, a classification is a *consensual lens* through which to view a given set of entities, such as the various genres, so it can serve as a way of pulling together disparate views, terminology, and scope. It establishes the range of the phenomenon being described, and it allows for communication about it in a standardized way. If genre information is to be incorporated into systems as document and query representations, then there must be a mechanism for doing so that is not totally ad hoc and impossibly variable.
- Second, a classification allows for *systematic conceptual manipulation*. For example, if a classification is structured as a hierarchy, with the most inclusive terms at the top and the most specific terms at the bottom, we can refine the specificity of a search and deal with genre complexity better. Do I search for letters (specific), or for correspondence (more general), or for love letters (even more specific than letters)? A hierarchical representation allows a user to easily move between these queries. As an added benefit, identification of the appropriate scale might help avoid having to identify hundreds of detailed genres, while still providing a basic level of distinction in areas of particular interest.
- Third, a classification that is thorough, conceptually sound and grounded in observation of real phenomena allows researchers to *identify gaps and missing items*. Consider the role of the Periodic Table of Elements in the discovery of new elements.
- Finally, classifications enable *clustering*. It is what makes it possible to request "more like this." It also makes it possible to browse, which is a type of navigation without a predetermined goal. Browsing is a good way of expanding or narrowing searches by identifying close neighbors, learning what the system has to offer, learning about the relationship of one thing to another, and generally being able to search and explore without specifying exactly what is required. Browsing is not possible (or at least not much fun) without some underlying organization to the information so that the user can navigate from one node to another along some definable paths.

There are many issues to consider in creating any classification, however, let alone one for so complex a concept as genre. These issues include determining the scope and extent of the domain being classified and the entities themselves—their scale (granularity) and the terminology used to describe them. Once these are established, a conceptual structure must be determined, since a classification is not merely a "loose bag of concepts" but rather, a collection of such concepts that are related to each other through classificatory relationships. One example of such a relationship is the genus/species relationship in a hierarchy. The conceptual structure of a classification is often determined by how theory or practice determines that the entities "go" with each other. Atomic theory guides the Periodic Table's structure, while theological beliefs guide the organization of the Choirs of Angels.

3. Creating a classification scheme for genres

The first practical issue in building a classification scheme is to determine the nature of the entities being classified. Put simply, this means determining what are the "things" that are being classified—in our case, genres. One can think of this step as concept harvesting. This means establishing a body of entities that when organized into a classificatory structure would clearly, completely and truly describe the phenomenon of "genre"—or at least do so in a way that would enable incorporation of genre metadata into information-access mechanisms.

A related task is to determine the unit. Many genres (such as a newsletter, for instance) can be viewed as composites of several genres (articles, editorials, calendars of events) and can be distinguished by both their components as well as by the unique assembly of components into an identifiable whole. From a classification point of view, this means establishing a scale for the scheme. How finely grained does the identification of genres (and their possible components) have to be? Conversely, how do we know when we have reached the boundaries of any given genre? When does a memo turn into a report or an abstract into a review?

There are basically two approaches to the task of genre identification: top down, and bottom up. In the top-down approach, one would gather genre names and their associated attributes from existing sources or from existing theoretical models (such as those in textual studies, librarianship, or rhetoric). There is a substantial body of work on analyzing genre in printed documents and some work studying them on the Web [e.g., 2, 6, 7-10]. These studies analyzed a set of documents based on theoretical principles or according to *a priori*

classifications. For example, Crowston & Williams [2] based their classification on the *Art and Architecture Thesaurus* [11] and a number of studies used the categories of the Brown Corpus.

A top-down approach to genre is problematic, though, 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. 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"). Researchers once thought of genres as rather static and familiar. We grew up learning what a letter was, what a bill of sale was, or a recipe. But, as pointed out by Dillon and Gushrowski [7, 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.

For this reason, we suggest that the bottom-up approach might be more valid in the case of an implementable scheme for genre. It is important to capture the users' own language and understanding of genres because if such information is to be incorporated into the retrieval process it must resonate with how genres are actually recognized and named. A few researchers have attempted to identify genres bottom-up through relatively small-scale user studies [e.g., 12, 13]. However, we do not as yet have a fully articulated set of data that reveals what genres people recognize nor for what tasks they find documents of specific genres useful.

So, as a first step in creating a classification of genres we suggest that, at a minimum, the following questions should be addressed:

- How do people talk about the genre of documents?
- How do people understand and make use of new, unnamed, emerging, and "colonized" genres [14] 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?

Once genres and their attributes have been identified, one can proceed to the next step, which is the organization of these entities into a conceptual structure.

3.1. Creating a facetted classification of genres.

Most organized lists of genres are structured as single hierarchies. For example, Figure 1 shows a small section of the hierarchy of genres of Web documents identified by Crowston and Williams [2]. Advertisements and announcements are both examples of declaratory document genres; classified advertisements are a special kind of advertisements, and so on.

The criticism of traditional hierarchies is that they rely on a single organizing principle, which may not be useful or appropriate for all cases. To overcome this problem we suggest using the facetted-analysis approach [15]. In suggesting the use of facetted analysis we follow the example of previous genre-identification studies such as Päivärinta [16], Tyrväinen and Päivärinta [17] and Karjalainen et al. [18] who looked at the management of enterprise documents, and Kessler, Nunberg and Schuetze [19] who sought to identify a limited set of facets for communicative purposes.

Facetted classifications are not really a different representational structure, but rather a different approach to the classification process. The notion of facets rests on the assumption that there is more than one way to view the world, and that even those classifications that are viewed as stable are in fact provisional and dynamic. The challenge is to build classifications that are flexible and can accommodate new phenomena. In the case of genres, a facetted classification is particularly appropriate because

<declaratory document genres>	
advertisements	
classified advertisements	Short paid announcements appearing in a periodical sorted according to the good or service being offered or requested
announcements	Printed or published statements or notices that inform the reader of an event or other news
custom 404 page	A Web page announcing that the requested Web page could not be found on the server
news bulletins	
press releases	Official or authoritative statements giving information for publication in newspapers or periodicals

Figure 1. A section of a hierarchy of document genres [from 2].

Period/Style	Place	Process	Material	Object
19th Century	Japanese	Raku	ceramic	vase
Arts & Crafts	American	Carved	oak	desk

Figure 2. A facetted analysis of artifacts [from 21].

we know that genres are not only complex, fusing content, form, and purpose, but they are also dynamic—new ones emerge, old ones morph into new ones.

Facetted classification has its roots in the works of S.R. Ranganathan, an Indian scholar, who posited that any complex entity could be viewed from a number of perspectives or *facets*. He suggested the fundamental categories of Personality, Matter, Energy, Space and Time [20]. Over the years, Ranganathan's facets have been reinterpreted in many contexts; they have been used to classify objects as disparate as computer software (for reuse), patents, books, and art objects [15].

Not all modern facetted classifications use Ranganathan's prescribed fundamental categories, but what they do have in common is the process of analysis. Consider the example in Figure 2 [from 21]. Figure 2 shows a possible solution to the classification and description of two objects of material culture, which in its diversity defies easy description and categorization. For purposes of demonstration this is a simplified version of the one used by the *Art and Architecture Thesaurus*. For any given artifact, there are many possible ways of representing it, let alone the "knowledge" that enabled its production or its value. The facetted approach follows the following steps:

- ***Choose facets***. One must decide on the important criteria for description. In principle, this approach requires several passes. The first pass identifies and labels facets that seem to be important. In the example we have Period, Place, Process, Material, and Object, following closely on what Ranganathan suggested, but for genres we might include form, content, source, style, implied use, and the relationship of that document to others. These basic facets would emerge from the user studies in which we observed how people name and differentiate genres, and would serve as starting points. 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"? If necessary, more data is collected and the analysis process repeated until saturation is reached (i.e., no new categories emerge).
- *Develop facets.* Once the fundamental categories of description have been identified, then each facet can be developed/expanded using its own logic and warrant and its own classificatory structure. In the example, the Period facet can be developed as a timeline; the Materials facet can be a hierarchy; the Place facet a part/whole tree, and so on. This is one of the strongest attributes of facetted classifications because it does not lock the designer in to one logical scheme. Since we know that genres are multidimensional, we can also assume that the dimensions will be quite different in kind one from the other. That is, building a sub-scheme for genre style might follow a different logic than developing one for genre source.
- *Analyze entities using the facets.* In analyzing an entity, one chooses descriptors from the appropriate facets to form a string, as shown above. Thus, the classification string for object 1 in Figure 1 is "19th Century Japanese raku ceramic vase" and the string for object 2 is "Arts & Crafts American carved oak desk." It is important to note that the process is not one of division (as in a hierarchy) where the entities are subdivided into ever more specifically differentiated categories. It is not a process of decomposition, either (as in a part/whole tree), in which the entities are broken down into component parts, each part different from the whole. Instead, the process of facet analysis is to view the object from all its angles—same object, but seen from different perspectives. So, in the example, the vase can be seen from the point of view of its period, the place in which it was made, the material and processes, and so on. A genre could be viewed from the perspective of its purpose, content, and form.

It should be noted that facet analysis is an ongoing process, and once the basic facets have been identified, the actual values within the facet can be adjusted as new knowledge emerges.

3.2. Extending the notion of facets to the description of genres: An example

Let us say that a person is searching the Web for documents dealing with botox treatments. Many "hits" are retrieved and the person must now start the process of distinguishing one type of document from another. By way of example, let's assume that the search yields the following:

- A scholarly journal article,
- A popular magazine article,
- A personal testimonial,
- A chat group,

Type of Document	What Users Invoke as Clues to Identification	Possible Facet
Scholarly paper	.edu in url	Source
	presence of journal name, volume, number	Source
	presence of abstract	Structure
	statistics, tables and figures in the text	Content (presence)
	particular style of photos (anonymous closeups)	Graphics
	scholarly language	Language level
	references	Structure/Content
	more than 5 pages long	Length
	formal unadorned layout	Structure/Layout
Popular magazine article	Artistic layout	Structure/Layout
	Everyday language	Language level
	Photos show actual human beings	Graphics
	No references	Content (absence)
	Short paragraphs with many headings	Structure
Chat	Sequence of short entries	Length/ Structure
	Presence of "tags" (People's nicknames)	Content (presence)
	"Chat" style language – incomplete sentences, colloquial expressions, chat abbreviations	Language level
	Reverse chronological dated entries	Content (presence)
	Subject lines	Content/Structure

Table 1. Possible clues to identifying a document's genre and facets represented.

- The website for a group medical practice specializing in cosmetic surgery,
- A pop-up advertisement for "lowest-cost botox treatments," and
- An short excerpt from a women's health newsletter that requires a subscription to get the full text.

Taking each in turn, we note what people tell us are the distinguishing features that allow them to tell one type of document from another. Table 1 shows a sampler of what such descriptions might comprise for three example genres.

Having collected an inventory of clues, such as the ones in Table 1 (and we anticipate that the lists would, in fact, be much longer), we could then proceed to building a set of facets or basic dimensions along which people make such descriptions. In the table we suggest preliminary facets dealing with content, structure, language, source, and so on, but there are probably others as well that would emerge as more and more genres were studied. Having a set of clues and facets, we could then proceeds to developing the particular classification scheme for the individual facets. A well-grounded set of such facets would allow a more complex and flexible approach to representing genres—one that could build a profile of a genre that includes form and communicative purpose.

3.3. Why is a facetted classification appropriate for genres?

As mentioned above, genre is a subtle and difficult-to-define notion. One of the most challenging obstacles to studying it is that we have no way of knowing when a complete set has been captured or whether we have tapped all the possible nuances of purpose and form. Without a strong foundational theory of genre to guide us, it is also problematic to set up a classification structure that will accommodate all genres, all purposes, all forms. Under these circumstances it is difficult, if not impossible, to build a single, unified Periodic Table of Genres, so to speak. Thus, a facetted classification is a useful tool because of the following characteristics:

- *It does not require complete knowledge.* In building a facetted scheme it is not necessary to know either the full extent of the entities to be accommodated by the scheme, or the full extent of the relationships among the facets. It is thus particularly useful in ill-defined domains, or domains that are apt to change.
- *It is relatively hospitable.* When a classification is hospitable it means it is capable of accommodating new entities smoothly. In a facetted scheme, if the fundamental categories are sound, new entities can be described and added. This is particularly important in the classification of genres, where we have no way of predicting the emerging genres that will be produced by the human imagination and the evolving nature of human endeavor in which the genres are invoked. If a genre recognized 100 years from now could be described by the fundamental categories of a facetted scheme, then that scheme will still be robust.
- *Facetted schemes have flexibility.* Since a facetted scheme describes each object by a number of independent attributes, these attributes can be invoked in an endlessly flexible way, in a sort of Lego

approach. "Let me see all the kinds of homepages for not-for-profits." "OK, now all newsletters." "OK, now newsletter for for-profits…" This flexibility can be used to discover new and interesting associations. The approach is called post-coordination, and means that attributes can be mixed and matched at the time of retrieval. It is in contrast to the pre-coordinated categories that are a requirement of most hierarchies, in which the rules for class inclusion are invoked at the time the entity is classified and stay fixed from there on. Put another way, categories can be produced on the fly without having to know in advance that the attributes will be put together in a fixed profile. At the same time, fixed profiles can be created if needed.

- *It allows for requisite expressiveness.* A facetted approach can be more expressive because each facet is free to incorporate the vocabulary and structure that best suits the type of knowledge represented by that facet. Thus, the designer has the freedom to build a structure that is as detailed or general as is necessary **for each facet**, rather than for the classification scheme as a whole. Since it isn't possible to describe every genre for every single purpose, some selectivity as to the level of description will be necessary. A facetted classification allows some facets to have more specificity, as required, without over-specifying where it is not useful to do so.
- *It does not **require** a strong theory.* In a facetted classification it is the individual facets that have classificatory structures, while the overall scheme may or may not have such a structure. For this reason, the overall facetted scheme does not have to have a "theoretical glue" to hold it all together and to guide the rules for association and distinction. It can be constructed more pragmatically, so long as the fundamental categories function well as pigeonholes for the main concepts. So, if we do not understand, for example, how the form of a genre is related to its purpose, we do not have to include information about that relationship in the scheme as we must do in a phylogenetic tree, for instance. There is a facet for *form*, and another for *purpose*, and we can associate them if we wish, but the viability of the entire scheme is not dependent on this.

  Having said this, facetted schemes can be instrumental in building theoretical understanding because they provide a mechanism for analysis, and subsequently synthesis, by presenting the dimensions in an organized and exhaustive way, but not in a way that is predetermined and therefore rigid.
- *It can accommodate a variety of theoretical structures and models.* A facetted approach makes it possible to represent a variety of conceptual frameworks because each facet can derive from a distinct body of thought. The study of genre draws from many disparate disciplines, which could not easily be accommodated under the umbrella of a single classificatory scheme. A facetted classification could allow for one facet to draw on the field of Communication to describe any given genre as a type of "communication act," for instance, while another to draw on the field of Education for the notion of "reading level."
- *Multiple perspectives.* One of the most useful features of a facetted approach is that it allows entities to be viewed from a variety of perspectives—a feature that is lacking in unitary classification structures. In a facetted analysis it is possible to describe a dog as an animal, as a pet, as food, as a commodity, and ad infinitum, so long as the fundamental categories have been established with which to do this.

3.4. What are some of the obstacles to creating a facetted scheme?

While the flexibility and pragmatic appeal of facetted classifications have made this a good candidate for genre classification, there are, nevertheless, some limitations:

- *Difficulty of establishing appropriate facets.* The strength of a facetted classification lies in the fundamental categories, which should be able to express all of the important attributes of the entities being classified. Without knowledge of the domain and of the potential users, this is often difficult to do. While it is possible to flexibly add entities, it is not a simple matter to add fundamental facets once the general classification is established. In the case of classifying genres, this is further complicated by the fact that people may not be aware of what allows them to recognize a given genre, and thus the determination of an adequate set of fundamental categories will be a challenge.
- *Lack of relationships among facets.* Most facetted classifications do not do a good job of connecting the various facets to each other in any meaningful way. Each facet functions as a separate kingdom, as it were, without much guidance as to how to put the parts together. For example, if we were to facet analyze motion pictures by genre, country, director, film process, and so on, we would still have no insight as to the meaningful relationships of, say, a particular country and the popular film genre there, or of a particular film process and the genres it supports. In terms of theorizing and model building, the facetted classification serves as a useful and multidimensional description, but does not explicitly connect this description in an explanatory framework. In the case of applying genre information in systems, this limitation is probably less important because we merely need to know whether a given dimension is important or not. However, it would be helpful to understand how the

facets function interdependently so that if it is easier to identify cues for a genre along one facet than another, it might make the implementation process more efficient.

- *Difficulty of visualization.* Other classification structures, such as clusters or hierarchies, can be visually displayed in such a way that the entities and their relationships are made evident. This is difficult to do for a facetted classification, especially if each facet is structured using a different internal logic. As a result, facetted schemes can only be viewed along one or two dimensions at a time, even though a more complex representation is actually incorporated into the descriptive strings. Thus it is difficult to see a vase in the context of other vases, of other Japanese artifacts, of other clay objects, of other *raku* objects and so on, all at the same time. Since we envision genre-enhanced retrieval results to be one of the ways in which genre recognition may help, the problem of visualizing a facetted scheme would have to be addressed.

Nevertheless, the facetted approach is useful because we recognize that it allows at least some systematic way of viewing the phenomenon without the necessity for a mature and stable framework from within which to view it.

3.5. Other considerations for identifying and classifying genres

So far we have described the basic and general process of approaching a facetted classification of genres, but of necessity we have limited the discussion to a representation of genres that is meaningful to human beings using them for the purpose of refining queries, enhancing searching, or interpreting results. There is another aspect of genre representation, though, that might not as easily fall into a semantic classification approach, such as the one described above. This is the problem of distinguishing between what cues a human needs to distinguish one genre from another [22], and what a machine might need to do the same thing. For instance, in some situations a human might find the form of a genre sufficient to identify it (such as a formal letter with a return address, a salutation, body and closing), but might require something else in addition to form in some other situation (such as a recipe).

A machine, on the other hand might do better with purely structural cues such as sentence length, presence or absence of certain punctuation and spacing, and so forth. Furthermore, in applications such as machine learning, it may not be necessary for the designers to even know what criteria a human finds to be useful cues. In this case, would a facetted classification of machine-friendly dimensions be useful or, indeed, possible?

We anticipate that humans and machines overlap considerably in the cues they use for recognition, even if they are not isomorphic. In any event, we would still need to know what it is users need to have presented to them in order to recognize a given genre, and for this a facetted scheme will provide a rich and complex description that can then be used in a variety of representational tasks.

4. Conclusion

A facetted approach to classifying genres is pragmatic and not dependent on any one conceptual perspective. It permits the designer to draw on a number of existing sources and models in creating a multidimensional description. It allows for the development of several associative structures using a number of fundamental dimensions, rather than just one. The results of this process would yield a classification that is flexible, expressive and hospitable to new genres and genre combinations. It would also allow a view of genres at a variety of conceptual levels, from the general and inclusive to the very specific, which will be useful in many genre-enhanced representations.

5. Bibliography

[1] K. Crowston and B. H. Kwasnik, "Can document-genre metadata improve information access to large digital collections?," *Library Trends*, In press.

[2] K. Crowston and M. Williams, "Reproduced and emergent genres of communication on the World-Wide Web," *The Information Society*, vol. 16, pp. 201–216, 2000.

[3] W. J. Orlikowski and J. Yates, "Genre repertoire: The structuring of communicative practices in organizations," *Administrative Sciences Quarterly*, vol. 33, pp. 541–574, 1994.

[4] H. Chen, C. Schuffels, and R. Orwig, "Internet categorization and search: A self-organizing approach," *Journal of Visual Communication and Image Representation*, vol. 7, pp. 88–102, 1996.

[5] D. Marcu, "From discourse structures to text summaries," presented at 14th National Conference on Artificial Intelligence (AAAI-97), 1997.

[6] I. Bretan, J. Dewe, A. Hallberg, and N. Wolkert, "Web-Specific genre visualization," presented at WebNet '98, Orlando, 1998.

[7] A. Dillon and B. Gushrowski, "Genres and the Web: Is the personal home page the first uniquely digital genre?," *Journal of the American Society for Information Science*, vol. 5, pp. 202–205, 2000.

[8] R. Furuta and C. C. Marshall, "Genre as Reflection of Technology in the World-Wide Web," Hypermedia Research Lab, Texas A&M, Technical Report 1996.

[9] J. Karlgren, I. Bretan, J. Dewe, A. Hallberg, and N. Wolkert, "Iterative information retrieval using fast clustering and usage-specific genres," presented at Eighth DELOS Workshop: User Interface in Digital Libraries, Stockholm, Sweden, 1998.

[10] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic text categorization in terms of genre and author," *Computational Linguistics*, vol. 26, pp. 471–498, 2000.

[11] T. Petersen, *Art and Architecture Thesaurus*. New York: Oxford, 1994.

[12] J. Dewe, J. Karlgren, and I. Bretan, "Assembling a Balanced Corpus from the Internet," presented at The 11th Nordic Computational Linguistics Conference, Copenhagen, Denmark, 1998.

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

[14] C. Beghtol, "The concept of genre and its characteristics," *Bulletin of the American Society for Information Science & Technology*, vol. 26, 2000.

[15] B. H. Kwasnik, "The legacy of facet analysis," in *Ranganathan and the West*, R. N. Sharma, Ed. New Delhi, India: Sterling, 1992, pp. 98–111.

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

[17] P. Tyrväinen and T. Päivärinta, "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, 1999.

[18] A. Karjalainen, T. Päivärinta, P. Tyrväinen, and J. Rajala, "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, 2000.

[19] B. Kessler, G. Nunberg, and H. Schuetze, "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*. Madrid: Morgan Kaufmann Publishers, 1997, pp. 32–38.

[20] S. R. Ranganathan, *Prologomena to library classification*, 3rd ed. Bombay: Asia Publishing House, 1967.

[21] B. H. Kwasnik, "The role of classification in knowledge representation and discovery," *Library Trends*, vol. 48, pp. 22–47, 1999.

[22] E. G. Toms, D. G. Campbell, and R. Blades, "Does Genre Define the Shape of Information? The Role of Form and Function in User Interaction with Digital Documents," presented at American Society for Information Science; ASIS '99, Washington, DC, 1999.