# Characterizing Genres of Web Pages: Genre Hybridism and Individualization

Marina Santini
*University of Brighton, UK*
*M.Santini@brighton.ac.uk*

## Abstract

*When dealing with genres of web pages, there are two important aspects to be taken into account. On the one hand, the web is fluid, unstable and fast-paced. On the other hand, genres on the web are instantiated in web pages, which are a complex type of document, more composite and unpredictable than paper documents. These two aspects are interwoven and often result in classification hurdles. In this paper, I suggest analyzing these classification problems in terms of two broad textual phenomena: genre hybridism and individualization. The identification of these two phenomena helps pinpoint the range of flexibility that an automatic classification system should have. More precisely, genre hybridism accounts for multi-genre variation within the individual web page, while individualization refers to absence of any recognized genre in a web page. In a few words, the aim of this paper is to show that web pages need a* zero-to-multi-genre classification scheme*, i.e. a scheme that allows zero genre or multi-genre classification, in addition to the traditional single-genre classification.*

## 1 Introduction

The aim of this paper is to show that web pages need a *zero-to-multi genre classification-scheme*, i.e. a scheme that allows zero genre or multi-genre classification, in addition to the traditional single-genre classification. More specifically, a zero genre classification accounts for those web pages that do not fit into any genre, and multi-genre classification is useful when a web page contains more than one genre. Previous work on web genres has assumed that a web page instantiates only a single web genre. This is clear both from automatic approaches (e.g. Shepherd et al., 2004 [34]; Meyer zu Eissen and Stein, 2004 [24]; Lim et al., 2005 [17]), and from qualitative analyses (e.g. Shepherd and Watters, 1998 [36]; Ashekave and Nielsen, 2005 [1]). In contrast with this assumption, in this paper I focus on how genres are instantiated in web pages, since it is often the case that a web page includes more than one genre or has no genre at all. More precisely, I suggest considering two attributes – genre hybridism and individualization – that might help characterize the genre of web pages more accurately. The main benefit that would follow from the inclusion of genre hybridism and individualization in the characterization of genre and from a zero-to-multi genre classification scheme is that they would account for the current classification intractability of many web pages, i.e. those web pages that cannot be classified using a single-genre label.

The argument that a single genre classification scheme is inappropriate for genre classification of web pages has already been developed by scholars and researchers. In particular, it shares some similarities with the multi-facetted classification of genre recently re-proposed in different, but neighboring, areas by Kessler et al. (1997) [21] (automatic genre classification), Tyrväinen and Päivärinta (1999) [40] (document management) and Crowston and Kwasnik, 2004 [8] (information studies). The main difference between the multi-facetted classification and the zero-to-multi genre classification scheme is that the multi-facetted approach is a multi-dimensional approach that highlights different aspects in a document, not necessarily different genres. For example, in automatic genre detection, Kessler et al. (1997) [21] proposed three facets, BROW, NARRATIVE, and GENRE, which relate respectively to the kind of language used in the text, the text typology, and the genre itself. Such an approach returns three independent and uncorrelated classifications, where the classification by genre corresponds to the traditional single-genre labeling. In other words, more than a multi-genre classification, this approach offers three computationally tractable views on a document. By contrast, I would like to focus on genre identification proper, i.e. on the assignment of zero-, one- or multi-genre labels to a web page.

It is important to stress that my view on genre of web pages is geared toward automatic genre identification. Since manual annotation is expensive and time-consuming, the benefits of automatic genre identification have been widely acknowledged for a variety of tasks related to digital or web environments, especially for information retrieval (e.g. Karlgren and Cutting, 1994 [19]; Kessler et al, 1997 [21], Stamatatos et al., 2000 [38]; Meyer zu Eissen and Stein, 2004 [24]; Lim et al., 2005 [17]), digital libraries (e.g. Rauber and Müller-Kögler, 2001 [27]), information extraction (e.g. Maynard et al., 2001 [22]), and other web-related applications. However, although a considerable amount of research has already

1530-1605/07 $20.00 © 2007 IEEE

been carried out in automatic genre classification, most previous work has considered genres as mutually exclusive categories, disregarding the fact that many documents, and in particular many web pages, do not fit into a single genre. This approach has been taken for the sake of practicality, but proves inadequate when dealing with complex documents, like web pages. For this reason, I started implementing an automatic approach that can handle a zero-to-multi genre classification scheme with promising results (Santini, 2006 [33] and Santini et al., 2006 [32]). In this paper, however, I would like to focus on the theoretical aspect and not on the practical implementation, and describe in more detail (a) why a single-genre classification scheme is inappropriate when dealing with web pages, and (b) in which way the two attributes of genre hybridism and individualization help pinpoint the range of flexibility that an automatic genre classification model should have.

This paper is organized as follows: Section 2 describes the characteristics of genres on the web, and some textual traits of web pages; Section 3 presents a discussion about the two attributes that I propose including in the characterization of genres of web pages; finally in Section 4 I draw some conclusions.

## 2 Genres of Web Pages

The difficulty of assigning a single genre to web pages is well expressed by Rosso (2005) [29], when he reports the comments of the participants to his studies:

> "*In summary, the comments provided much insight into participants' experiences of single-genre webpage categorization: problem with pages fitting multiple categories, problems with pages fitting no categories, and general recognition of the characteristic formal elements of many of the web pages.*" (Rosso, 2005: 116)

This difficulty has been pointed out also by other scholars and researchers carrying out surveys on genres of web pages, like Haas and Grams (1998) [15], Crowston and Williams (2000) [9], Karlgren (2000: 99 ff.) [18], Roussinov et al. (2001) [30], and Meyer zu Eissen and Stein (2004) [24].

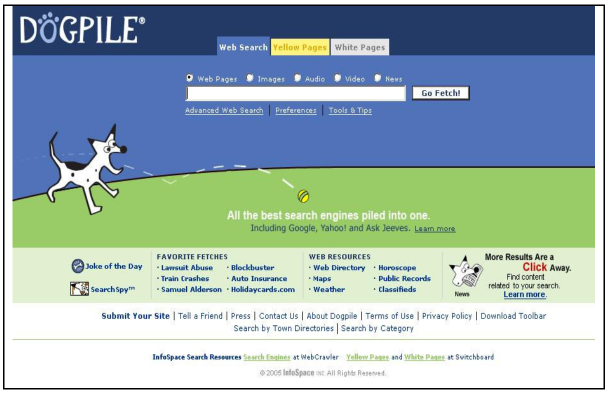
When dealing with genres of web pages, there are two important aspects to be taken into account. On the one hand, the web is fluid, unstable and fast-paced. On the other hand, genres on the web are instantiated in web pages, which are a complex type of document, more composite and unpredictable than paper documents. These two aspects are intertwined and often result in classification intractability because many web pages show more than one genre or do not have any. In the next two subsections I will analyze these two aspects in more detail.

### 2.1 Web Pages

Do web pages differ from documents on other media? Genres on the web are instantiated in web pages, and web pages are complex objects. Even when taken individually, web pages appear to be a composite type of document, with a visual organization of the space, where different communicative purposes and different functions are included at the same time. The intertwining of visual and verbal is not new. What is new is the frequency of use of such a solution. While the linear organization of most of paper documents is still reflected in traditional electronic corpora, like the British National Corpus (BNC), web pages have a visual organization that allows the inclusion of several functions or several texts with different communicative purposes in a single document. For example, the space on a web page can be divided into different sections, organized by lists of links – mainly isolated noun structures or verbal elements (Haas and Grams, 2000: 186-187 [16]) – and snippets of text scattered around the main body of the document, such as navigational buttons, menus, ads, and search boxes, that are visually dislocated in different areas of a single page. Additionally, the effect of hyperlinking, interactivity and multi-functionality can affect the textuality of web pages, which heavily rely also on the use of images and other graphical elements. Although the use of fonts of different types, sizes, and colors, as well as the use of formatting devices, like columns, lines separating different sections of a document, pictures, etc. is not new (cf. Waller, 1987 [41] for a detailed description of the role of both language and typography in the formation of document types), a NEWSPAPER ARTICLE organized in columns and headlines does not lose its specific linguistic and textual characteristics when it is included in a corpus like the BNC. The same is not true for many web pages, because the visual structure of a web page incorporating a NEWSPAPER ARTICLE in most cases cannot be flattened out or ignored without losing important information or functions (cf. Watters and Shepherd, 1997 [42]).

Hence, a web page can be considered as a sort of container of multiple texts – so much so that in coding the pages of their sample, Haas and Grams (2000) [16] repeatedly encountered pages that could be interpreted as comprising more components. Artificially separating what is considered to be the main body from the rest is an arbitrary operation and it would not make sense in many cases, for example in a web page similar to that shown in Figure 1.

In sum, in a web page not all the elements necessarily belong together but they all contribute to form a whole, even without any linear progression.


DÖGPILE®
Web Search
Yellow Pages
White Pages
● Web Pages ○ Images ○ Audio ○ Video ○ News
Go Fetch!
Advanced Web Search | Preferences | Tools & Tips
Joke of the Day
SearchSpy™
FAVORITE FETCHES
· Lawsuit Abuse · Blockbuster
· Train Crashes · Auto Insurance
· Samuel Alderson · Holidaycards.com
WEB RESOURCES
· Web Directory · Horoscope
· Maps · Public Records
· Weather · Classifieds
News
Learn more.
Submit Your Site | Tell a Friend | Press | Contact Us | About Dogpile | Terms of Use | Privacy Policy | Download Toolbar
Search by Town Directories | Search by Category
InfoSpace Search Resources Search Engines at WebCrawler Yellow Pages and White Pages at Switchboard
© 2005 InfoSpace INC All Rights Reserved.

All the best search engines piled into one.
Including Google, Yahoo! and Ask Jeeves. Learn more

More Results Are a Click Away.
Find content related to your search.

**Figure 1. Example of a web page**

The complexity of web pages can be explained by the flexibility provided by HTML and the simplicity of such a language (cf. also Furuta and Marshall, 1996 [12]) that allows the creation of more complex texts without much effort or expertise. This textual or multi-modal complexity is then incorporated in the upgrading of conventions of existing genre, or in the creation of fresh conventions for unprecedented genres, whose introduction is spurred by new communication needs caused by the rapid evolution of the web.

The textual complexity of web pages accounts for the 'malleability' of genre. As Orlikowski and Yates (1994) [25] have stressed, genres are rarely homogeneous. On the contrary, they tend to overlap and mix. In an open communication space, like the web, where many communities meet, each with its own genre system and repertoire (cf. Crowston and Williams, 2000 [9]), phenomena like genre colonization (Beghtol, 2001 [3]), genre combination (Østerlund, 2006 [26]) and genre contamination, common also in other environments, are likely to occur.

For all these reasons, web pages result often more unpredictable and difficult to sort into a single genre than documents in other media, where often social rigidity, work practices or stable settings favor more controlled and standardized text production (cf. Yates and Sumner, 1997 [44]).

## 2.2 Genres on the Web

Do genres on the web differ from genres on other media? The influence of a new communication medium on genre evolution and creation has been historically proved (cf. also Yates and Orlikowski, 1992 [43]). For example, the introduction of printing in the XV century, which entailed a passage from hand-written manuscripts to printed books, radically enlarged and transformed the potential for written genres. Printing is considered an important factor in the development of the NOVEL as we know it today, and for the spreading of the SONNET as a poetic genre throughout Europe.

As printing, the web is a new communication medium that was invented only a few decades ago[1]. The web is also a large and heterogeneous community and a new virtual environment where the interaction among the members (the *internauts* or web users) and the possibility offered by the technology modify existing genres or create new ones, which better satisfy the communication needs brought about by these new conditions.

As shown by Crowston and Williams (2000) [9], who were among the first to study the development of genres on the web, the web has had a substantial impact on the genre repertoire. In their survey, they document the genres in use on the web by sampling and classifying randomly selected web pages in 1996. Crowston and Williams (2000) explicitly refer to Yates and Orlikowski (1992) [43] and Orlikowski and Yates (1994) [25] and define genres as "social type of communicative actions, characterized by a socially recognized communicative purpose and common aspect of form". Crowston and Williams (2000) [9] carry out their survey with three main objectives in mind: (a) trace the development of genres caused by the introduction and the fast evolution of the new medium; (b) observe the effects of lack of central management, enforced, for examples, in organizations; (c) study the effects of the encounter of many communities on the web, using different genres systems. They identify four types of genres:

1. Reproduced genres: 60.6%
2. Adapted genres: 28.6%
3. Novel genres: 5.3%
4. Unclassified web pages: 5.6%

Predictably, most of genres come from previous traditions. When moving to a new medium, it is normal to use forms available on existing media, before elaborating new ones, more tailored to the specific characteristics of the new medium (cf. Askehave and Nielsen, 2004 [2]). Most genres on the web in 1996 were still borrowed (reproduced genres) from other media, while a large proportion appeared to be adapted (adapted genres) to the needs and capability of the new media. Interestingly, in the survey an equal amount of novel genres and unclassified web pages are recorded. Even if the proportions could slightly vary across the different genre typologies according to different criteria of assessment[2], it

[1] The World Wide Web began life in 1989 at CERN, the European Laboratory for Particle Physics in Geneva, Switzerland.

[2] There are several criteria for assessing novel genres. For example, Crowston and Williams (2000) [9] considered FAQs to be a reproduced genre, while in Shepherd and Watters (1998) [36], this genre is mentioned as an example of novel and spontaneous genres.

is nonetheless interesting that novel genres and unclassified web pages have a similar proportion. Both novel genres and unclassified web pages represent the most advanced manifestations of genre evolution. Following the interpretation suggested by Crowston and Williams (2000) [9], novel genres (namely, HOME PAGE, HOTLISTS, PAGES ABOUT WEB SERVERS, AND INTERACTIVE PAGES) serve communicative purposes unique to the web, while some of the unclassified pages may be seen as examples of genres in the process of adaptation to the web or, in other words, as genres still emerging, not fully formed yet. This study returns a picture of a highly dynamic genre repertoire, where changes are still under way and have not fully coalesced yet.

Similar, in some respects, is the interpretation of genre evolution on the web by Shepherd and Watters (1998) [36], who coined a special name – cybergenre – to designate genres in the new medium, more specifically those genres created by the combination of the use of the computer and the Internet. Although nowadays the word 'cybergenre' seems to lose ground in favor of the plainer compound 'web genre' (the authors themselves used this variant as a synonym of cybergenre in Shepherd and Watters., 2004 [35]), it still keeps a fresh connotation. Cybergenre is characterized by the triple <content, form, functionality>, where the first two elements are common to traditional genres, while functionality refers exclusively to the capabilities offered by the web. Shepherd and Watters (1998) [36] propose a hierarchical taxonomy that accounts for the evolution of genres on the web. The driving force behind genre evolution on the web is, in their view, the functionality attribute afforded by the new medium. According to the level of functionality they show, cybergenres can be extant (i.e. based on existing genres) or novel (i.e. not like any existing genre in any other medium). Extant genres can be replicated, i.e. based on genres existing in other media, or variant, i.e. a modification of existing genres. Novel genres can be emergent, i.e. derived but significantly different from existing genres, or spontaneous, i.e. never employed in other media. The key evolutionary aspect is the functionality attribute, which is illustrated by two examples, the NEWS cybergenre and the MATH DICTIONARY cybergenre. Shepherd and Watters (1999) [37] exemplify what they intend by the attribute of functionality. In a survey of 96 complete web sites (and not individual web pages as in Crowston and Williams, 2000 [9]), they identify five cybergenres out of the six they had in their palette (namely, HOME PAGE, BROCHURE, RESOURCE, CATALOGUE, SEARCH ENGINE, and GAME). Each of these cybergenres show different level of functionality, defined in terms of browsing, email facility, search, discussion, interaction, email ordering/enquiring, online ordering, online enquiring, interactivity, and collaborative computing. Illuminatingly, Shepherd and Watters (1998, 1989) [36] [37] emphasize that functionality has caused a leap in genre evolution creating a new species, the web genre. However, they assume that a web page instantiates only a single web genre. This assumption is clear in their practical experiments (Shepherd et al., 2004 [34] and Kennedy and Shepherd, 2005 [20]), where they carry out single-label classification of HOME PAGES.

Crowston and Williams (2000) [9] and Shepherd and Watters (1998, 1999) [36] [37] use complementary classification schemes. On one side, Shepherd and Watters (1998, 1999) never mention unclassified web pages, while Crowston and Williams (2000) [9] realistically find them in their sample, and consequently include them in their classification scheme. On the other side, while Crowston and Williams (2000) [9], drawing on Yates and Orlikowski (1992) [43], suggest that new genres are mostly derived from earlier genres that might have seemed appropriate to the situation, Shepherd and Watters (1998, 1999) [36] [37] introduce the typology of spontaneous cybergenres, which do not have any counterpart in other media. This view is supported also by Haas and Grams (1998) [15] as they state: "The Web, with its multimedia capabilities, has also spawned page types that have no equivalent in the print world, such as he home page of a corporation, a page containing audio or video clips, or interactive pages". As also pointed out by Dillon and Gushrowski (2000) [10], it is unlikely that just mimicking existing genres may support adequate design of new information types that the digital world enables, because this would mean underutilize the power of the new medium to provide innovative information structures.

By combining the two classification schemes, the one by Shepherd and Watters (1998, 1999) [36] [37], and the other by Crowston and Williams (2000) [9], we get a wider range of genre typologies on the web:

1. reproduced/replicated genres,
2. adapted/variant genres,
3. emergent/novel genres,
4. spontaneous genres,
5. unclassified web pages.

Together with the functionality attribute suggested by Shepherd and Watters (1999) [37], there is another important attribute that characterize web genres: the use of hypertext and HTML.

For Askehave and Nielsen (2004, 2005) [1] [2] the use of hypertext has created a new way of reading, the hypertext reading. They introduce the concept of 'modal shift' between reading mode and navigating mode for analyzing the HOME PAGE web genre, thus inaugurating the two-dimensional perspective on genre analysis.

For Rehm (2006) [28] different types of hypertexts can be conceptualized as individual hypertexts type. Each hypertext type corresponds to a web genre. This leads to the equation "one web site = one web genre", where an entire web site (including one or more pages), and not the individual web page, is used as a unit of analysis. From an evolutionary point of view, Rehm distinguishes between automatically converted and manually prepared HTML documents. The influence of the genre on the automatic conversion of existing documents is straightforward and the conversion process does not significantly alter the genre itself, so that we can assume that a document's original genre is directly transferred into the web. The evolutionary processes that shape and form web genre with regard to the manual creation of web sites includes four phases: production, modification, change and reception. His assumption is that when building a site with a specific web genre, an author incorporates, consciously or unconsciously, elements of related web sites. Over a period of time, this process generates web specific conventions and rules that authors choose to apply, extend or break. In other words, Rehm suggests that the process of imitation maintains stability in the genre repertoire, while change is determined by the extension or the break of conventions. This view complies with Baktin's interpretation of language, where there is always a tension between 'centripetal' forces, allowing the continuity of communication, and 'centrifugal' forces, allowing change and evolution. In this respect, for Rehm too, new genres are always the outcome of transformation of earlier genres, similarly to Crowston and Williams (2000) [9]. In contrast with most analysts of web genres, who mainly consider genre evolution on the web as a fast process, Rehm sees the process of emerging rules and conventions for specific web genres as a "slow-going" progress. This view is somehow opposed to Shepherd and Watters' (2004) [35], for whom the identification of web genres is as difficult as hitting a moving target, because of the continuous rearrangement of the genre repertoire caused by the functionality afforded by the new medium. Similarly, Haas and Grams (2000) [16] see the web as quickly moving target, where analyzing format, genres and the design conventions that go into web pages is tricky because the technology is constantly upgraded. Dillon and Vaughan (1997) [11] suggest relying on the attribute of 'shape' of a document to overcome this disorientation on the web, while Yates and Sumner (1997) [44] argue that such a rapid advance in new technologies results in the evolution of increasingly well-defined genres to better support the new communicative needs and practices, thus providing 'fixity', i.e. stability, in the production and transmission of texts. That the web is a fast-paced medium and that genre repertoire follows its rhythm by being upgraded and updated with the introduction of novel genres is also proved by the many genre analyses on new kinds of texts (e.g. cf. the recent analyses of WIKIS by Mehler, 2006 [23] and Copestake, 2006 [6]).

In conclusion, the web can be defined as a "moving target" (Shepherd and Watters, 2004 [35]), although owning some kind of "self-regulatory force" (Yates and Sumner, 1997 [44]), and probably this is the metaphor that best suits this stage of the web evolution.

Additionally, I would like to pinpoint one effect of the fast evolution: the presence of emerging genres.

### 2.2.1 Emerging genres

Emerging genres represent a transitional phase in genre evolution. They are genres still in formation, without a name, and not fully standardized or acknowledged. They should not be confused with novel but emerged genres, i.e. new genres that are acknowledged by an audience and whose genre conventions can be singled out and coherently described. By contrast, emerging genres are still in a phase of formation, where it is not yet clear whether they will ever coalesce into a new communication object. They can be considered as hypothesised genres.

Although evolving on all media and in all historical period, genres are assumed to be slow-forming, often emerging only over generations of producers and consumers, and proving resistant to change (cf. Dillon and Gushrowski, 2000 [10]). However, since the web is recent, fluid and evolving at fast pace, the emergence of novel genres is much more rapid than in other media, as observed by Boese (2005: 14) [5] "new genres are emerging frequently and a lot of old ones are in a state of flux".

The concept of emerging genres has not been explicitly formulated in the genre literature. However, it results useful because it might account for unclassified or unclassifiable web pages. As shown by the findings reported in Crowston and Williams (2000) [9], there is a number of web pages that cannot be classified by genre, for different reasons. More specifically, they state: "some of these unclassified pages may in fact be emerging genres". It thus appears that one of the reasons that explains the difficulty of classifying web pages by genre is the existence of web pages that may belong to emerging genres, i.e. to textual patterns without any clear or acknowledged genre convention. The hectic introduction of new genres creates transitional phases where genre conventions in web pages are not clear. For this reason, many web pages are difficult to sort into acknowledged genres and remain unclassified. For example, before 1998, WEB LOGS (or BLOGS) were already present on the web, but they were still not identified as a genre. They were just "web pages" with similar characteristics and functions. In 1999, suddenly a community sprang up using this new genre (Blood, 2000 [4]). Only at this point, the genre label WEB LOG or BLOG started being spread and being recognized.

However, it is worth repeating with Crowston and Williams (2000) [9] that the emergence of a novel genre depends on social acceptance. Consequently, it is impossible to predict exactly when an emerging genre will become a fully independent genre or even whether it will coalesce at all. It is nonetheless important to have a concept that captures the transitional phase that currently characterizes so many web pages, i.e. those web pages that cannot be sorted into any acknowledged genre.

As stated earlier, emerging genre are hypothesized genres. Although we can confirm and disconfirm these hypotheses only in future, some quantitative cues may indicate that a new genre is ‘in formation’. Computationally, we might suspect an emerging genre when there is a recurrent textual pattern without an acknowledged name. Emerging genes do not have a name because a genre name becomes acknowledged when the genre itself has an active role and a communicative function in a community or society (Swales, 1990: 54-57 [39]; Görlach, 2004: 9 [13]; Görlach, 2002 [14]).

If we see the web as a dynamic environment, we could say that there are three forces interacting: what we bring from the past (reproduced genres), what is new or adapted to the new environment (novel genres and adapted genres), what is going to emerge and is not fully formed yet (emerging genres) (Figure 2).

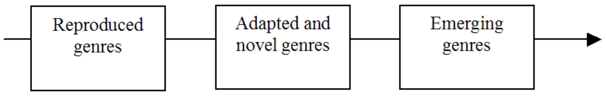


**Figure 2. Genre evolution on the web**

This representation of the genre evolution can be virtually applied to any communication medium. On the web this process is much more evident because the web is fast and we can see many changes taking place in the short term. I suggest postulating emerging genres as the pushing force behind genre change and creation. This view of genre evolution may complement previous studies on the same subject.

## 2.3 Summary

In summary, I suggested that the fluidity and fast-paced dynamism of the web together with the complexity of web pages affect the genre system and genre repertoire on the web. More precisely:

1. As the web is a new and still developing communication medium, mechanisms of re-adjustments – such as the adaptation of existing genres to the new conditions and creation of novel genres – are common. These re-adjustments bring about a transitional phase where genre conventions are unclear, and web pages do not belong to any recognizable genre.
2. The mixture of several genres, coming from different traditions, in a single web page is an easy and fast operation.
3. The lack of any institutionalized control (like, for example, in an organizational intranet) can stimulate authorial creativity and individualization.

In order to account for: (a) unclear genre conventions, (b) genre mixture and (c) authorial creativity, I suggest including in any characterization of genre of web pages the two attributes of genre hybridism and individualization, discussed in the next section.

## 3 Genre Hybridism and Individualization

One effect of this eventful and fermenting scenario is the classification intractability of many web pages. I suggest breaking down this classification intractability into two broad textual phenomena: genre hybridism and individualization.

Broadly speaking, genres show sets of standardized or conventional characteristics that make them recognizable, and this identity raises specific expectations. Together with conventions and expectations, genres have many other traits. I would like to focus on three traits, namely hybridism, individualization and evolution.

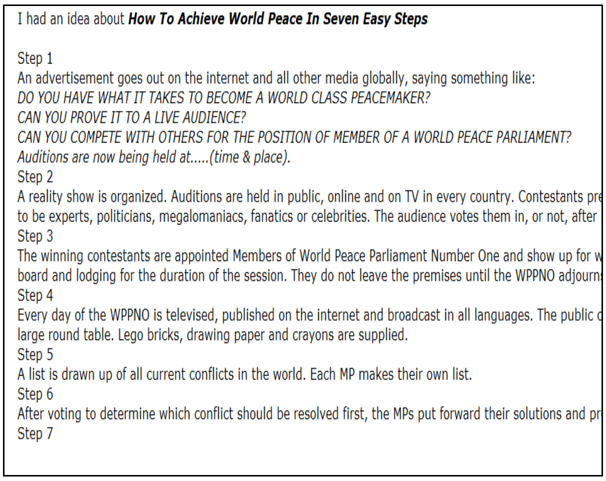
First, genres are not mutually exclusive and different genres can be merged into a single document, generating hybrid forms. Second, genres allow a certain freedom of variation, and consequently can be individualized. Finally, genre repertoires are dynamic, i.e. they change over time, thus triggering genre change and evolution. It is also important to notice that before genre conventions become fully standardized, a genre does not have an official name. A genre name becomes acknowledged when the genre itself has an active role (Swales, 1990: 54-57 [39]; Görlach, 2004: 9 [13]; Görlach, 2002 [14]). Before this acknowledgement, a genre shows hybrid or individualized forms, and indistinct functions

All these traits can be accounted for by the two attributes of genre hybridism and individualization.

Genre hybridism is broad term accounting for several phenomena. It has often been pointed out that genres are not discrete systems (e.g. Gledhill as cited in Chandler, 1997 [6]). A number of genre combinations are possible and common. For example, a mixed genre, like the TRAGI-COMEDY, is a genre having its own blending aspects of two or more genres. Multi-genre documents are documents where two or more genres overlap without creating a specific and more standardized genre, as in the case of ESHOPS, which are often also SEARCH PAGES, as noted also by Meyer zu Eissen and Stein (2004) [24]. Some genres are intrinsically mixed, such as the NEWSLETTER, which contains EDITORIALS, REPORTS, INTERVIEWS and so on. An additional problem concerns the fuzziness of genre labels because, for example, the same document can be named NEWS BULLETIN or PRESS RELEASE, as noted by Roussinov et al. (2001) [30] (cf. also the difficulties of developing a genre palette described in Rosso, 2005: 67 ff.).

Hybrid genres abound and are very common in the mass media (as noted by Fairclough, cited in Chandler, 1997 [6]). But in an open environment, like the web, where many needs and communities meet, genre mixture appear to be more extended. For instance, Figure 3 shows a personal BLOG that includes a HOW-TO. The attribute of genre hybridism does not break down if this page is both a BLOG and a HOW-TO, or a HOW-TO within a BLOG or vice-versa, or half a BLOG and half a HOW-TO. The concept of genre hybridism as intended in this paper simply helps pin down when a web page contains more than one genre, regardless how these genres relate to each other.

The acknowledgement that a web page can be hybrid is important when dealing with automatic genre identification because traditional single-label classification algorithms are usually confused by hybrid genre conventions.



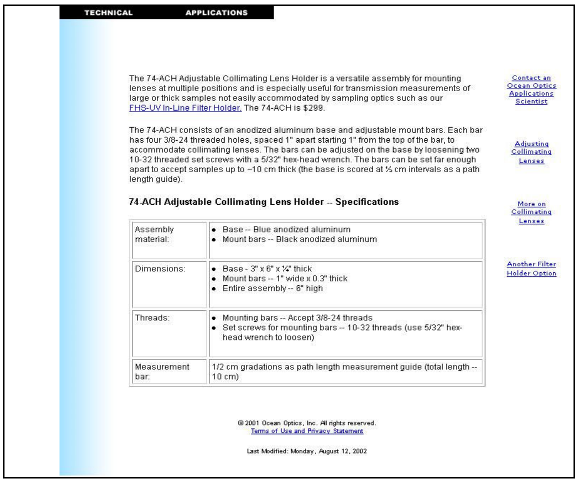
I had an idea about ***How To Achieve World Peace In Seven Easy Steps***

Step 1
An advertisement goes out on the internet and all other media globally, saying something like:
*DO YOU HAVE WHAT IT TAKES TO BECOME A WORLD CLASS PEACEMAKER?*
*CAN YOU PROVE IT TO A LIVE AUDIENCE?*
*CAN YOU COMPETE WITH OTHERS FOR THE POSITION OF MEMBER OF A WORLD PEACE PARLIAMENT?*
*Auditions are now being held at.....(time & place).*
Step 2
A reality show is organized. Auditions are held in public, online and on TV in every country. Contestants pre
to be experts, politicians, megalomaniacs, fanatics or celebrities. The audience votes them in, or not, after
Step 3
The winning contestants are appointed Members of World Peace Parliament Number One and show up for w
board and lodging for the duration of the session. They do not leave the premises until the WPPNO adjourn
Step 4
Every day of the WPPNO is televised, published on the internet and broadcast in all languages. The public c
large round table. Lego bricks, drawing paper and crayons are supplied.
Step 5
A list is drawn up of all current conflicts in the world. Each MP makes their own list.
Step 6
After voting to determine which conflict should be resolved first, the MPs put forward their solutions and pr
Step 7

**Figure 3. A web page showing genre hybridism**

Individualization refers to the impact of authorial experimentation. Although there is a process of imitation that favors similarities (cf. Furuta and Marshall, 1996 [12]; Yates and Sumner, 1997 [44]; Askehave and Nielsen, 2004 [1]; Rehm, 2006 [28]), authors of web pages are virtually free to invent or propose any genre variations. It is so much so that many web pages cannot be classified into any genres (cf. Haas and Grams, 1998 [15]). For instance, a web page like the one shown in Figure 4 does not find much consensus on the genre label. According to a user study based on 135 subjects– fully documented in Santini, *Forthcoming* [31] – this web page has been classified as follows: ABOUT PAGE (28 users), ESHOP (20 users), HOW-TO (14 users), TUTORIAL (9 users), CORPORATE HOME PAGE (5 users), FAQs (3 users), NET AD (3 users), BLOG (1 user), ONLINE FORM (1 user), and SEARCH PAGE (1 user). The majority of users (36) added their own labels, which ranged from 'content page' to 'product manual'[3]; and finally a number of users (14) frankly declared that they did know its genre.

The concept of individualization covers web pages like the web page shown in Figure 4, where genre conventions are not clear, and the range of oscillation is large. I ascribe this oscillation to authorial creativity, because a web page creator is virtually free to publish any kind of text in any kind of format, while in the paper world constraints on genre conventions are stronger, as noted by Yates and Sumner (1997) [44].



The 74-ACH Adjustable Collimating Lens Holder is a versatile assembly for mounting lenses at multiple positions and is especially useful for transmission measurements of large or thick samples not easily accommodated by sampling optics such as our FHS-UV In-Line Filter Holder. The 74-ACH is $299.

The 74-ACH consists of an anodized aluminum base and adjustable mount bars. Each bar has four 3/8-24 threaded holes, spaced 1" apart starting 1" from the top of the bar, to accommodate collimating lenses. The bars can be adjusted on the base by loosening two 10-32 threaded set screws with a 5/32" hex-head wrench. The bars can be set far enough apart to accept samples up to ~10 cm thick (the base is scored at ½ cm intervals as a path length guide).

**74-ACH Adjustable Collimating Lens Holder -- Specifications**

| | |
|---|---|
| Assembly material: | • Base -- Blue anodized aluminum<br>• Mount bars -- Black anodized aluminum |
| Dimensions: | • Base - 3" x 6" x ¼" thick<br>• Mount bars -- 1" wide x 0.3" thick<br>• Entire assembly -- 6" high |
| Threads: | • Mounting bars -- Accept 3/8-24 threads<br>• Set screws for mounting bars -- 10-32 threads (use 5/32" hex-head wrench to loosen) |
| Measurement bar: | 1/2 cm gradations as path length measurement guide (total length -- 10 cm) |

Contact an Ocean Optics Applications Scientist
Adjusting Collimating Lenses
More on Collimating Lenses
Another Filter Holder Option

© 2001 Ocean Optics, Inc. All rights reserved.
Terms of Use and Privacy Statement

Last Modified: Monday, August 12, 2002

**Figure 4. A web page showing individualization**

It is worth noting that the border between genre hybridism and individualization remains fuzzy, because an author can use a mixture of existing genres, or employ individualized and unprecedented solutions to create a personalized web page.

In conclusion, if we use Baktin's metaphor, genre conventions can be seen as the centripetal force that keeps stability in genre repertoire, while hybridism and individualization are the centrifugal forces that de-

[3] The complete list of labels added for the web page shown in Figure 4 includes: *content page, information page, online product information, product catalogue, product documentation, product information, product manual, product specification page, sub page of an online store, tech specifications, technical description, technical documentation, technical information page, technical instructions, technical product description,* and *(normal) webpage* (sic).

stabilize the system. This struggle between "stasis" and "change" (cf. also Yates and Sumner (1997) [44]) gives rise to an intermediated phase in genre evolution – the transitional phase of emerging genres – that shows unclear genre conventions, which can also be explained in terms of genre hybridism and individualization.

Unclear genre conventions of emerging genre, genre mixture (cf. Figure 3) and authorial creativity (cf. Figure 4) justify the need for broadening the characterization of genre. I propose the following broad theoretical characterization of genres of web pages: genres are named communication artifacts characterized by conventions, raising expectations, showing hybridism or individualization, and undergoing evolution.

This characterization is flexible enough to encompass not only genres of web pages and other digital genres, but also paper genres, both literary and practical genres.

In conclusion, the identification of the two attributes of genre hybridism and individualization help characterize genre of web pages more accurately, and has a practical purpose: genre hybridism accounts for multi-genre classification, and individualization for no-genre assignment.

# 4 Conclusion

In this paper, I proposed a characterization of genres of web pages that includes the two textual attributes of genre hybridism and individualization. I explained that they are useful because they help describe more accurately how genre categories are instantiated in web pages. The web is a complex scenario where:

- the mixture of several genres in a single web page is fast operation;
- the lack of any institutionalized control can stimulate authorial creativity and individualization;
- the constant introduction of web technologies brings about the transitional phase of emerging genres, where genre conventions are unclear.

The identification of the two attributes of genre hybridism and individualization help understand how flexible a classification system for web pages should be: genre hybridism accounts for multi-genre classification, and individualization for zero-genre classification.

# 5 References

[1] Askehave I. and Nielsen A. E. (2005). "What are the Characteristic of Digital Genres? – Genre Theory from a Multi-modal Perspective". *Proceedings of the 38$^{th}$ Hawaii International Conference on System Sciences.*

[2] Askehave I. and Nielsen A. E. (2004), *Web-Mediated Genres - A Challenge to Traditional Genre Theory*. Working Paper nr. 6. Center for Virksomhedskommunikation, Aarhus School of Business, Denmark.

[3] Beghtol C. (2001). "The Concept of Genre and Its Characteristics". *Bulletin of The American Society for Information Science and Technology*, Vol. 27, No. 2.

[4] Blood R. (2000). Weblogs: A History and Perspective. Rebecca's Pocket. 07 September 2000. <http://www.rebeccablood.net/essays/weblog_history.html">

[5] Boese E. (2005). Stereotyping the Web: Genre Classification of Web Documents, M.S. Thesis, Computer Science Department, Colorado State University, Fort Collins, CO, March 2005.

[6] Chandler D. (1997). An Introduction to Genre Theory. <http://www.aber.ac.uk/media/Documents/intgenre/intgenre.html>

[7] Copestake A. (2006). "Errors in wikis". *Proceedings of the workshop on NEW TEXT. Wikis and blogs and other dynamic text sources* (EACL 2006), Trento.

[8] Crowston K. and Kwasnik B. (2004). "A Framework for Creating a Facetted Classification for Genres: Addressing Issues of Multidimensionality". *Proceedings of the 37th Hawaii International Conference on System Science*.

[9] Crowston K. and Williams, M. (2000). "Reproduced and emergent genres of communication on the World-Wide Web". *The Information Society*, Vol. 16, No. 3, pp. 201-216.

[10] Dillon A. and 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*, Vol. 51, No. 2.

[11] Dillon A. and Vaughan M. (1997). "It's the journey and the destination: Shape and the emergent property of genre in evaluating digital documents". *New Review of Multimedia and Hypermedia*, Vol. 3, pp. 91-106

[12] Furuta R. and Marshall C. (1996). "Genre as reflection of technology in the World-Wide Web. In Hypermedia Design". *Proceedings of the International Workshop on Hypermedia Design* (IWHD 95). 182-95, Springer-Verlag.

[13] Görlach M. (2004). *Text Types and the History of English*, Mouton de Gruyter, Berlin - New York.

[14] Görlach M. (2002). "What's in a Name? Terms Designating Text Types and the History of English". In Fisher A., Tottie G. and Lehmann H. M. (eds.).

*Text Types and Corpora*. Gunter Narr Verlag. Tübingen (Germany), pp. 17-27.

[15] Haas S. and Grams E. (1998). “Page and Link Classifications: Connecting Diverse Resources”. *Proceedings of Digital Libraries ’98* – Third ACM Conference on Digital Libraries, pp. 99-107.

[16] Haas S. and Grams E. (2000). “Readers, Authors, and Page Structure: A Discussion of Four Questions Arising from a Content Analysis of Web Pages”. *Journal of the American Society for Information Science*, Vol. 51, No. 2, pp. 181-192.

[17] Lim C. S., Lee K. J. and Kim G. C. (2005). “Automatic Genre Detection of Web Documents”. Su K., Tsujii J., Lee J., Kwong O. Y. (eds.) *Natural Language Processing – IJCNLP 2004*, Springer, Berlin Heidelberg.

[18] Karlgren J. (2000). Stylistic Experiments for Information Retrieval, Thesis submitted for the degree of Doctor of Philosophy, Department of Linguistics, Stockholm University, Sweden.

[19] Karlgren J. and Cutting D. (1994). “Recognizing Text Genre with Simple Metrics Using Discriminant Analysis”. *Proceedings of the 15th International Conference on Computational Linguistics* (COLING 1994*)*, Kyoto, Japan.

[20] Kennedy A. and Shepherd M. (2005). “Automatic Identification of Home Pages on the Web”. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*.

[21] Kessler B., Numberg G. and Shütze, H. (1997), “Automatic Detection of Text Genre”. *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*.

[22] Maynard D., Tablan V., Ursu C., Cunningham H. and Wilks Y. (2001). “Named Entity Recognition from Diverse Text Types”. *EuroConference in Recent Advances in NLP* (RANLP - 2001), Tzigov Chark, Bulgaria.

[23] Mehler A. (2006). “Text Linkage in the Wiki Medium - A Comparative Study”. *Proceedings of the workshop on NEW TEXT. Wikis and blogs and other dynamic text sources* (EACL 2006), Trento, Italy.

[24] Meyer zu Eissen S. and Stein B. (2004). “Genre Classification of Web Pages: User Study and Feasibility Analysis”. Biundo S., Fruhwirth T. and Palm G. (eds.), KI 2004: *Advances in Artificial Intelligence*, Springer, Berlin Hedelberg New York, pp. 256-269.

[25] Orlikowski W. and Yates J. (1994), “Genre repertoire: The structuring of communicative practices in organizations”. *Administrative Science Quarterly*, Vol. 39, No. 4, pp. 541-574.

[26] Østerlund C. (2006). “Combining Genres: How Practice Matters”. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*.

[27] Rauber A. and Müller-Kögler A. (2001). “Integrating Automatic Genre Analysis into Digital Libraries”. *Proceedings of ACM/IEEE joint Conference on Digital Libraries 2001*, Roanoke, USA.

[28] Rehm G. (2006). “Hypertext Types and Markup Languages”. Metzing D. and Witt A. (eds.), *Linguistic Modelling of Information and Markup Languages*. Springer, 2006 (in preparation).

[29] Rosso M. (2005). Using Genre to Improve Web Search, PhD dissertation submitted for the degree of Doctor of Philosophy, University of North Carolina, Chapel Hill, USA.

[30] Roussinov D., Crowston K., Nilan M., Kwasnik B., Cai J. and Liu X. (2001), “Genre Based Navigation on the Web”. *Proceedings of the 34th Hawaii International Conference on System Sciences*.

[31] Santini, *Forthcoming*. Automatic Identification of Genre in Web Pages, PhD thesis, University of Brighton, United Kingdom.

[32] Santini M., Power R. and Evans R. (2006). “Implementing a Characterization of Genre for Automatic Genre Identification of Web Pages”. *Proceeding of Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics* (COLING/ACL 2006), Poster Session, Sydney, Australia.

[33] Santini M. (2006). “Identifying Genres of Web Pages”. *Proceeding of TALN 2006* (Traitement Automatique des Langues Naturelles), Leuven, Belgium.

[34] Shepherd M., Watters C. and Kennedy A. (2004). “Cybergenre: Automatic Identification of Home Pages on the Web”. *Journal of Web Engineering* Vol. 3, No. 3 & No. 4, pp 236-251

[35] Shepherd M. and Watters C. (2004). “Identifying Web Genre: Hitting a Moving Target”, *Proceedings of the WWW2004 Conference. Workshop on Measuring Web Search Effectiveness: The User Perspective*, New York, USA.

[36] Shepherd M. and Watters C. (1998). “The Evolution of Cybergenre”. *Proceedings of the 31st Hawaii International Conference on System Sciences*.

[37] Shepherd M. and Watters C. (1999). “The Functionality Attribute of Cybergenres”. *Proceedings of the 32nd Hawaii International Conference on System Sciences*.

[38] Stamatatos E., Fakotakis N. and Kokkinakis G. (2000). “Text Genre Detection Using Common Word Frequencies”. *Proceedings of the 18th International Conference on Computational Linguistics* (COLING 2000), Saarbrücken, Germany.

[39] Swales J. (1990), *Genre Analysis. English in academic and research settings*. Cambridge University Press, Cambridge.

[40] Tyrväinen P. and Päivärinta T. (1999). “On Rethinking Organizational Document Genres for Electronic Document Management”. *Proceedings of the 32nd Hawaii International Conference on System Sciences*.

[41] Waller R. (1987). The typographic contribution to language, Thesis submitted for the degree of Doctor of Philosophy, University of Reading, UK.

[42] Watters C. and Shepherd M. (1997). “The Digital Broadsheet: An Evolving Genre. *Proceedings of The 30th Annual Hawaii International Conference on System Sciences*.

[43] Yates J. and Orlikowski W. (1992), “Genres of organizational communication: A structural approach to studying communications and media”. *Academy of Management Review*, Vol. 17, No. 2, pp. 229-326.

[44] Yates S. and Sumner T. (1997). “Digital Genres and the New Burden of Fixity”. *Proceedings of the 30th Hawaii International Conference on System Sciences*.