Bridging the Gap: A Genre Analysis of Weblogs

Susan C. Herring
Lois Ann Scheidt
Sabrina Bonus
Elijah Wright
School of Library and Information Science
Indiana University, Bloomington
{herring,lscheidt,sbonus,ellwrigh}@indiana.edu

Abstract

Weblogs (blogs)—frequently modified web pages in which dated entries are listed in reverse chronological sequence—are the latest genre of Internet communication to attain widespread popularity, yet their characteristics have not been systematically described. This paper presents the results of a content analysis of 203 randomly-selected weblogs, comparing the empirically observable features of the corpus with popular claims about the nature of weblogs, and finding them to differ in a number of respects. Notably, blog authors, journalists and scholars alike exaggerate the extent to which blogs are interlinked, interactive, and oriented towards external events, and underestimate the importance of blogs as individualistic, intimate forms of self-expression. Based on the profile generated by the empirical analysis, we consider the likely antecedents of the blog genre, situate it with respect to the dominant forms of digital communication on the Internet today, and advance predictions about its long-term impacts.

1. Introduction

Weblogs (blogs), defined here as frequently modified web pages in which dated entries are listed in reverse chronological sequence, are becoming an increasingly popular form of communication on the World Wide Web. Although some claim that the earliest blog was the first website created by Tim Berners-Lee in 1991[27], the present-day format first appeared in 1996,[1] and the term weblog was first applied to it in 1997.[2] Since mid-1999, blogging as an online activity has been increasing exponentially, enabled by the release of the first free blogging software (Pitas), and fueled by reports from the mainstream media of the grassroots power of blogs as alternative news sources, especially in the aftermath of 9/11/01 and during the U.S.-led invasion of Iraq. Current estimates place the number of sites calling themselves blogs at over 1.3 million [1], of which about 870,000 are actively maintained. Moreover, as blogging software becomes easier to use, the number of bloggers continues to increase. In the several months during spring 2003 in which we conducted the research for the present paper, the number of publicly available blogs more than doubled.

As with other Internet communication protocols that have blossomed into seemingly sudden, intense popularity (e.g., email; the WWW; peer-to-peer file transfer), blogs are being hailed as fundamentally different from what came before, and as possessing a socially-transformative, democratizing potential. Journalists see blogs as alternative sources of news and public opinion [22]. Educators and business people see them as environments for knowledge sharing [14]; blogs created for this purpose within an organization or institution are sometimes called k(nowledge)-logs. Last but not least, private individuals create blogs as a vehicle for self-expression and self-empowerment [6]. According to Blood [6], blogging makes people more thoughtful and articulate observers of the world around them. All of this is purportedly brought about by the technical ability that blogging software affords to update web pages rapidly and easily.

In this paper, we seek to characterize the properties of the emergent blog genre, and situate it with respect to offline genres, as well as with respect to the broader genre ecology of the Internet [13]. Our primary goal in so doing is to provide an empirical snapshot of the weblog in its present stage, as a

[1] A site created by Dave Winer as part of the 24 Hours of Democracy Project [27].
[2] By Jorn Barger. The clipping 'blog' came into use after Peter Merholz started pronouncing 'weblog' as 'wee-blog' in early 1999 [6].

historical record for purposes of comparison with future stages of evolution. A further goal is to contribute to a theoretical understanding of how technological changes trigger the formation of new genres, which in turn may affect the genre ecology of a larger domain such as the Internet. Our analysis suggests that the blog is neither fundamentally new nor unique, but that it—along with other emergent genres driven by interactive web technologies—occupies a new position in the Internet genre ecology. Specifically, it forms a de facto bridge between multimedia HTML documents and text-based computer-mediated communication, thereby blurring the traditional distinction between these two dominant Internet paradigms, and potentially contributing to its breakdown in the future.

2. Background

2.1. Genre analysis

The present research is premised on the assumption that recurrent electronic communication practices can meaningfully be characterized as genres, a perspective pioneered by Yates and Orlikowski [28] in their analysis of organizational uses of email. Yates and Orlikowski's work, along with much of the subsequent work it has inspired, draws on traditional models of genre from rhetoric (see, e.g., Miller's definition of a genre as "typified rhetorical action based in recurrent situations" [24]). At its core, genre analysis is an exercise in classification of "typified acts of communication" based on their form and substance [28]. Similarly, Swales [27] characterizes a genre as "a class of communicative events" having "a shared set of communicative purposes" and similar structures, stylistic features, content and intended audiences. In addition, Swales notes that a genre is usually named and recognized by members of the culture in which it is found. According to these criteria, weblogs are a good *prima facie* candidate for genre status, in that they are named, and—as we will show—tend to exhibit common structures and substance.

2.2. Web genres

Recent years have seen a growing interest in the identification of genres on the World Wide Web (e.g., [10], [25]). As a type of web document, blogs are related to—and some would claim, replacing—personal home pages: both are typically created and maintained by a single individual, and their content tends to focus on the creator or his/her interests. While to our knowledge ours is the first systematic genre analysis of blogs, personal home pages have received considerable attention from Web genre analysts. Crowston and Williams [10] cite personal home pages as an example of an "emergent" (rather than a "reproduced") web genre; for Dillon and Gushrowski [11], it is the first uniquely web-based genre. Bates and Lu [3], Chandler [8], and Dillon and Gushrowski [11] identify structural characteristics of personal home pages, including the presence of personal information about the creator, number and patterns of hyperlinks; layout; presence of formulaic welcome messages; and iconographic and technical features. (See Döring [12] for a useful overview of this literature.) Content analyses have been conducted of the home pages of business web sites [16]. Arnold and Miller [2] find gender differences in the structure and content of home pages created by academic professionals.

The findings of the latter two studies suggest that older practices from related off-line genres carry over into the web genres, making them at least partially "reproduced" in the sense of Crowston and Williams [10]. A question that arises is whether blogs are an emergent or a reproduced genre. Our analysis suggests that blogs are neither unique nor reproduced entirely from offline genres, but rather constitute a hybrid genre that draws from multiple sources, including other Internet genres.

2.3. Previous blog research

Most descriptions of the blog genre to date emanate from blog authors themselves. Blog authors tend to define blogs around their characteristic entries-posted-in-reverse-chronological-order format, which is derived from the software used to create and maintain blogs [20]. Updates should be frequent: according to Rebecca Blood [6, p. 9], one of the most visible bloggers to publish in print about blogs, "most webloggers make a point of giving their readers something new to read every day." In terms of patterns of use, the prototypical blog is focused around links to other sites of interest (or other blogs) on the Web,[3] with blogger commentary for added value [5, 6]. This type of blog, in which the blogger "pre-surfs" the Web and directs readers to selected content, is known as a filter. Blood [6] distinguishes three basic types of weblogs: filters, personal journals, and notebooks. The content of filters is external to the blogger (world events, online happenings, etc.), while the content of personal journals is internal (the blogger's thoughts and internal workings); notebooks may contain either

[3] A high incidence of links is central to Blood's definition of blogs: "I would go so far as to say that if you are not linking to your primary material when you refer to it—especially when in disagreement—no matter what the format or update frequency of your website, you are not keeping a weblog" [6, pp. 18-19].

external or internal content, and are distinguished by longer, focused essays. Blood adds that although the earliest blogs were filters, the journal type has now become more common. For Blood, blogs are unique (in her term, "native") to the Web, rather than carried over from off-line genres.

Among its practitioners, blogging is also frequently characterized as socially interactive and community-like in nature. Not only do blogs link to one another [7], but some blogs allow readers to post comments to individual entries, giving rise to "conversational" exchanges on the blog itself [6]. Blood claims that social interactivity is highest in journal-type blogs.

Although empirical research on blogs is thus far limited, the results of two descriptive studies bear on the general claims advanced about blogs above. Halavais [17] found that popular news stories—external content—were the most common topics of discussion in a random sample of 125 blogs. Krishnamurthy [21] analyzed patterns of posting to a community news blog in the week immediately following the events of 9/11/2001, and found that the daily number of posts increased (from an average of 28 to 75), while the number of links per post decreased (from an average of 1.89 to 1.16) and the average number of comments received per post remained the same (about 17 per day). In general, according to Krishnamurthy, "the posts that are most insightful or controversial get the most comments." The findings of these studies are consistent with the predominant view of blogs as news filters, and bloggers as highly interconnected.

As background to his study, Krishnamurthy proposed a classification of blogs into four basic types according to two dimensions: personal vs. topical, and individual vs. community. His schematic representation is reproduced as Figure 1.

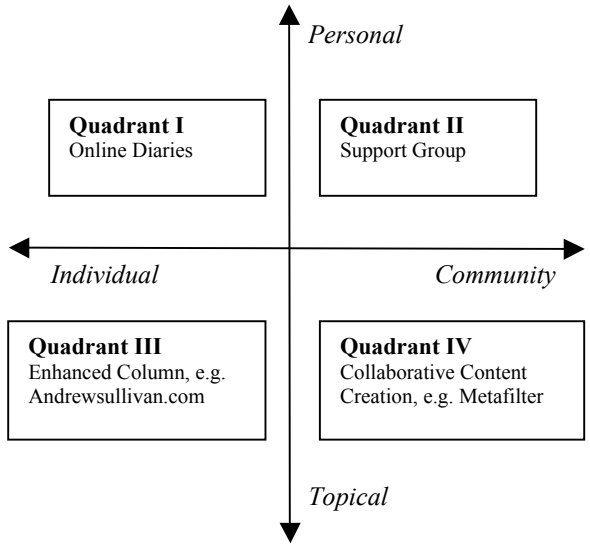

Figure 1. Types of blogs [21]

The community blog analyzed by Krishnamurthy (Metafilter) falls into quadrant IV. The personal journals found on LiveJournal.com are examples of quadrant I. Halavais' [17] study included examples from quadrant III, a type also known as 'filter' blogs because they select and provide commentary on information from the web. A group of friends collaboratively blogging about personal matters would constitute an example of quadrant II. In the present analysis, we identify blog sub-types from a random sample of sites that call themselves 'blogs' (but excluding LiveJournal and other online diary sites). Notably, in our sample, the types in quadrants I and III are well-represented, but few examples are found of quadrants II and IV. In addition, we find types not represented in Krishnamurthy's two-dimensional model (notably, the k-log).

3. Data

The present study is based on an analysis of a random sample of 203 blogs collected from March through May of 2003 using the randomizing feature of the blog-tracking website blo.gs. The blo.gs site was selected as the data source because it tracks a large number of blogs from diverse sources.[4] Lists of updated weblogs are imported by blo.gs every hour from antville.org, blogger.com, pitas.com, and weblogs.com. Thus blo.gs tracks currently active weblogs. In addition, blog owners can individually ping the blo.gs site when they update if they wish their blog to be listed on the site. At the time of this writing (3:00 p.m. CDT on September 26, 2003), the site claimed to be tracking 710,755 blogs, roughly 82% of the number of active blogs given on the NITLE blog census site [1].

Of the blogs selected randomly by the site during the data collection process, the vast majority were in English.[5] To create a coherent corpus, we excluded blogs in other languages,[6] photo and audio blogs that did not also contain a significant amount of text, and uses of blog software for non-blog purposes (e.g., community center events announcements; news; retail). We also excluded blogs that contained fewer than two entries, so that the practices of neophyte bloggers would not bias the sample at a time when new blogs are being created daily. Thus the blogs

[4] The blo.gs site defines a blog as "a type of web site (or page) that is organized much like a diary or journal—short nuggets of writing added regularly (or not) as a running commentary on almost any subject."

[5] The NITLE Blog Census site reported at 12:15 a.m. CDT on September 21, 2003 that 800,037 blogs (67% of all the blogs it had visited) appeared to be in English.

[6] Spanish, German, French, Portuguese, Russian and Arabic are some of the other languages in which blogs were found.

Figure 2. A sample blog

selected for analysis were established, English-language, text-based blogs. An estimated 60% of the randomly selected blogs met these combined criteria. As we were primarily interested in active blogs, we also excluded any blogs that had not been updated within the two weeks prior to data collection; this resulted in the elimination of several additional blogs.

Figure 2 shows the first screen of the home page of a typical blog from our corpus.[7] Its title, Lazy Gnome, appears in the header of the page, along with two other text descriptions: 'living at the bottom, lost in the middle' and 'Lazy Diary'. The remainder of the page consists of two columns, one devoted to the presentation of entries in reverse chronological order, and the other a sidebar containing links, an interactive poll, and an image with an updatable 'Iraq Body Count'. Although it is not visible in this image, the page also has a footer containing the author's full name.

4. Methodology

We employed content analysis [4] to identify and quantify structural and functional properties of the blogs in the corpus. The coding categories we employed were determined by multiple means.

First, to situate the genre within a community of users, we coded for demographic *characteristics of the blog authors*, to the extent that these could be determined from the blogs themselves. We were also interested in how much information about the blog author appears in the blogs, as a point of comparison with personal home pages, in which the site owner's identity is typically in focus.

Second, because purpose is a key criterion for defining a genre, we coded for the overall *purpose of the blog*: filter, personal journal, k-log, mixed purpose, or other. We did not code for the notebook (long, focused essay) type of blog described by Blood [6], in that her criterion for defining it in terms of entry length seemed problematic to us (entries can vary significantly in length within a single blog); nor did such a type emerge naturally from our data.

Exemplars of a genre also share structural features, thus *structural analysis of the blogs* was carried out. The structural features we selected for coding were adapted from previous content analytic research on Web genres (e.g., [3, 8]), such as number of links, images, presence of a search feature, and advertisements. In addition, to allow unique characteristics of the blog genre to emerge, we used a grounded theory approach [15] based on our initial inspection of the blogs in the corpus. This led us to add coding categories that we observed to be present

[7] This image is used with the permission of the blog creator and maintainer, Iain Cuthberson. The full blog can be viewed at: http://www.lazygnome.net.

in some of the blogs, but that otherwise were not described in the literature, such as the type of blogging software used, the ability for readers to post comments to entries, and the presence of a calendar, archives, and badges (small icons, often functioning as hypertext links, which represent the blogger's affiliation with a product—such as blogging software—or group of users). In our initial inspection, it appeared to us that comments, calendars, archives and badges were potentially useful indicators that a web site was a blog.

We also took into consideration popular definitions of blogs, incorporating means to measure the purported defining characteristics of blogs as articulated by Rebecca Blood and other bloggers: frequency of links, links to other blogs and news sources, numbers of actual comments on entries, and message length. The first two of these features were coded both for the blog home page and for the most recent entry in each blog.

Finally, to evaluate claims about the frequency with which blogs are updated, and determine the average age of blogs currently available on the Web, we coded for three types of *temporal information*: recency of update (in relation to the time of sampling); interval of update (between the most recent and the previous entry); and age of the blog, as determined from the date of the oldest posting on the site.

In all, 44 features were coded for each of the 203 blogs in the sample. All four authors participated in the coding: 10 blogs were coded by four coders and 40 others by pairs of coders in three successive cycles of refinement of the coding categories, until an 80% rate of inter-rater agreement was achieved. The remaining blogs were then coded by individual authors, and the composite results were checked by the first author before counts of each feature were made.

5. Results

The results of the content analysis support some previous claims made about blogs, but paint quite a different picture from them in other respects. In this section, we present quantitative summaries of the results of the analysis, as an empirical snapshot of the blogs in our sample. In the subsequent discussion, we interpret these results qualitatively and compare blogs with other genres of online and off-line communication.

5.1. Blog author characteristics

Upon initial examination, the characteristics of blog authors do not appear to be significantly different from the demographics of users of other public Internet communication protocols such as discussion forums [19] and personal homepages on the Web [12]. That is, they tend to be young adult males residing in the United States. Also as in other forms of Internet communication, the authors provide considerable information about their real-life identities, although some are more self-revealing than others. The main characteristics of the blog authors in the sample are summarized in Table 1.[8]

Table 1. Blog author characteristics

Characteristic	Frequency	Percentage
One author	196	90.8
Male	110	54.2
Adult (20 years or older)	115	59.6
Student	73	57.5
Located in USA	104	69.8
Name on first page (other than pseudonym)	127	67.6
Other personal information on first page	108	54.0
Graphical representation on first page	34	17.5

The overwhelming majority of blogs (90.8%) in the sample were created and maintained by a single individual. Gender can be determined in 91.2% of the blogs, with more bloggers being male (54.2%) than female (45.8%). In the 85.8% of blogs for which blogger age was apparent, roughly 60% were adult and 40% teenagers, although many of the adults indicate that they are in their early 20's. Perhaps not surprisingly given that we only examined blogs in English, nearly 70% of the 62% of bloggers whose geographic location could be determined are in the United States, followed by Singapore (7.4%), the UK (6.0%), Canada (3.4%) and Australia (2.7%). Blog author occupation was mentioned in 55% of the blogs; the most frequent occupation by far is student (secondary or tertiary level) at 57.5%; technology-related occupations such as Web developer, system administrator, and computer programmer come in second at 18.9%.

The above information in some cases had to be inferred from the content of entries or by following links elsewhere. In addition, many bloggers include explicit personal information on the first page of their blogs. A striking 92.2% provide a name: a full name (31.4%), a first name (36.2%), or a pseudonym (28.7%). More than half (54%) provide some other explicit personal information (e.g., age, occupation, geographic location), and another 16.2% link to such information elsewhere. Thus the identity of the author is apparent to some extent in most blogs. However, in contrast to personal home pages, a rather low percentage of sites—17.5%—display graphical

[8] In this and other tables that list the results of multiple coding categories, the percentage is calculated out of the total number of individuals or blogs for which the category was able to be coded, excluding 'unknowns' and other problematic instances.

representations of the author (including photos) on the first page, and only 10.9% link to such representations elsewhere. This is consistent with the relatively low frequency of images found on these blogs overall.

It is beyond the scope of the present paper to analyze variation within each category. However, it is interesting to note that the gender of blog authors varies according to age. Among bloggers of known gender classified as 'adult', 63% are male. Conversely, a majority of 'teen' bloggers of known gender are female (58%). These two sub-populations pattern differently with respect to blog purpose, as described below.

5.2. Purpose

Table 2 summarizes the distribution of blog types according to their primary purpose.[9] Although filter blogs in which authors link to and comment on the contents of other web sites are assumed by researchers, journalists and members of the blogging community to be the prototypical blog type, the blogs in our sample are overwhelmingly of the personal journal type (70.4%), in which authors report on their lives and inner thoughts and feelings (see the example in Figure 2). This result is all the more notable in that we excluded journal sites such as LiveJournal.com and Diaryland.com from our data collection, so that their popularity would not overshadow the other blogs in the sample. Even so, filter blogs account for only 12.6% of the sample, and k-logs are the least frequent at 3.0%. Mixed blogs (9.5%) combine the functions of two or more of the first three types, and 'Other' accounts for 4.5% of blogs which serve miscellaneous other functions.[10]

Table 2. Blog type by primary purpose

Type	Frequency	Percentage
Personal journal	140	70.4
Filter	25	12.6
K-log	6	3.0
Mixed	19	9.5
Other	9	4.5
	199	100

It is likely that k-logs are more common than these data indicate, in as much as they may be restricted to members of a specific community of practice [23], and thus may not be publicly available on the Web. By the same logic, personal journals should also be rare, and filters should be frequent, because of their presumably private and public natures, respectively. However, the opposite is the case.

How could the most common blog type (by far) be so overlooked and underrepresented in discussions about the nature of blogs? Blood [6] suggests one possible explanation: the personal journal blog is a newer type that is gaining ground at the expense of the earlier filter type, as blogging software becomes easier for anyone to use. This trend may have accelerated in the year or so since Blood offered this observation. At the same time, online web journals have been around since the mid-1990s, and thus the explanation that a "new" blog type appeared is not entirely satisfactory. It is also possible that journal-style blogs are not new, but that they are considered less interesting than filter-style blogs, and thus the latter have been selectively embraced over the former in popular descriptions of the phenomenon.

There are also gender and age differences in blog purpose. While bloggers of both genders and all ages create personal journals, females and teens create them somewhat more than do males and adults. Conversely, filter blogs, k-logs, and 'mixed' blogs are created almost exclusively by adult males.

This variation notwithstanding, on the whole, the blogs in this sample share a common purpose: to express the author's subjective, often intimate perspective on matters of interest to him or her. In the case of most blogs, the matters of interest concern the authors and their daily lives.

5.3. Temporal measures

The blogs in the present sample had all been updated within two weeks prior to collection, according to our sampling criteria. In fact, in a majority of cases the most recent update was less than one day old, and the mean number of days since last update was only 2.2 for the entire sample. However, this number could reflect a sampling bias on the part of the blo.gs site, which tracks blogs when they are updated. A more representative measure of the frequency with which these blogs are updated is the mean number of days between the most recent and the next-most recent entry, or 5.0 days for the sample as a whole. The mode for this measure is one day, lending some support to Blood's claim that "most" blogs are updated daily, although the range of update frequency is wide (0-63 days). On the whole, the blogs in this sample appear to be quite actively maintained.

Moreover, their authors maintain this level of activity over an extended period. Our sampling method only required that a blog have at least two entries, and we found several that had been started on

[9] Four blogs could not be classified according to purpose, thus the total number of blogs analyzed in this category is 199.

[10] These include a blog consisting entirely of the author's poetry (mostly rough drafts); a blog devoted to song lyrics that the author can't get out of her head; a blog containing notes for a class on urban planning; a blog archiving quotes about a film actor; and blogs that serve as conversation boards for two or more authors.

the same day we sampled. However, the average blog in the sample is considerably older—163 days (five and a half months)—and the oldest blog had been in existence for 990 days (two years and nine months), with 16.6% being more than one year old, and 5% being more than two years old. These results (shown in Table 3) suggest that maintaining a blog represents a non-trivial time commitment for many authors.

Table 3. Temporal measures

Measure	Mean (days)	Mode (days)	Range (days)
Recency of update at time of data collection	2.2	0	0-11
Interval between two sequential entries	5.0	1	0-63
Age of blog	163.0	n/a[11]	0-990

Personal journal blogs are equally or more frequent than other blog types for every six-month period represented in the sample, starting in the second half of 2000. Their lead has steadily increased over time, with a sharp increase in frequency in the first half of 2003. This effect may have been partially triggered by developments in blogging software, as discussed below.

5.4. Structural characteristics

In this section, the results for the coding of structural characteristics are presented for two units of analysis: the blog home page (first page of each site) as a whole, and the most recent entry in each blog.

5.4.1. Home page. Table 4 lists the frequencies of structural features hypothesized to be characteristic of blogs, as coded for the home page of each blog. The home pages of the blogs in the sample differ from those of personal home pages in several respects. Blogs appear to be less likely to contain a guest book, a search function, and advertisements than are personal home pages [3]. Blogs are relatively image-poor as well, compared to other genres which make greater use of the multimedia potential of the Web. At the same time, blogs exhibit features that personal home pages lack. Archives (links in the sidebar to older entries; 73.5%) and badges (small icons in the sidebar, header or footer advertising a product or group affiliation; 69%) are found in a clear majority of blogs. These are not, to our knowledge, characteristic of any other Web genre, at least not in combination. In contrast, while a calendar in the sidebar was perceived by us initially to be a typical blog feature, it turned out to be less frequent than we had thought (13%), as did the feature of allowing readers to comment on entries (43%).

Table 4. Structural features

Feature	Frequency	Percentage
Archives	139	73.5
Badges	138	69.0
Images	133	58.6
Comments allowed	85	43.0
Link to email blog author	63	31.3
Ads	48	25.1
Search function	35	18.5
Calendar	25	13.0
Guest book	9	4.5

The presence or absence of the above features is determined in part by the blog creation software used by the blog author. Blog software imposes a one- to three-column format and the display of entries in reverse chronological order. In addition, it incorporates defaults (such as comments on entries, archives, the presence of a badge for the blog software) that inexperienced bloggers tend to preserve, if only because they do not know how to change them. Table 5 shows the breakdown of the sample according to the brand of software used to create the blog. The most common blog software in our sample is Blogger, by Pyra. Blogger's popularity is accounted for by the fact that it is free, easy to use, and requires little of the user.[12] The predominance of Blogger blogs in the sample biases the results in relation to particular structural features: for example, Blogger by default does not allow comments, and does not provide a calendar in the sidebar. At the same time, in as much as this is the software most users are choosing, the results fairly represent the "average" blog at the present time.

Table 5. Blog software used

Software name	Frequency	Percentage
Blogger	122	63.2
Movable Type	22	11.4
Pitas	13	6.7
Radio Userland	6	3.1
All others combined	14	7.3
Unknown	16	8.3
	193	100

Next we consider patterns of linking from the home page of the blogs. We coded for the presence of links (yes or no) of different types, classified in terms of their destination. These results are shown in Table 6.

[11] No clear central tendency for age of blog can be observed because the data are sparsely distributed over a wide range.

[12] For example, it does not require the user to host the blog on their own server, but rather provides free space on a public server.

Table 6. Links from home page

Destination	Frequency	Percentage
To websites by others	117	53.7
To other blogs	106	51.2
To news sites	74	36.1
To websites created by or about self	33	17.2
To webrings	9	4.8

Most of the links lead to (non-blog) websites created by other than the blog author, although the number of blogs that do so (53.7%) is lower than one might expect, given that blogs are often defined in terms of linking to content elsewhere on the Web. Only about half of the blogs (51.2%) link to other blogs, and fewer yet (36.1%) link to news sites, in contrast to the popular characterization of blogs as heavily interlinked and oriented towards external events. Nor does it appear that blogs participate in webrings of bloggers; instead, 'blogrolls', or lists of blogs the author claims to read regularly, are included in the category 'links to other blogs'. We also found blogs (N=17) with no external links of any kind on their homepage, not even email contact links or badges.[13] In general, although some blogs contain many links, the extent to which blogs link to other content is not as great as it is popularly represented in previous characterizations of blogs.

5.4.2. Most recent entry. The heart of a blog is its entries; these are the 'frequently updated content' that readers visit the site on a regular basis to read. In order to characterize blog entries, we coded the most recent (i.e., top) entry in each blog in the sample in some detail. These results are presented in Tables 7-9 under the categories 'entry headers and footers', 'entry body features' and 'entry body text'.

Table 7 describes the types of information contained in the header and the footer of the entry. This is determined by the software used, and is consistent across all entries within a blog. The most frequent information contained in the entry header is the date and title of the entry; the footer typically contains the time of posting, the author's name (or pseudonym), and links to a permanent copy of the entry stored elsewhere on the site ('permalinks'). A link to add or read comments, if present, usually appears in the footer.

The last line of Table 7 indicates that the average entry in our sample received .3 comments, and the majority of entries received none. The entry that received the *highest* number of comments (N=6) still received fewer than the *average* number of comments reported by Krishnamurthy for a typical Metafilter entry (N=17). For purposes of comparison, we also made similar counts of comments for the oldest entry on the home page of each blog, on the hypothesis that the newest entries had not yet had sufficient time to collect comments. To our surprise, the results for the oldest entries were nearly identical: mean = .3 , mode = 0, range = 0-7 comments. It appears that entries do not continue to collect comments over time, but rather are only commented on while they are new. In general, the evidence of readers commenting on blog entries is less than previous claims about blog interactivity and community had led us to expect.

Table 7. Entry header and footer

Information contained		Frequency	Percentage
Header		331	99.5
	date	176	93.6
	title	84	44.7
	time	30	16.0
	author's name	21	11.2
Average number of header features per blog		*1.8*	
Footer		481	92.0
	time	148	78.7
	author's name	121	64.4
	internal links	109	58.0
	comments	61	32.4
	date	22	10.6
Average number of footer features per blog		*2.6*	
Number of comments per entry	mean .3	mode 0	range 0-6

As important as comments are in the popular perception of blogs, links within entries are even more important. Blood [6], for example, defines a blog entry as centered around a link to external content. Thus it is striking that fewer than one-third of blog entries (31.8%) contain any links at all, and that the central tendency is for an entry to have none (see Table 8). The mean number of links per entry is .65, as compared with 1.89 as reported for Metafilter by Krishnamurthy [21]. When links are present, moreover, they rarely lead to news sites or other blogs, although they do lead to other websites. It is theoretically possible to include as many links as one wants in any blog entry; the choice is the author's, not the software's. The low incidence of links in entries appears in part to be a reflection of the prevalence of personal journal type blogs in the sample.[14]

The blog entries analyzed contain few images (9.2%); however, this may be due in part to our

[13] If we exclude badges from the count, the number of blog home pages with no other links rises to 62 (30.5%) of the blogs in the corpus.

[14] Personal journal type blogs are least likely to allow comments, and not just because 68% use Blogger software, which has 'no comments' as a default. Of the 95 Blogger personal journals found, only 24% allow comments, as compared with 69.2% of the filter blogs created with Blogger (N=13). All three k-logs created with Blogger allow comments.

sampling procedure, which led us to reject any blog that did not contain text in the most recent entry. It is our impression that a random sampling of blog entries without regard to this criterion would reveal a somewhat higher incidence of images.

Table 8. Entry body features

Feature	Frequency	Percentage[15]
Images	18	9.2
Links		
to websites by others	54	27.7
to news sites	16	8.2
to other blogs	13	6.7
to internal to blog	6	3.1
to websites created by or about self	4	2.1
Number of links per entry mean .65	mode 0	range 0-11

The final set of measures involves structure at the level of the text of the blog entry; these are summarized in Table 9. At 210.4 words, the average blog entry is somewhat shorter than an email posting to an academic discussion list [18]. Its sentences, at 13.2 words, are three words shorter than those of private email messages exchanged in a university setting [9]. Quoted content of any kind (regardless of whether enclosed in quotation marks) accounts for only 18 words per message on average, and sentences in quotes are shorter (7.8 words)—a reflection perhaps of a higher incidence of sentence fragments (headings, etc.) and pithy sayings in quoted than in non-quoted content.

Table 9. Entry body text measures

Measure	Total	Avg	Range
Words	42930	210.4	1-1262
Sentences or fragments	3260	16.0	1-117
Words per sentence		13.2	
Paragraphs	709	3.5	0 - 21
Words in quotations	3681	18.0	0 - 430
Quoted sentences/fragments	468	2.3	0 - 40
Quoted words per sentence		7.9	

The following example of an entry from a blog in our corpus is provided to illustrate the patterns summarized above. (The home page for this blog, entitled 'Lazy Gnome', was shown in Figure 2.) This message is shorter than average at 99 words and 10 sentences (counting each link as a sentence fragment), and it contains more links (four—in this case to content (photos) produced by the blog author) than average, but is within the normal range of blog entries analyzed in the corpus. The entry header provides the date, time, and a title ('trigger happy hippy with a Canon AE-1'), and the footer contains a comment link, the question mark after the word 'Comment' and the absence of a number indicating that no responses have yet been posted to this entry.

Friday, 13th June 2003
3.08pm - trigger happy hippy with a Canon AE-1

If I go away, I take my camera. Standard practice.
So, for your viewing displeasure, there are 4 new gallerys to view:

watery times
my 1st b+w shoot
Swanage area + 1
sea and air

First two are from my latest trip to Edinburgh to see my little sweetie. The second two were taken from my 4 day trip to the south coast with my parents in their campervan (I had a 4 man tent all to myself!).

Seeing as the last 'family holiday' I can remember was about 8 years ago, it was a real treat for me.

Comment ?

The content of this entry is typical of that for a journal-style blog, with reference to the author's recent activities involving his girlfriend and his family. In a link in the sidebar entitled 'about the boy' (highlighted in Figure 2), the author gives his full name and indicates that he is 24 years old, works as a systems administrator, and resides in the UK.

6. Discussion and Conclusion

6.1. Antecedents of blogs

We return now to the question of the blog's origins. Is the blog uniquely digital, "native to the Web", as Blood [5] claims? Blood proposed that blogs are directly descended from 'what's new?' or 'cool links' pages that provided lists of links ('hotlists') to sites deemed by the site creator to be of interest in the early days of the Web. While it is beyond the scope of the present paper to conduct a historical analysis of blogs, we believe that our findings, together with commonsense reasoning, can shed some light on this question.

Blood's claim about the origins of the blog is based on the assumption that blogs are link-centered filters of Web content. Our findings show that this assumption misrepresents most blogs at the present time. Blood herself recognizes that journal-type blogs were more common even at the time she wrote. Personal journal blogs contain few links and do not focus on Web content; thus it is unlikely that they trace their genesis to lists of links. Rather, as

[15] Of most recent entries coded (N=195).

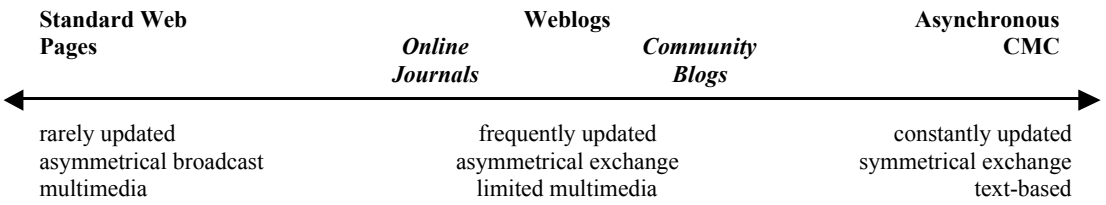

Figure 3. Weblogs on a continuum between standard Web pages and CMC

suggested by Blood herself, they resemble the online journals found since the mid-1990s, a genre which is itself reproduced from the centuries-old genre of handwritten diaries. Thus the journal blog can be seen to have off-line antecedents.

Similarly, the filter blog type, especially when used to self-publish commentary on current events, has an off-line precedent in editorials and letters to the editor in print newspapers—with which this type of blog is often compared in the popular press [22]. K-logs functionally resemble hand-written project journals in which a researcher or project group makes observations, records relevant references, and so forth about a particular knowledge domain. Less common uses of blogs can also be related to off-line genres: travel blogs resemble travelogues and photo albums; memory blogs in which the author keeps track of information for later use function in some respects like post-it notes to oneself; and blogs created for conversation between two or more individuals resemble email exchanges, which in turn have taken over the function of personal letters.

Moreover, the blog shares similarities with other digital genres, including the personal home page, which prior to the creation of blogs, was the preferred way to present oneself and one's views on the Web. Blogs convey demographic information about their authors primarily in sidebars and through links, reserving the entry column for the author's musings of the moment, but their overall functionality is similar. All of this suggests that blogs, rather than having a single source, are in fact a hybrid of existing genres, rendered unique by the particular features of the source genres they adapt, and by their particular technological affordances.

6.2. Blogs in the genre ecology of the Internet

As noted above, blogs have features in common with personal homepages. Blogs also share features with asynchronous discussion forums, a text-based form of interactive computer-mediated communication. This observation leads us to a final argument for the hybrid nature of blogs, and a closing speculation.

Interactive text-based CMC is generally held to be a fundamentally different type of Internet communication from the static, single author, multimedia HTML documents that are the standard means of communication on the World Wide Web. In recent years, efforts have been made from both sides to bridge this gap (e.g., HTML mail and encoding schemes for multimedia attachments from the CMC side, and links to chat and discussion forums in HTML documents from the Web side). However, HTML-enhanced CMC and CMC-enhanced Web pages still remain essentially different technologies—they do not meet in the middle. Weblogs, in contrast, bridge this technological gap along several dimensions. This can be represented schematically as a continuum, as shown in Figure 3.

The three dimensions of comparison in Figure 3 are frequency of update, symmetry of communicative exchange, and multimodality. Weblogs are situated at an intermediate point between standard Web pages and asynchronous CMC along each of these dimensions. Web pages may be updated only once every few months, but weblogs are typically updated several times a week, and discussion forums are updated every time a conversational participant posts a message. Author and reader roles in web pages are highly asymmetrical, in contrast with the fully symmetrical give and take of unmoderated discussion forums; blogs allow limited exchanges (in the form of comments), while according blog author and readers asymmetrical communication rights—the author retains ultimate control over the blog's content. Finally, blogs can incorporate multimedia elements as desired, like Web pages, but tend to preserve a mostly textual focus, like CMC. That the relationship is a continuum, rather than three discrete points, is further suggested by the placement of two genres closely related to blogs, but that we excluded from the present study in order to examine blogs in a more focused manner: online journal sites (such as LiveJournal.com) and 'community blog' sites (such as Metafilter and Slashdot). Journal sites, with their lesser interactivity, are closer to standard Web pages than are blogs. Community sites are closer to online discussion groups than are individually-maintained blogs in their frequency of activity and exchange of messages among multiple participants. The two major types of blogs defined by purpose in the present study could also be placed along the contin-

uum, with journal-style blogs closer to online journals, and filter-style blogs closer to community blogs.[16]

The "intermediate" characteristics of blogs make them attractive to users. In particular, they allow authors to experience social interaction while giving them control over the communication space. Combined with the unprecedented opportunity blogs provide for ordinary people to self-express publicly, these characteristics suggest that blogs will continue to grow in popularity in the future, and that they will be put to increasingly diverse uses.

Ultimately, we believe that blogs have the potential to change the way we think about the Web and about CMC, by rendering obsolete any hard-and-fast distinction between the two. At the root of this transformative potential are two technical enhancements provided by weblog software, neither of them revolutionary in itself. By enabling faster and easier content modification that does not require knowledge of HTML, blogs can be used by almost anyone, and be responsive to people's daily needs. Second, by enabling readers to post comments, blog soft-ware makes Web pages truly interactive, even if that interactive potential has yet to be fully exploited. Moreover, the flexible, hybrid nature of the blog format means that it can express a wide range of genres, in accordance with the communicative needs of its users. This analysis thus illustrates how technological changes, even incremental ones, can have wider consequences. One of those consequences, in the case of weblogs, is the potential to reshape the genre ecology of the Internet.

References

[1] (2003). NITLE Blog Census. http://www.blogcensus.net/?page=Home

[2] Arnold, J. & Miller, H. (1999). Gender and web home pages. http://ess.ntu.ac.uk/miller/cyberpsych/cal99.htm

[3] Bates, M. J. & Lu, S. (1997). An exploratory profile of personal home pages: Content, design, metaphors. *Online and CD Review, 21,* 331-340.

[4] Bauer, M. W. (2000). Classical content analysis: A review. In M. W. Bauer & G. Gaskell (Eds.), *Qualitative Researching with Text, Image, and Sound: A Practical Handbook* (pp. 131-151). London: Sage Publications.

[5] Blood, R. (2002). Introduction. In J. Rodzvilla (Ed.), *We've Got Blog: How Weblogs are Changing Our Culture* (pp. ix-xiii). Cambridge MA: Perseus Publishing.

[6] Blood, R. (2002). *The Weblog Handbook: Practical Advice on Creating and Maintaining Your Blog*. Cambridge MA: Perseus Publishing.

[7] Cavanaugh, T. (2002). Let slip the blogs of war. In J. Rodzvilla (Ed.), *We've Got Blog: How Weblogs are Changing Our Culture* (pp. 188-197). Cambridge, MA: Perseus.

[8] Chandler, D. (1998). Personal homepages and the construction of identities on the Web. http://www.aber.ac.uk/media/Documents/short/webident.html

[9] Cho, N. (2003). Linguistic features of electronic mail. In S. C. Herring (Ed.), *Computer-Mediated Conversation* (Cresskill NJ: Hampton Press.

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

[11] Dillon, A. & Gushrowski, B.A. (2000). Genre and the Web: Is the personal home page the first uniquely digital genre? *Journal of The American Society for Information Science, 51,* 202-205.

[12] Döring, N. (2002). Personal home pages on the Web: A review of research. *Journal of Computer-Mediated Communication, 7.*

[13] Erickson, T. (2000). Making sense of computer-mediated communication (CMC): Conversations as genres, CMC systems as genre ecologies. In *Proceedings of the 33rd Hawaii International Conference on System Sciences*.

[14] Festa, P. (2003). Blogging comes to Harvard. http://news.com.com/2008-1082-985714.html?tag=fd_nc_1

[15] Glaser, B. & Strauss, A.L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine Publishing.

[16] Ha, L. & James, E.L. (1998). Interactivity reexamined: A baseline analysis of early business web sites. *Journal of Broadcasting and Electronic Media, 42,* 457-474.

[17] Halavais, A. (2002). Blogs and the "social weather". Maastricht, The Netherlands: Internet Research 3.0.

[18] Herring, S.C. (1996). Two variants of an electronic message schema. In S.C. Herring (Ed.), *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives* (pp.81-106). Amsterdam: Benjamins.

[19] Herring, S.C. (2003). Gender and power in online communication. In J. Holmes & M. Meyerhoff (Eds.), *The Handbook of Language and Gender*. Oxford: Blackwell.

[20] Hourihan, M. (2002). What we're doing when we blog. http://www.oreillynet.com/pub/a/javascript/2002/06/13/megnut.html

[21] Krishnamurthy, S. (2002). The Multidimensionality of Blog Conversations: The Virtual Enactment of September 11. In Maastricht, The Netherlands: Internet Research 3.0.

[22] Lasica, J. D. (2001). Blogging as a form of journalism. *USC Annenberg Online Journalism Review*.

[23] Lave, J. & Wenger, E. (1991). *Situated Learning. Legitimate peripheral participation*. Cambridge: University of Cambridge Press.

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

[25] Shepard, M. & Watters, C.R. (1998). The evolution of cybergenres. In: *Proceedings of the 31st Annual Hawaii International Conference on System Sciences* (pp. 97-109).

[26] Swales, J. (1990). *Genre Analysis: English in Academic Settings*. Cambridge University Press.

[27] Winer, D. (2002). The history of weblogs. http://newhome.weblogs.com/historyOfWeblogs

[28] Yates, J. & Orlikowski, W.J. (1991). Genres of organizational communication: An approach to studying communication and media. MIT Sloan School of Management.

[16] It is not immediately apparent where k-logs should be situated along this continuum. Additional dimensions may be required to distinguish it from the other blog types.