# Weblogs as a bridging genre

Susan C. Herring, Lois Ann Scheidt, Elijah Wright and Sabrina Bonus

*School of Library and Information Science, Indiana University, Bloomington, Indiana, USA*

## Abstract

**Purpose** – Aims to describe systematically the characteristics of weblogs (blogs) – frequently modified web pages in which dated entries are listed in reverse chronological sequence and which are the latest genre of internet communication to attain widespread popularity.

**Design/methodology/approach** – This paper presents the results of a quantitative content analysis of 203 randomly selected blogs, comparing the empirically observable features of the corpus with popular claims about the nature of blogs, and finding them to differ in a number of respects.

**Findings** – 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.

**Originality/value** – Based on the profile generated by the empirical analysis, considers the likely antecedents of the blog genre, situates it with respect to the dominant forms of digital communication on the internet today, and suggests possible developments of the use of blogs over time in response to changes in user behavior, technology, and the broader ecology of internet genres.

**Keywords** Communication, Communication technologies, Worldwide web, Internet

**Paper type** Research paper

## 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 web site created by Tim Berners-Lee in 1991 (Winer, 1999), what is commonly recognized as the present-day format first appeared in 1996[1], and the term weblog was first applied to it in 1997[2]. Blogging as an online activity has been increasing exponentially since mid-1999, enabled by the release of the first free blogging software (Pitas, Blogger, and Groksoup; Blood, 2000), and fueled by reports from the mainstream media of the grassroots power of blogs as alternative news sources, especially in the aftermath of 9/11/2001 and during the 2003 US-led invasion of Iraq. Current estimates place the number of sites calling themselves blogs at over 2.1 million, of which 66 percent are actively maintained (NITLE Blog Census, 2004). Moreover, as blogging software becomes easier to use, the number of bloggers continues to increase by the day. 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. e-mail; 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 (Lasica, 2001). Educators and business

Information Technology & People
Vol. 18 No. 2, 2005
pp. 142-171

0959-3845
DOI 10.1108/09593840510601513

people see them as environments for knowledge sharing (Festa, 2003; Ray, 2003); 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 (Blood, 2002a). According to Blood (2002a), blogging makes people more thoughtful and articulate observers of the world around them. These effects are purportedly brought about by the technical ability that blogging software affords users to update web pages rapidly and easily.

As yet, however, little empirical research has sought to determine what blogs are actually like or what uses of blogs are most common. In this study, 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 (cf. Erickson, 2000). Our primary goal in so doing is to provide an empirical snapshot of the weblog in its present stage, as a 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 other genres within a functional domain. Our analysis suggests that the blog is neither fundamentally new nor unique, but that it – along with other emergent genres expressed through 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, blurring the traditional distinction between these two dominant internet paradigms, and potentially contributing to its future breakdown.

## Background

### *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 (1992) in their analysis of organizational uses of e-mail. They analyze e-mail as a continuation of the memorandum genre, pointing out that e-mail has reshaped the genre through its technical format and norms of use. Yates and Orlikowski's research, along with much of the subsequent research it has inspired, draws on traditional models of genre from rhetoric, especially Miller's (1984) definition of a genre as "typified rhetorical action based in recurrent situation." Thus for Yates and Orlikowski (1992), genre analysis is an exercise in classification of "typified acts of communication" based on their form and substance. Similarly, Swales (1990) 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 the name "weblog" and its variant "blog" are recognized by internet users, and – as we will show – weblogs tend to exhibit common structures and substance.

Orlikowski and Yates (1994) further observe that genres exist, and are defined and modified, in relation to other genres in use within a shared domain. They employ the term "genre repertoire" to refer to "[t]he set of genres that community members use (and don't use) to conduct their interaction," and maintain that:

> [a]n in-depth examination of a genre repertoire explores the nature and source of genres that are recognized and accepted by a community of practice as legitimate forms of working and interacting, and helps to explain when, how, and why established norms and practices shift over time.

Adopting this perspective, we propose that considering weblogs in relation to other genres – both electronic and traditional – as part of a larger communicative domain helps to explain their nature, norms of use, functional antecedents, and evolutionary trajectory. This, in turn, can help us to assess the likely future impact of weblogs.

*Web genres*

Recent years have seen a growing interest in the identification and description of genres on the World Wide Web (e.g. Crowston and Williams, 2000; Shepherd and Watters, 1998). Studies have characterized genres as diverse as health sites (McMillan, 1999), presidential candidate sites (Foot and Schneider, 2002), political satire sites (Warnick, 1998), and Holocaust denial sites (Polger, 2003), all of which arguably have off-line antecedents. According to Crowston and Williams (2000), such genres are "reproduced."

The most-studied web genre thus far, however, has been the personal home page. Crowston and Williams (2000) cite personal home pages as an example of an "emergent" web genre, i.e. one that did not exist prior to the creation of the web. Similarly, Dillon and Gushrowski (2000) assert that the personal home page as the first uniquely web-based genre. Bates and Lu (1997), Chandler (1998), and Dillon and Gushrowski (2000) identify structural characteristics of personal home pages hypothesized to define them as a genre, 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 (for a useful overview of this literature, see Döring, 2002). In other content analyses, Arnold and Miller (1999) identify gender differences in the structure and content of home pages created by academic professionals, and Ha and James (1998) find a relatively low frequency of "interactivity" features in the home pages of business web sites.

Like weblogs, personal home pages are typically created and maintained by a single individual, and their content tends to focus on the creator or his/her interests. Chandler (1998) compares personal home pages with "the bedroom walls of young people in the West, with their diverse arrays of graphics and text in the form of posters, postcards, snapshots, sports insignia and so on." This observation, along with the findings of Arnold and Miller (1999) and Ha and James (1998) mentioned above, illustrate that older practices from related off-line genres may carry over into web genres, making them at least partially "reproduced" in the sense of Crowston and Williams (2000). 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 communication genres.

*Previous blog research*

The earliest descriptions of the blog genre have as their source blog authors themselves, a number of whom have reflected on their practices in blog entries, online media reports, and less commonly, print publications (e.g. Blood 2000, 2002a, b;

Rodzvilla, 2002; Winer, 1999, 2001). 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 (Hourihan, 2002). Updates should be frequent: according to Blood (2002a, p. 9), "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 (Blood, 2002a, b). This type of blog, in which the blogger "pre-surfs" the web and directs readers to selected content, is known as a filter. Blood (2002a) 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 about the blogger (the blogger's activities and internal states); notebooks may contain either external or personal content, and are distinguished by longer, focused essays. 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 (Cavanaugh, 2002), but some blogs allow readers to post comments to individual entries, giving rise to "conversational" exchanges on the blog itself (Blood, 2002a). Blood claims that social interactivity is highest in journal-type blogs.

Although empirical research on blog structure and content is thus far limited, the results of two quantitative studies bear on the general claims advanced about blogs above. Halavais (2002) found that popular news stories – external content – were the most common topics of discussion in a random sample of 125 blogs. Krishnamurthy (2002) analyzed patterns of posting to a community news blog, Metafilter, 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.

The community blog analyzed by Krishnamurthy (Metafilter) falls into quadrant IV. The personal journals found on LiveJournal.com are examples of the type represented by quadrant I. Halavais's (2002) 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 the blog type represented in 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).

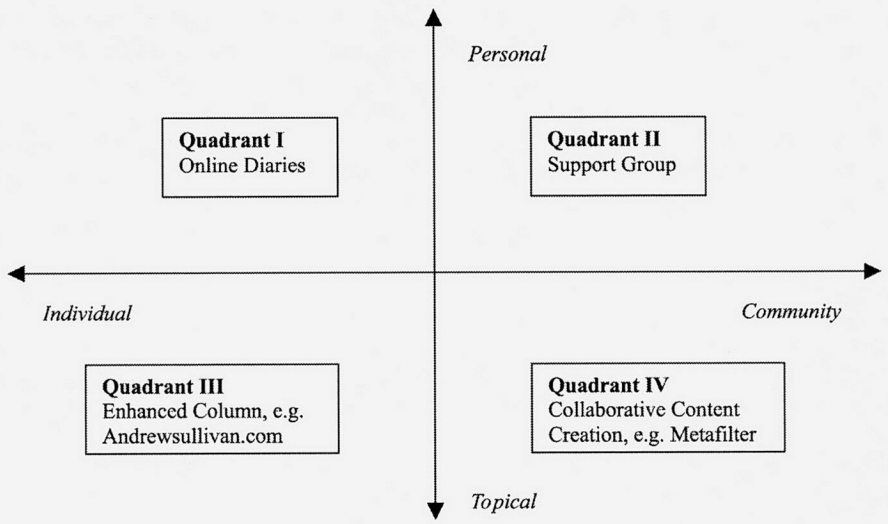



**Figure 1.**
Types of blogs

**Source:** Krishnamurthy (2002)

## Study description

*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 web site 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 beginning of the period of our data collection, the site claimed to be tracking approximately 400,000 blogs; by the end, that number had doubled.

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 were being created daily. Thus the blogs selected for analysis were established, English-language, text-based blogs. An estimated 60 percent 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.

In determining how best to collect an interpretable, random sample of weblogs, we made the decision to exclude online journals hosted on the popular services LiveJournal, DiaryLand, and Xanga, which numbered over one million at the time. There were two reasons for this. First, at the time of our data collection, those services

self-identified as hosting "journals" or "diaries" rather than "weblogs." Second, blog tracking services such as blo.gs did not index LiveJournal, etc. journals, a further indication of their less-than-prototypical weblog status. Although both of these situations have since changed, at the time we deemed it best to leave aside online journal hosting sites in the interest of sampling only sites that could be described uncontroversially as weblogs.

*Methodology*

Consistent with other empirical analyses of web genres (e.g. Bates and Lu, 1997; Ha and James, 1998; McMillan, 1999), we employed methods of content analysis (Bauer, 2000) to identify and quantify structural and functional properties of the blogs in the corpus. The coding categories we employed were designed to provide an overall characterization of the genre as well as to test popular claims about weblogs, and 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. Purpose was determined from the nature of the content in entries posted on the first page of each blog and categorized according to a modified version of Blood's (2002a) three blog types: filters, personal journals, and k-logs[7]. Filter blogs were operationalized as primarily containing observations and evaluations of external, typically public events; an example of a filter is given in Figure 2. Personal journals were defined as primarily reporting events in the blogger's life and the blogger's internal states and/or reflections; see Figure 3 for an example. We coded a blog as a k-log in cases where its primary content was information and observations focused around a(n external) topic, project or product; an example of a k-log from our sample is shown in Figure 4.

A blog was coded as one of the three types described above only if a clear majority of its entries exhibited the requisite content; blogs with a roughly balanced mix of content were coded as "mixed," and blogs whose content indicated a purpose other than those described above were coded as "other."

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. Bates and Lu, 1997; Chandler, 1998), 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 (Glaser and Strauss, 1967) based on our initial inspection of the blogs in the corpus. This led us to add coding categories that we observed to be present in some of the blogs, but that otherwise were not described in the web genre 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

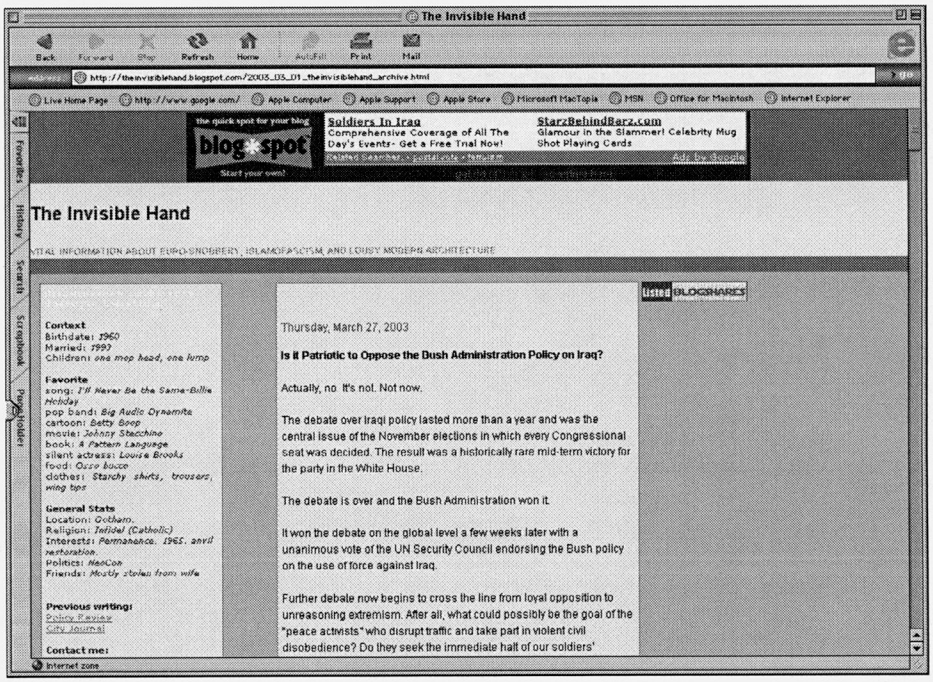



**Figure 2.**
A filter blog (description in subtitle: "Vital information about Euro-snobbery, Islamofascism, and lousy modern architecture")

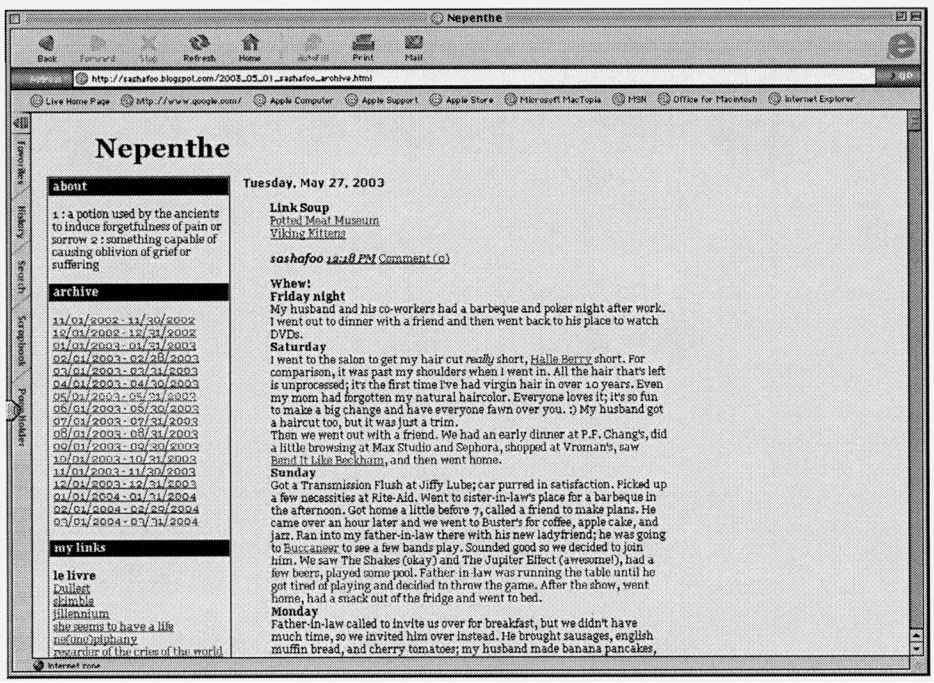



**Figure 3.**
A personal journal blog (first line of first entry: "My husband and his co-workers had a barbeque and poker night after work.")

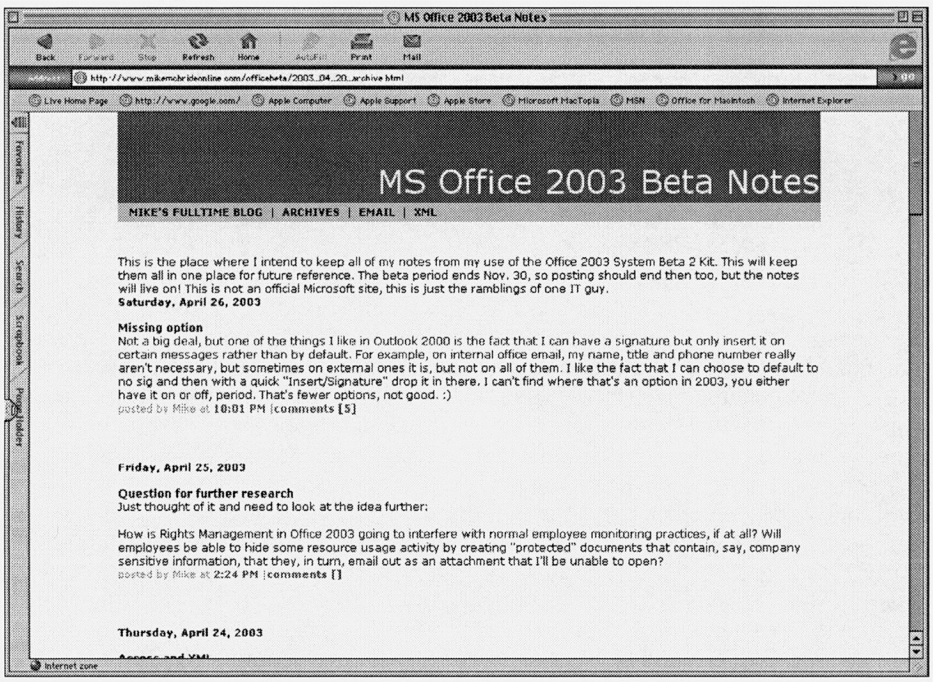



**Figure 4.**
A k-log (first sentence: "This is the place where I intend to keep all of my notes from my use of the Office 2003 System Beta 2 Kit.")

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 Blood (2002a) 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 newest and previous entry); and age of the blog, as determined from the date of the oldest posting on the site.

In all, 55 features were coded for each of the 203 blogs in the sample (see the Appendix). All four authors participated in the coding: ten 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 percent 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.

## 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.

*Blog author characteristics*

Upon initial examination, the characteristics of blog authors resemble the demographics of users of other public internet communication protocols such as discussion forums (cf. Herring, 2003) and personal homepages on the web (Döring, 2002). That is, they tend to be young, adult males residing in the USA. 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 I[8].

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

In some cases the above information 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 percent provide a name: a full name (31.4 percent), a first name (36.2 percent), or a pseudonym (28.7 percent). More than half (54 percent) provide some other explicit personal information (e.g. age, occupation, geographic location), and another 16.2 percent 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, 35 percent of which were found to contain photographs of the author by Bates and Lu (1997), only 17.5 percent of the blogs in our sample display graphical representations of the author of any kind on the

**Table I.**
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 |

first page, and only 10.9 percent 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 classified as "adult" for whom gender is known, 63 percent are male. Conversely, a majority of "teen" bloggers for whom gender is known are female (58 percent). These two sub-populations pattern differently with respect to blog purpose, as described below (see also Herring *et al.*, 2004).

*Purpose*

Table II 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 percent), in which authors report on their lives and inner thoughts and feelings. This result is all the more notable in that we excluded journal sites such as LiveJournal.com and Diaryland.com from our data collection; had they been included, the frequency of personal journal blogs would have been even higher. Even with this exclusion, filter blogs account for only 12.6 percent of the sample, and k-logs are the least frequent at 3.0 percent. 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 (Lave and Wenger, 1991), 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. Mixed blogs (9.5 percent) combine the functions of two or more of the first three types (most commonly filter and personal journal), and "Other" accounts for 4.5 percent of blogs which serve miscellaneous other functions[10].

The blog types found in our sample differ somewhat from those in Krishnamurty's blog classification scheme presented in Figure 1. K-logs are not included in Krishnamurty's scheme, even though they are more frequent than "support group" blogs (Krishnamurty's quadrant II), of which we found none in our sample. This comparison suggests that additional dimensions may be needed to classify currently available blog types, such as a dimension of topical focus versus topical heterogeneity.

How could the most common blog type (by far) be so overlooked and underrepresented in discussions about the nature of blogs? Blood (2002a) suggests one possible explanation: the personal journal blog is a newer type that is gaining

| 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 |
| Total | 199 | 100 |

**Table II.**
Blog type by primary purpose

ground at the expense of the earlier filter type, as blogging software becomes easier for anyone to use. This trend may have accelerated 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. (We return to the question of the antecedents of the blog in the following section.) Journal-style blogs may also be considered less interesting than filter-style blogs, and thus the latter may have been selectively promoted over the former in popular discourses about weblogs (Herring *et al.*, 2004).

There are also gender and age differences with respect to 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[11].

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.

*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 a minimum of two entries, and we found several that had been started on the same day on which 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 percent being more than one year old, and 5 percent being more than two years old. These results (summarized in Table III) suggest that maintaining a blog represents a considerable time commitment for many authors.

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

| 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[a] | 0-990 |

**Table III.**
Temporal measures

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

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.

*Structural characteristics*

A genre is defined in part by structural features typically shared by its exemplars. 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.

*Home page*. 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 (cf. Bates and Lu, 1997). Blogs are relatively image-poor as well, compared to personal home pages, nearly 80 percent of which were found to contain graphics in Bates and Lu's sample. At the same time, blogs exhibit features that personal home pages lack. 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 percent), as did the feature of allowing readers to comment directly on entries (43 percent). In contrast, archives (links in the sidebar to older entries; 73.5 percent) and badges (small icons in the sidebar, header or footer advertising a product or group affiliation; 69 percent) 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. Table IV lists the frequencies of these structural features as coded for the home page of each blog.

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 V 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, created by Pyra (which has since been purchased by the search engine company Google). Blogger's popularity is apparently 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. In as much as this is the

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

**Table IV.**
Structural features

software most users are choosing, these results reflect what a majority of blogs look like at the present time.

Next we consider patterns of linking from the home page of the blogs. We initially set out to analyze all the links and where they led, but soon found this to be too time-consuming. The sidebars of many blogs contain numerous links, each one of which would have had to be followed and manually coded. Instead, we coded for the presence of links (yes or no) of different types, classified in terms of their destination. These results are reported in Table VI.

Blogs are often defined in terms of linking to content elsewhere on the web. Most of the links we counted lead to (non-blog) web sites created by other than the blog author, although the number of blogs that do so (53.7 percent) is lower than one might expect. Only about half of the blogs (51.2 percent) link to other blogs, and fewer yet (36.1 percent) 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 e-mail contact links or badges. If we exclude badges from the count, the number of blog home pages with no links rises to 62 (30.5 percent, or nearly one-third) of the blogs in the corpus. In general, although some blogs contain many links, the extent to which the blogs in our sample link to other content is not as great as has been represented in previous characterizations of blogs.

*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 VII-IX under the categories "entry headers and footers," "entry body features" and "entry body text."

**Table V.**
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 |
| Total | 193 | 100 |

**Table VI.**
Links from home page

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

| 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 | |

**Notes:** Number of comments per entry: mean, 0.3; mode, 0; range, 0-6

**Table VII.**
Entry header and footer

| Feature | Frequency | Percentage[a] |
|---|---|---|
| Images | 18 | 9.2 |
| Links | | |
| to web sites 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 web sites created by or about self | 4 | 2.1 |

**Notes:** Number of links per entry: mean, 0.65; mode, 0; range, 0-11. [a] Of most recent entries coded ($n = 195$)

**Table VIII.**
Entry body features

| Measure | Total | Average | Range |
|---|---|---|---|
| Words | 42,930 | 210.4 | 1-1,262 |
| Sentences/fragments | 3,260 | 16.0 | 1-117 |
| Words per sentence | | 13.2 | |
| Paragraphs | 709 | 3.5 | 0-21 |
| Words in quotations | 3,681 | 18.0 | 0-430 |
| Sentences/fragments in quotations | 468 | 2.3 | 0-40 |
| Words per sentence in quotations | | 7.8 | |

**Table IX.**
Entry body text measures

Table VII 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 as well.

The last line of Table VII indicates that the average entry in our sample received 0.3 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 ($n = 17$)[13]. 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 = 0.3, mode = 0, range = 0-7 comments. It appears that entries do not continue to collect comments over time, but rather are only commented upon while they are new. Nor do the older blogs in our sample attract more comments than the newer blogs; no relationship between age of blog and number of comments received on the most recent entry was observed.

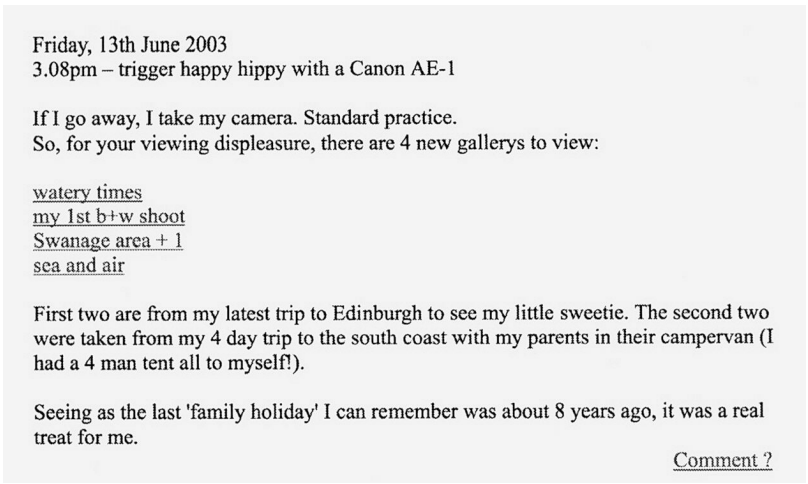
The above measures include blogs which do not allow readers to post comments. (Recall that this is the default setting for Blogger software.) If we exclude such blogs, the mean number of comments received per entry for blogs that allow comments is still less than 1 (0.9 for the most recent entry, and 0.8 for the oldest entry on the home page). Overall, therefore, we found fewer reader comments than previous claims about blog interactivity and community had led us to expect.

As important as comments may be in the popular perception of blogs, links within entries are even more important. Blood (2002a), 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 percent) contain any links at all, and that the central tendency is for an entry to have none (see Table VIII). The mean number of links per entry is 0.65, as compared with 1.89 as reported for Metafilter by Krishnamurthy (2002). When links are present, moreover, they rarely lead to news sites or other blogs as is widely claimed, although they do lead to other web sites. It is theoretically possible to include as many links as one wants in any blog entry; the choice is the author's[14], 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[15].

The blog entries analyzed contain few images (9.2 percent); however, this may be due in part to our 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 higher incidence of images.

The final set of measures involves structure at the level of the text of the blog entry; these are summarized in Table IX. At 210.4 words, the average blog entry is somewhat shorter than an e-mail posting to an academic discussion list (Herring, 1996). Its sentences, at 13.2 words, are three words shorter than those of private e-mail messages exchanged in a university setting (Cho, in press). 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.

The example of an entry from a blog in our corpus (Figure 5) is provided to illustrate the patterns summarized above. This message is shorter than average at 99 words and ten 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 still within the normal range of blog entries analyzed in the corpus. Its header



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 ?

**Figure 5.**
Example of an entry from a blog in our corpus

provides the date, time, and a title ("trigger happy hippy with a Canon AE-1"), and its 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. The content of the 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," the author provides his full name and indicates that he is 24 years old, works as "a systems admin for a linux/windows network," and resides in the UK. It is not unusual for a blog in our sample to provide this amount of personal information.

## Summary

This section has presented a quantitative characterization of the blog genre based on a corpus of 203 randomly selected English-language blogs. While our corpus does not include all available variants of blogs, we believe that it represents some currently prominent tendencies, notably the tendency for younger authors (especially students) to use standard blogging software to produce frequently updated personal content which contains relatively few links per entry and receives relatively few comments. Table X summarizes characteristics shared by a clear majority (more than 60 percent) of the blogs in our sample, excluding those determined by blogging software, and other general characteristics of the sample, including the absence of expected features.

Stated somewhat differently, the random corpus turned up few examples of the external-content-focused, densely interconnected journalistic or knowledge-sharing blogs that have been hailed as socially transformative (Cavanaugh, 2002; Festa, 2003; Glenn, 2003; Lasica, 2001), but many examples of blogs' self-expressive function (Blood, 2002a). This finding has implications for the antecedents and future trajectory of weblogs.

## Situating the weblog in an ecology of genres

### *Antecedents of weblogs*

Are weblogs an emergent, or a reproduced, genre? Blood (2002a) claims that the weblog is uniquely digital, "native to the web." Along with Dave Winer, who is sometimes

| Characteristic | Percent of blogs |
|---|---|
| Single-authored | 90.8 |
| Personal content | 70.8 + |
| Archives | 73.5 |
| Badges | 69.0 |
| Blogger's name on first page | 67.8 |
| Uses blogger software | 62.3 |
| *Other general characteristics:* | |
| Updated frequently; | |
| few advertisements; | |
| few links in entries; | |
| few comments on entries | |

**Table X.**
Common characteristics of blogs in sample

credited with creating the first weblog around 1996, Blood (2000) proposes 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. The first web site created by Tim Berners-Lee in 1991, a list of all web sites available at the time, was of this type (Winer, 1999). We find this claim to be problematic for several reasons.

First, the claim is premised on the assumption that blogs are link-centered filters of web content. However, the results of our analysis indicate that this is not the case for most blogs at the present time. Blood herself notes that journal-type blogs were more numerous than filters even at the time she wrote (Blood, 2002a). 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, they resemble the online journals found since the mid-1990s (Flynn, 2003), in that both focus on personal (often intimate) content, are updated daily or nearly so, have few links, and list entries in reverse chronological sequence. Figure 6 is a screen capture from an early online journals started in 1995.

In the entry shown in Figure 6, the author – an aspiring science fiction writer – describes her recent activities, which include giving her boss a hat for her baby for Christmas, selling a novella to a publisher, and visiting her parents, followed by a poem "to an old lover." The format and personal content of this journal resemble in many respects those of the personal journal weblog illustrated previously in Figure 3.

As the name of the journal in Figure 6 ("An ongoing, erratic diary") suggests, the online journal is itself reproduced from the centuries-old genre of handwritten diaries (McNeill, 2003). Diaries flourished in seventeenth-century England; a famous example is the diary kept by Samuel Pepys from 1660 to 1669 (Latham and Matthews, 1983). As in many weblogs, Pepys's diary contains a combination of intimate, banal content and comments on the political events of the day. While it is beyond the scope of this article to develop a full historical account, we find it revealing (and amusing) that Pepys's diary has been reformatted in recent years as a weblog, replete with entries re-ordered in reverse chronological sequence, links, and a discussion forum (www.pepysdiary.com). A reformatted entry from October 30, 1660 is shown in Figure 7.

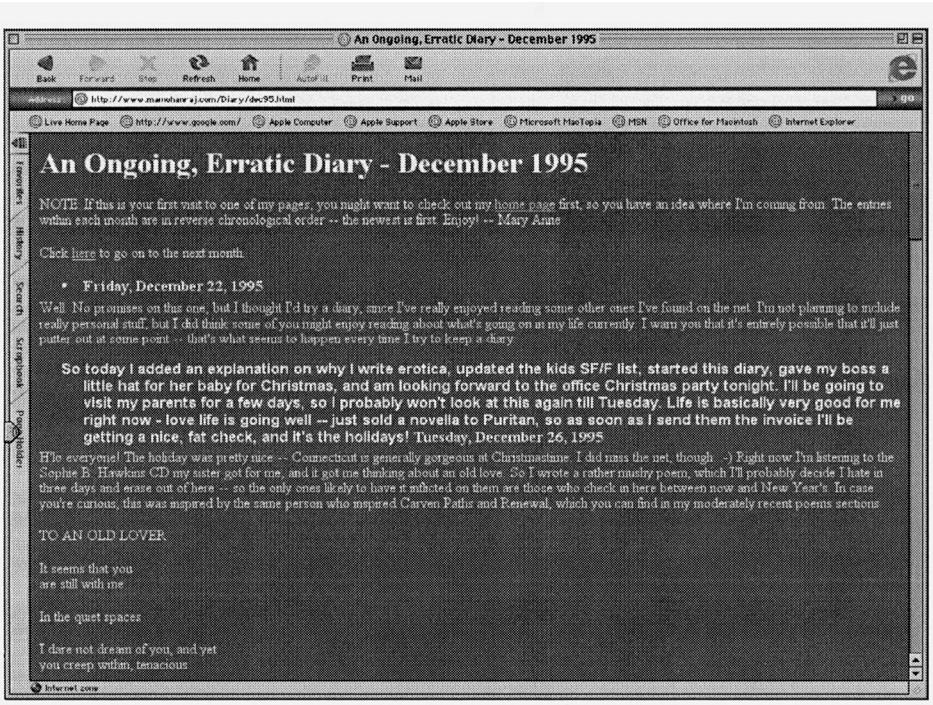



**Figure 6.**
An early online journal

If journal blogs are related to online journals, which in turn have off-line antecedents, the blog genre is not uniquely digital; in Crowston and Williams's (2000) terms, it is (at least partially) reproduced.

Second, the filter blog type, especially when used to self-publish commentary on current events, also has off-line antecedents, e.g. in editorials and letters to the editor in print newspapers, with which this type of blog is often compared in the popular press (Lasica, 2001). Similarly, 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 frequent uses of blogs, coded as "other" in our content analysis, 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 e-mail exchanges, which in turn have taken over the function of personal letters. Thus not only do blogs bear functional and structural resemblances to earlier off-line genres, they have multiple off-line antecedents.

Third, weblogs share characteristics with other digital genres. Most notably, these include the personal home page, which prior to the creation of blogs was the preferred means though which to present oneself and one's views on the web (Chandler, 1998). Although 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, their overall functionality is similar to that of personal home pages. Less obviously, perhaps, "community blogs" such as Metafilter and Slashdot, to which

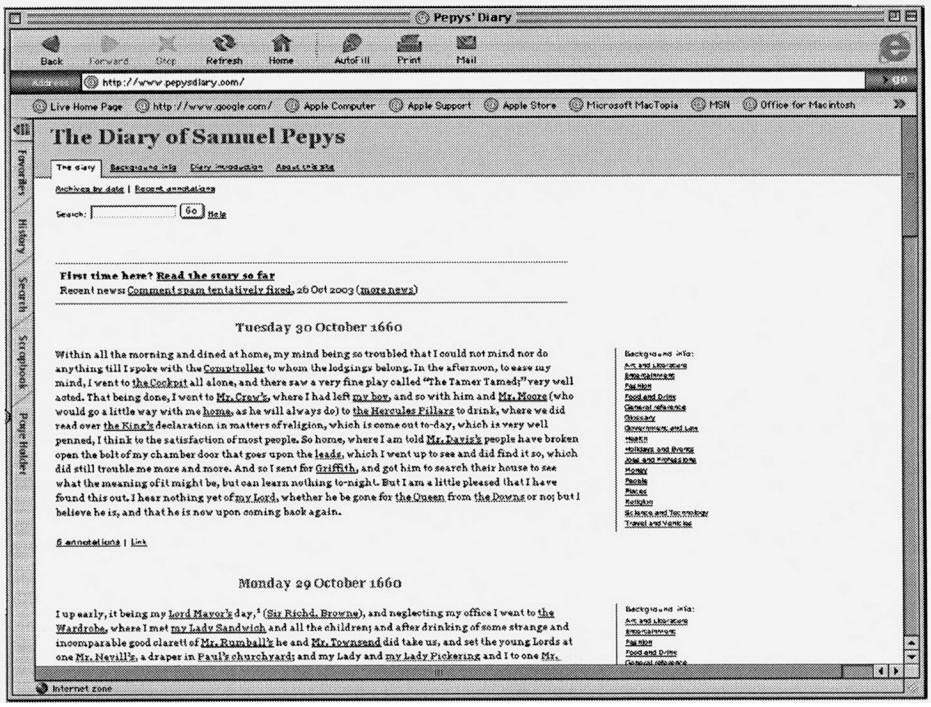



**Figure 7.**
An entry from Samuel Pepys's diary in weblog format

anyone can post entries, share features with asynchronous discussion forums, a text-based form of interactive computer-mediated communication, in that both are multi-participant, public, text-based, dynamically interactive online communication environments (cf. Smith's (1999) description of Usenet newsgroups). Not only community blogs have these features; individual blogs that allow comments also resemble threaded discussion forums in some respects. Taken together, these observations lead us to propose that blogs, rather than deriving from a single source, are in fact a hybrid of existing genres, rendered unique by the combination of features of the source genres they adapt, and by their distinctive technical affordances.

Weblogs exhibit hybrid properties at multiple levels. The genre as a whole includes identifiable sub-types which can be traced to multiple off-line antecedents. It is not sufficient, however, to posit that personal journal blogs derive from hand-written diaries, filters from newspaper editorials, k-logs from project journals, etc. – rather, individual blogs incorporate the functions of multiple genres. A number of blogs in our sample combine two or more purposes; we coded these as "mixed." Even blog types coded for primarily one purpose tend to incorporate elements of the other types; a pure filter or k-log is rare[16]. Many blogs are thus a hybrid of public and private, personal and professional. The hybrid nature of blogs comes about, we believe, because the technical ability afforded by blog software to modify web content quickly and easily makes it an attractive tool for a wide range of communicative uses, including some traditionally carried out in private. In blogging for purposes traditionally associated

with other genres, people inevitably carry over some of the practices and conventions of those genres into their blogging practices, resulting in a heterogeneous blend.

*Blogs in the genre ecology of the internet*

In the previous section, we compared weblogs with personal home pages, which are (typically) single-author HTML documents, and asynchronous newsgroups, a text-based form of interactive computer-mediated communication. These three genres comprise a genre repertoire (Orlikowski and Yates, 1994), in the sense that many internet users make use of all three; they can thus be considered interdependent components of a single ecology or system (Erickson, 2000). This observation leads us to a final argument for the hybrid nature of blogs.

Interactive text-based computer-mediated communication (CMC) is generally held to be a fundamentally different type of internet communication from the relatively 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: for example, HTML mail and encoding schemes have been developed to enable multimedia attachments from the CMC side, and links to chat and discussion forums have been incorporated into 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 8.

Figure 8 compares weblogs with standard HTML documents and asynchronous CMC along three dimensions: frequency of update, symmetry of communicative exchange, and multimodality. Weblogs are intermediate between standard web pages and asynchronous CMC with respect to each of these dimensions. Personal home pages may be updated only once every few months, while newsgroups are updated every time a conversational participant posts a message; weblogs are typically updated several times a week. Author and reader roles in standard HTML documents are highly asymmetrical, in contrast with the fully symmetrical give and take of unmoderated newsgroups; blogs allow limited exchanges (in the form of comments), while according blog author and readers asymmetrical communication rights – the author retains ownership of, and ultimate control over, the blog's content. Finally, blogs can incorporate multimedia elements as desired, like standard 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)

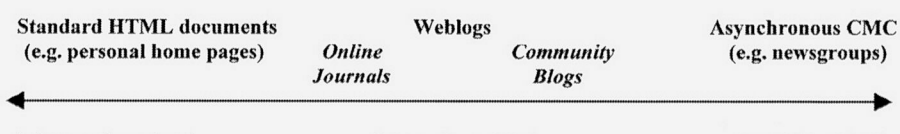



**Figure 8.**
Weblogs on a continuum between standard HTML documents and CMC

and "community blog" sites (such as Metafilter and Slashdot). Journal sites, with their more individual focus, are closer to standard web pages than are blogs. Community sites are closer to online discussion groups in their frequency of activity and exchange of messages among multiple participants than are individually maintained blogs. The two most frequent types of blogs defined by purpose in the present study could also be situated along the continuum, with journal-style blogs closer to online journals, and filter-style blogs closer to community blogs[17].

This analysis is intended to situate blogs in relationship to two popular modes of internet communication; it is not our intention to suggest that all forms of internet communication can be fit along a single continuum. Nor are blogs the only interactive HTML genre; social software systems such as Friendster and Orkut that allow users to post and manage content about themselves while interacting dynamically with other users also occupy a position intermediate between traditional web pages and CMC. We believe that the "bridging" properties of such systems make them attractive to users: they unite the best of two worlds. In particular, they allow authors to experience social interaction in ways that are otherwise difficult to achieve through web pages, while giving them ownership of, and control over, the communication space that is difficult to achieve in CMC. Combined with the unprecedented opportunity blogs provide for ordinary people to self-express publicly and at length, 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 the blog format has 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 created by almost anyone, and be responsive to people's daily communication needs. Second, by enabling readers to post comments directly to the blog, blog software 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 needs and interests of its users. While these features are not sufficient in and of themselves to bring about change in human behavior – technology changes nothing independent of its use by people – they are important triggers, as evidenced by the fact that the more cumbersome means for updating web content that have been previously available, and the practice of embedding CMC applications into web sites rather than making site content interactive, have not resulted in similar patterns of adoption and use. This analysis thus illustrates how technological changes, even incremental ones, can have widespread consequences through enabling new patterns of use. One of those consequences, in the case of weblogs, is the potential to reshape the genre ecology of the internet.

## Conclusions and future directions

In the limited descriptions previously available, weblogs have been characterized primarily as link-centered, highly interconnected, filters of web content. They have been hailed by some as a fundamentally new phenomenon with socially and cognitively transformative potential. In this study, we have sought to characterize the

weblog genre empirically, through content analysis of a corpus of randomly selected blogs, and theoretically, through consideration of its relationship to other genres – offline and online, past and present – with a particular focus on the functional domain occupied by blogs in relation to HTML documents and text-based CMC.

Our analyses revealed less evidence than expected of blogs as interlinked, interactive, and oriented towards external events; rather, most of the blogs in our corpus are individualistic, even intimate, forms of self-expression, and a surprising number of them contain few or no links. Based on the profile generated by the empirical analysis, we traced the historical antecedents of weblogs back to hand-written diaries. We also pointed out the hybrid nature of weblogs, suggesting that the technical affordances of the weblog format make it readily adaptable to multiple purposes of use. Finally, we suggested that these same affordances bridge, and ultimately blur the boundaries, between HTML documents and text-based CMC, as blogs and other interactive web-based communication systems replace some of the functions of traditional internet genres and give rise to new functions.

This analysis rests on three assumptions. First, it assumes that the sample of 203 blogs is representative of a majority of established, recently updated, text-based, English-language weblogs. While the sample is obviously too small to include all available blog variants, or to examine low-frequency variants (such as filters and k-logs) in depth, we nonetheless believe that it represents some dominant tendencies. Another sample of 155 blogs selected randomly from blo.gs in September 2003 and analyzed for 24 of the original 44 coding categories displayed similar results in every category (Herring *et al.*, 2004). Second, the analysis assumes that weblogs of this type are central to the weblog phenomenon as a whole. In this, we followed previous descriptions (e.g. Blood, 2002a, b; Rodzvilla, 2002), which focus on actively maintained, text-based, English-language blogs; these are also pulled up most often by the random blog selection feature of blo.gs. Third and last, we make the assumption that a genre is meaningfully defined by central tendencies discoverable through empirical analysis; that is, that a genre is a generalization about what a majority of its practitioners do. This assumption is consistent with previous empirically based genre research (e.g. Biber, 1993).

Support for the analysis comes from the ongoing evolution of weblogs. Since we began our data collection in the spring of 2003, the number of weblogs has more than quintupled, evidence that blogs continue to grow in popularity. The use of Blogger software has increased, supporting our assertion that ease of creation is a factor in blog growth. More people of all ages and both genders are creating personal journals in preference to other blog types (Herring *et al.*, 2004). Relatedly, online journal hosting sites such as LiveJournal, Diaryland and Xanga are now considered blogs (Henning, 2003), suggesting that the popular conception of the weblog has broadened to take account of the importance of online journals. At the same time, blog software is being used increasingly for non-blog, including commercial, purposes, as predicted by our analysis of the weblog as a socio-technical format open to multiple uses. We have also observed an increase in the number of advertisements on blog homepages and commercial spamming of blogs through comments. Commercialization could affect the degree of spontaneity and trust that goes into blog creation, potentially discouraging the posting of intimate content. As such, it is a trend to watch in the future.

This study is intended as an initial contribution to a systematic approach to weblog analysis and description. Given the importance of personal journal-type blogs, there is a need for empirical studies of online journal hosting sites such as LiveJournal, which are enormously popular and which appear to constitute semi-self-contained subcultures within the larger universe of blogs (Paolillo and Wright, 2004). The number of blogs in languages other than English is growing rapidly; these should be addressed in future studies as well. Photo blogs and audio blogs, which were excluded from the present corpus, also deserve to be characterized in their own terms. Finally, follow-up studies are needed to explore in depth some of the observations that emerge from the present study. For example, we suspect that weblogs facilitate social interconnection to a greater extent than is indicated by simple counts of links and comments; more focused methodologies could be employed to address this question[18].

We also believe that it is important to study the evolution of weblogs over time, to deepen our understanding of how technologically mediated genres emerge and develop. If our predictions are correct, the purposes to which weblog software will be put and the conventions that will arise around them will become so diverse in the future that it will no longer be meaningful to speak of weblogs as a single genre. Rather, as Yates and Orlikowski (1992) propose for e-mail, weblogs will become a "medium," or in our term, a socio-technical format, whose convenience and general utility support a variety of uses.

## Notes

1. A site created by Dave Winer as part of the "24 hours of democracy" project (Festa, 2003).
2. By Jorn Barger. The clipping "blog" came into use after Peter Merholz started pronouncing "weblog" as "wee-blog" in early 1999 (Blood, 2002a).
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 web site, you are not keeping a weblog" (Blood, 2002a, pp. 18-19).
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. At the time of our data collection in Spring 2003, English-language blogs constituted 65 percent of the blogs tracked by the NITLE Blog Census site (www.blogcensus.net/).
6. Spanish, German, French, Portuguese, Russian and Arabic are some of the other languages in which blogs were found.
7. We did not code for the "notebook" – long, focused essay – type of blog described by Blood (2002a), in that her criterion for defining it in terms of entry length seemed problematic: we observed that entries can vary significantly in length within a single blog. Nor did such a type emerge naturally from our data.
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 could be coded, excluding "unknowns" and other problematic instances.
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; blogs that document travel; and blogs that serve as conversation boards for two or more authors.
11. For further discussion of variation in blog type according to gender and age of blog author, see Herring *et al.* (2004).
12. For example, it does not require users to host blogs on their own servers, but rather provides free space on a public server.
13. This suggests either that community blogs (as compared to single-author blogs) tend to receive more comments, or that Metafilter is atypical in this regard.
14. With most blog software, however, the author must know how to create a link using HTML.
15. Personal journal type blogs are least likely to allow comments, and not just because 68 percent use Blogger software, which has "no comments" as a default. Of the 95 Blogger personal journals found, only 24 percent allow comments, as compared with 69.2 percent of the filter blogs created with Blogger ($n = 13$). All three k-logs created with Blogger allow comments. This finding is inconsistent with Blood's (2002a) assertion that personal journals involve the most message exchanges.
16. Both usually contain some personal journalizing.
17. 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.
18. For an example of one such approach employing social network analysis, see Herring *et al.* (2005).

## References

Arnold, J. and Miller, H. (1999), "Gender and web home pages", available at: http://ess.ntu.ac.uk/miller/cyberpsych/cal99.htm

Bates, M.J. and Lu, S. (1997), "An exploratory profile of personal home pages: content, design, metaphors", *Online and CD-ROM Review*, Vol. 21 No. 6, pp. 331-40.

Bauer, M. (2000), "Classical content analysis: a review", in Bauer, M. and Gaskell, G. (Eds), *Qualitative Researching with Text, Image and Sound*, Sage, Thousand Oaks, CA, pp. 131-51.

Biber, D. (1993), "The multi-dimensional approach to linguistic analyses of genre variation: an overview of methodology and findings", *Computers and the Humanities*, Vol. 26, pp. 331-45.

Blood, R. (2000), "Weblogs: a history and perspective", September 7, available at: www.rebeccablood.net/essays/weblog_history.html

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

Blood, R. (2002b), "Introduction", in Rodzvilla, J. (Ed.), *We've Got Blog: How Weblogs Are Changing Our Culture*, Perseus Publishing, Cambridge, MA, pp. ix-xii.

Cavanaugh, T. (2002), "Let slip the blogs of war", in Rodzvilla, J. (Ed.), *We've Got Blog: How Weblogs Are Changing Our Culture*, Perseus Publishing, Cambridge, MA, pp. 188-97.

Chandler, D. (1998), "Personal homepages and the construction of identities on the web", available at: www.aber.ac.uk/media/Documents/short/webident.html

Cho, N. (in press), "Linguistic features of electronic mail in the workplace: a comparison with memoranda", in Herring, S.C. (Ed.), *Computer-Mediated Conversation*, Hampton Press, Cresskill, NJ.

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. 201-16.

Dillon, A. and 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*, Vol. 51 No. 2, pp. 202-5.

Döring, N. (2002), "Personal home pages on the web: a review of research", *Journal of Computer-Mediated Communication*, Vol. 7 No. 3, available at: www.ascusc.org/jcmc/vol7/issue3/doering.html

Erickson, T. (2000), "Making sense of computer-mediated communication (CMC): conversations as genres, CMC systems as genre ecologies", *Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS-33), Maui, HI*, IEEE Press, Los Alamitos, CA, available at: www.pliant.org/personal/Tom_Erickson/genreEcologies.html

Festa, P. (2003), "Blogging comes to Harvard", *CNET News.com*, February 25, available at: http://news.com.com/2008-1082-985714.html?tag = fd_nc_1

Flynn, S.I. (2003), "Scribe tribes and shape shifters: an ethnographic study of online journal communities", unpublished doctoral dissertation, Department of Anthropology, Yale University, New Haven, CT.

Foot, K.A. and Schneider, S.M. (2002), "Online structure for political action: exploring presidential campaign web sites from the 2000 American election", *Javnost – The Public*, Vol. 9 No. 2, pp. 43-60.

Glaser, B. and Strauss, A.L. (1967), *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Aldine Publishing Co., Chicago, IL.

Glenn, D. (2003), "Scholars who blog", *The Chronicle of Higher Education*, June 6, pp. A14-A16.

Ha, L. and James, E.L. (1998), "Interactivity reexamined: a baseline analysis of early business web sites", *Journal of Broadcasting and Electronic Media*, Vol. 42 No. 4, pp. 457-74.

Halavais, A. (2002), "Blogs and the 'social weather'", paper presented at Internet Research 3.0, Maastricht, October.

Henning, J. (2003), "The blogging iceberg – of 4.12 million hosted weblogs, most little seen, quickly abandoned", Perseus Development Corp. White Papers, available at: www.perseus.com/blogsurvey/thebloggingiceberg.html

Herring, S.C. (1996), "Two variants of an electronic message schema", in Herring, S.C. (Ed.), *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, John Benjamins, Amsterdam, pp. 81-108.

Herring, S.C. (2003), "Gender and power in online communication", in Holmes, J. and Meyerhoff, M. (Eds), *The Handbook of Language and Gender*, Blackwell Publishers, Oxford.

Herring, S.C., Kouper, I., Scheidt, L.A. and Wright, E. (2004), "Women and children last: the discursive construction of weblogs", in Gurak, L., Antonijevic, S., Johnson, L., Ratliff, C. and Reyman, J. (Eds), *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*, available at: http://blog.lib.umn.edu/blogosphere/

Herring, S.C., Kouper, I., Paolillo, J.C., Scheidt, L.A., Tyworth, M., Welsch, P., Wright, E. and Yu, N. (2005), "Conversations in the blogosphere: an analysis 'from the bottom up'", *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS-38),*

*Big Island, HI*, IEEE Press, Los Alamitos, CA, available at: http://ella.slis.indiana.edu/~herring/blogconv.pdf

Hourihan, M. (2002), "What we're doing when we blog", June 13, available at: www.oreillynet.com/pub/a/javascript/2002/06/13/megnut.html

Krishnamurthy, S. (2002), "The multidimensionality of blog conversations: the virtual enactment of September 11", paper presented at Internet Research 3.0, Maastricht, October.

Lasica, J.D. (2001), "Blogging as a form of journalism", *USC Annenberg Online Journalism Review*, May 24, available at: www.ojr.org/ojr/workplace/1017958873.php.

Latham, R. and Matthews, W. (Eds) (1983), *The Diary of Samuel Pepys*, Vols 10-11, Bell & Hyman, London and University of California Press, Berkeley, CA and Los Angeles, , CA (Vols 1-9 published by G. Bell and Sons, London and University of California Press, London, Berkely, CA and Los Angeles, CA, 1970-1976).

Lave, J. and Wenger, E. (1991), *Situated Learning: Legitimate Peripheral Participation*, Cambridge University Press, Cambridge.

McMillan, S.J. (1999), "Health communication and the internet: relationships between interactive characteristics of the medium and site creators, content, and purpose", *Health Communication*, Vol. 11 No. 4, pp. 375-90.

McNeill, L. (2003), "Teaching an old genre new tricks: the diary on the internet", *Biography: An Interdisciplinary Quarterly*, Vol. 26, pp. 24-48.

Miller, C.R. (1984), "Genre as social action", *Quarterly Journal of Speech*, Vol. 70, pp. 151-67.

NITLE Blog Census (2004), available at: www.blogcensus.net/?page = Home

Orlikowski, W.J. and Yates, J. (1994), "Genre repertoire: norms and forms for work and interaction", Working Paper, No. 3671-94, MIT Sloan School of Management, Cambridge, MA, available at: http://ccs.mit.edu/papers/CCSWP166.html-genre5

Paolillo, J.C. and Wright, E. (2004), "The challenges of FOAF characterization", paper presented at the 1st Workshop on Friend of a Friend, Social Networking and the (Semantic) Web, Galway, September 2, available at: www.w3.org/2001/sw/Europe/events/foaf-galway/papers/fp/challenges_of_foaf_characterization/

Polger, M.A. (2003), "Re-writing the holocaust online: a discourse analysis of holocaust denial web sites", paper presented at the Association of Jewish Libraries Annual Conference, Toronto, June, available at: http://mark.degrassi.ca/ajl.ppt

Ray, T. (2003), "Why blogs haven't stormed the business world", *E-Commerce Times*, April 29, available at: www.ecommercetimes.com/perl/story/21389.html

Rodzvilla, J. (2002), *We've Got Blog: How Weblogs Are Changing Our Culture*, Perseus Publishing, Cambridge, MA.

Shepherd, M. and Watters, C.R. (1998), "The evolution of cybergenres", *Proceedings of the 31st Annual Hawaii International Conference on System Sciences (HICSS '98), Kohala Coast, HI*, Vol. II, IEEE Press, Los Alamitos, CA, pp. 97-109.

Smith, M.A. (1999), "Invisible crowds in cyberspace: mapping the social structure of the Usenet", in Smith, M.A. and Kollock, P. (Eds), *Communities in Cyberspace*, Routledge, London, pp. 195-219.

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

Warnick, B. (1998), "Appearance or reality? Political parody on the web in campaign '96", *Critical Studies on Mass Communication*, Vol. 15 No. 3, pp. 306-24.

Winer, D. (1999), "The history of weblogs", available at: http://newhome.weblogs.com/historyOfWeblogs

Winer, D. (2001), "What are weblogs?", November 16, available at: http://newhome.weblogs.com/personalWebPublishingCommunities

Yates, J. and Orlikowski, W.J. (1992), "Genres of organizational communication: a structurational approach to studying communication and media", *Academy of Management Review*, Vol. 17 No. 2, pp. 299-326.

## Appendix. Coding categories

*Overall identification*

A Blog #: assigned sequentially to random blogs as they are retrieved from blo.gs.

B Acquisition date: date the blog was retrieved for coding.

C Acquisition time: time the blog was retrieved for coding.

D URL of blog homepage: (give URL).

E Title in header (include all plausibly-related text, even if in blog software logo): none (0); give title and any associated description.

F Title in title tag: none (0); give title and any associated description.

G Title in URL (excluding personal names and domain names of other web sites): none (0); give title and any associated description.

H Title elsewhere on first page: none (0); give title and any associated description.

*Blog author(s)*

I Blogger's name (from first page only): none (0); pseudonym (1); first name (or transparently derived nickname) (2); full name (3); other (4); first name + initial (5); initial + last name (6).

J Blogger's name – location (code all that apply): no name visible from first page (0); in header (1); in title tag (2); in URL (3); in sidebar (4); in entry header or footer (5); in body of entry (6); other (7).

K Number of blog authors: (give number).

*Code L-Q separately for each blog author*

L Gender (from any available source): unknown (0); male (1); female (2).

M Age (from any available source): unknown (0); adult (1); teen (aged 13-19) (2); child (3).

N Occupation (from any available source): unknown (0); other (describe).

O Geographical location (from any available source): unknown (0); other (give country).

P Personal information about blogger (including explicit indication of gender, age, occupation, location, etc.): none (0); on first page (1); one click away from first page (2); elsewhere (3).

Q Graphic representation of blogger: none (0); on first page (1); one click away from first page (2); elsewhere (3).

*History and activity level of blog*

R Date of current entry (at time of sampling).

S Time of current entry.

T Date of next-most-current entry (at time of sampling).

U Time of next-most-current entry.

V Date of oldest entry in blog.

W Time of oldest entry in blog.

*Technical features*

X Blog software: unknown (0); other (give software name).

Y Comments on entries: not allowed (0); allowed (1).

Z Search (on first page): no (0); yes (1).

AA Calendar (on 1st page): no (0); yes (1).

AB Archives (text and/or date links outside calendar; on 1st page): no (0); yes (1).

AC Links to CMC (on 1st page; code all that apply): none (0); e-mail address (1); guest book (2); mailing list (3); other (4).

*Overall content*

AD Blog type (based on predominant content from entries on first page): unknown (0); personal journal (1); filter (2); k-log (3); mixed (4); other (5).

AE Genuineness of presentation (is the site what it appears to be?): unknown (0); genuine, including humor sites presented straightforwardly as such (1); non genuine, including satire, irony, or deceptively-presented sites (2).

*For AF-AI, code all that apply*

AF Ads: none on first page (0); in header (1); in sidebar (2); in footer (3); in entries (4); other.

AG Images (non-background): none on first page (0); in header (1); in sidebar (2); in footer (3); in entries (4); other (5).

AH Badges: none on first page (0); in header (1); in sidebar (2); in footer (3); in entries (4); other (5).

AI Links to webrings (i.e. to homepage of webrings or to a list of thematically-related links, excluding generic themes such as "friends"): none on first page (0); in header (1); in sidebar (2); in footer (3); in entries (4); other (5).

*For AJ-AM, code only for direct links, chosen by the blogger*

AJ Links to other blogs: none on first page (0); in header (1); in sidebar (2); in footer (3); in entries (4); other (5).

AK Links to non-blog content: news sources: none on first page (0); in header (1); in sidebar (2); in footer (3); in entries (4); other (5).

AL Links to non-blog content: other web sites created by the blogger: none on first page (0); in header (1); in sidebar (2); in footer (3); in entries (4); other (5).

AM Links to non-blog content: other web sites created by others: none on first page (0); in header (1); in sidebar (2); in footer (3); in entries (4); other (5).

*First entry*

AN Entry header (code all that apply): none (0); title of entry (1); blogger name (2); date (3); time (4); internal links (other than comments, but including permalinks) (5); external links (6); comments (7); other (8).

AO Entry footer (code all that apply): none (0); title of entry (1); blogger name (2); date (3); time (4); internal links (other than comments, but including permalinks) (5); external links (6); comments (7); other (8).

AP Number of words in entry body (including quotes, etc.): (give number).

AQ Number of sentences or sentence fragments in entry body (including quotes, etc.): (give number).

AR Number of paragraphs in entry body (including quotes, etc.): (give number).

AS Number of words in quotations (count all and only true quotations, regardless of whether in quotation marks): (give number).

AT Number of sentences or sentence fragments in quotations: (give number).

AU Images in entry body: (give number).

AV Links in entry body (put in by blogger): (give number).

AW Links – type (code all that apply): none (0); to other blogs (1); to news sources (2); to other web sites created by blogger (3); to other web sites created by others (4); internal to blog (5).

*Comments*

AX Most recent entry: number of comments: none (0); one or more (give number).

AY Most recent entry: number of unique commenters: none (0); one or more (give number).

AZ Most recent entry: number of comments by blogger: none (0); one or more (give number).

BA Oldest entry on page: number of comments: none (0); one or more (give number).

BB Oldest entry on page: number of unique commenters: none (0); one or more (give number).

BC Oldest entry on page: number of comments by blogger: none (0); one or more (give number).

BD Other remarks.