Marianne K. Gouge. Blogs as a Means of Preservation Selection for the World Wide Web. A Master's Paper for the M.S. in I.S degree. July, 2004. 42 pages. Advisor: Jeffrey Pomerantz

Currently, there is not a very strong system of selection in place when looking at the Web as a whole. This study is an examination of the blogging community for the possibility of utilizing the decentralized and distributed nature of link selection that takes place within the community as a means of preservation selection. The purpose of this study is to compare the blog aggregators, Daypop, Blogdex, and BlogPulse, for their ability to collect content which is of archival quality. This study analyzes the content selected by the aggregators to determine if any content which is linked most frequently for a given day is of archival quality. Archival quality is determined by comparing the content from the aggregator lists to criteria assembled for the study from a variety of archival policies and principles.

Headings:

    World Wide Web

    Web site--Weblog

    Preservation of Library Materials--Electronic Data Archives/Conservation and Restoration

    Archival Material--Computer Files

    Preservation--Computer Files

BLOGS AS A MEANS OF PRESERVATION SELECTION FOR THE WORLD WIDE
WEB

by
Marianne K. Gouge

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

July 2004

Approved by

_____

Jeffery Pomerantz

Table of Contents

Introduction

For the past few years archivists and preservationists have been struggling with

the preservation of digital records. Electronic media has become a part of almost all

aspects of business and culture as more and more information is being produced digitally.

The digital environment has a unique set of problems pertaining to preservation. The

principles that archivists have been developing for digital preservation come from the

traditional world of physical archives. Paul Conway "suggests a framework for

understanding preservation in the digital context by creating a bridge from the five core

principles of traditional preservation practice: longevity, choice, quality, integrity, and

access" (1999). These principles are achievable when applied to discrete digital objects or

to a collection of related digital objects, but it becomes increasingly more difficult to

apply these principles when looking at the Web as a whole. Tim Berners-Lee envisioned

the web as a community of shared information. "Its universality is essential: the fact that

a hypertext link can point to anything, be it personal, local or global, be it draft or highly

polished" is the essence of the Web (Berners-Lee, 1998). In the environment of the Web

documents will then potentially have placement in a variety of collections through a

variety of relationships.

How then do you capture this collection of information? One method is to try and

capture it all and save a replica of the Internet for future generations. Another method is

through selection, which can be time consuming and requires a great deal of personal

judgment. Traditionally those judgments are made by professionals in the library and

archival community. It would take an enormous staff of these professionals to review the documents created on the Web to determine their preservation value. Though without going through the process of selection, value is not added, and "choice involves defining value, recognizing it in something, and then deciding to address preservation needs in the way most appropriate to that value" (Conway, 1996).

However, there is a community which exists on the Web which contains professionals from a variety of disciplines already looking though the information provided on the Web. This is the community of weblogs. This community could provide assistance in the task of selection.   The community of weblogs has grown in tandem with the growth of the Web reflecting the communities that have become a part of what was once a place limited to those who were only comfortable with technology. Dave Winer, author of Scripting News, even refers to the first website, http://info.cern.ch/ as the first weblog (Winer, 2002). A weblog is "literally, a "log" of the web - a diary-style site, in which the author (a "blogger") links to other web pages he or she finds interesting using entries posted in reverse chronological order" (Perrone, 2004). Blogs offer the service of choice or selection because of their hypertext nature and because of the networked community they form. The information from these communities is being harvested by a few sites that aggregate the most linked to content into list. Could blog aggregators be used to predict content which is of archival quality?

Purpose of Study

The purpose of this study is to compare these aggregators for their ability to collect content which is of archival quality. Three aggregators were selected for this study to offer enough content for comparison and because each aggregator offers a list

indicating the most linked to content for the day from their slice of the blogosphere. Also, the method and source used to determine this content varies from aggregators to aggregators as well as length of time the aggregators has been in operation. The three search aggregators selected were [Daypop](#), [Blogdex](#), and [BlogPulse](#).

Daypop and Blogdex are two of the most established of sites that offer this feature of producing a list of most linked to information on the Web. The other established blog sites such as Technorni, do not offer the same feature of providing a list of the most linked to content on the Web. For example, Technorni divides the content by news or current event instead of by a straight system of ranking. BlogPulse is a newer less established site, but has a body of research to support the efforts behind the work the site is doing. From this information the choice was made to use these three sites for comparison rather than other sites on the Web.

This study will analyze the content selected by the aggregators to determine what content is of archival quality. Archival quality will be determined by comparing the content from the lists to criteria assembled for the study from a variety of archival policies and principles. The aggregators will also be analyzed to see what content is selected by all the aggregators, what content is different, and to see if any of the aggregators demonstrate a particular strength or weakness in their ability to select content of a particular archival criterion.

Content is defined as any digital object that a blogger has provided a link to in his/her blog as part of his/her post. These links may point to articles that are part of a news service, pictures, or even short films that are part the information being passed around the blogosphere. The more bloggers point to the same content, the more likely

that content will end up in the lists which are compiled from the aggregators' sample of the blogosphere.

Blogosphere is a term coined by William Quick intended to define the entirety of the space on the Web occupied by bloggers (2001). In this space the writer and the content – the blog and the blogger are difficult to separate and often interchangeable. The content of the blog becoming a projection of the author's identity as it is seen on the Web.

Internet Preservation

There are two approaches currently being used when preserving information created on the World Wide Web. One approach is the "save everything" approach while the other is to make selective decisions about what to save. The Internet Archive (IA) is an example of the save everything approach. In 1996 Brewster Kahle conceived of a method of capturing everything available on the World Wide Web for the sake of preservation. According to Kahle, "the Internet is a medium for artifacts that are considered to be ephemeral, both bibliographic, and nonbibliographic" (Edwards, 2004). The Internet Archive contains over 300 terabytes of data and is currently growing at a rate of 12 terabytes per month (Edwards, 2004). IA contains "a general collection of Web sites, two collections of Web sites from the 1996 and 2000 presidential elections, Web pages containing content related to the terrorist attacks on the World Trade Center and the Pentagon on September 11, 2001(news, reactions, commentary), and a Web Pioneers collection highlighting early commercial and organizational sites on the Web" (Edwards, 2004). The general collection of Web sites can only be accessed though entering a specific URL, so that accessing information from the archive can be clunking

and incomplete despite the efforts to capture everything. The reason for this is that "the

Web crawlers are limited by how many times they may visit a domain to collect new or

modified pages; while the intervals of [collecting] new archival material for a site have

been drastically reduced, there are still sizable gaps in content that is lost between each

archived page" (Edwards, 2004). Even if it becomes technically and economically

feasible, should we take the attitude of "saving everything"? In the end it may be

impossible to capture everything because "the Web is both an information archive and a

social network" (Burbules, 2001). Collecting the documents that represent the Web may

not fully represent the community and usage of the information because of the social

environment of the Web.

The other model of Internet preservation currently in use is when we selectively

choose what should be preserved from the Web. Most of these efforts are "based on a

primary strategy to preserve national material on official government Web sites as well as

Web pages that are registered under the nation's domain" (Edwards, 2004). PANDORA

is one example of these efforts which is being undertaken by the National Library of

Australia and its partners. PANDORA is focusing on publications produced by the

Australian government, Australian Institutions, and individual Australians through items

such as journals, newspapers, conference proceedings, or other substantial literary works.

The library acknowledges the disadvantages of the selective method but hopes that

through selection, "each item in the archive is quality assessed and functional to the

fullest extent permitted by current technical and resource capabilities; each item in the

archive can be fully catalogued and there fore become a part of the national bibliography,

and each item in the archive can be made accessible"(Phillips, 2003). Sweden, Canada, and Tasmania are undertaking similar projects.

Neither one of these methods are ideal. The Internet Archive may or may not be sustainable or scalable. Selection is only feasible when applied to a specific set of documents found within the corpus on the Web, and even then can be overwhelming and difficult to properly staff. This paper explores a third method which uses a combination of the two. Bloggers manually cull the Internet each day linking to objects of interest. This behavior of the community of Bloggers is basically performing the task of selection on the mass of information produced on the Web. Once a useful and interesting link is made within the blogging community, it will be linked to by other blogs and spread throughout the community. This phenomenon of linking will be discussed further in the section below about applying credibility to the blogging community. With the use of aggregators this spread of links could be harvested to create a manageable body of work representative of the social culture on the World Wide Web. This collection would capture the daily interests and conversations spreading across the blogosphere and would also capture the essence of the Web and its shared information that would then be preserved.

Defining Blogs

"11.30.03
Adventures in Vermont living, chapter thirty-four:
CHIMNEY FIRE!
It is *bitter* out this morning, a raw-wind day, little pellets of wet ice spitting from the grey sky. I came home from Meeting laden with groceries (I stopped for groceries after Meeting; the Quakers do not distribute groceries to Meeting-goers or anything) and bundled them all in the house and set them on the floor of the front room and thought, boy howdy, a fire would be just the ticket to-day" (Baker, 2003).

This is what most people think of when discussing a weblog. Yet blogs, as they are commonly called, have matured into more than diaries of personal musings. Since the advent of free blogging applications the ease of publication has transformed the world of blogs from a group of HTML savvy techies to a reflection of social culture as a whole (Blood, 2002). Members of a variety of communities have adopted blogging as a method of disseminating and filtering information. "These blogs . . .filter out the noise of the web, drawing back readers with new, interesting, and useful content everyday" (Carver, 2003). Blogs provide like-minded people with timely news, Web sites, and information, which helps build a small community around the blog. "Blogs succeed largely because they are extremely native to the Web as Tim Berners-Lee conceived it in the first place" (Searls & Sifry, 2003). Blog content is most often built around an out link to content on the Web which is then linked to from other blogs creating a place of shared information across the Web. The blogosphere has exponentially grown in the past few years making it more and more difficult to ignore. . "Recent Pew Internet and American Life research suggests that 11 percent of American Net users have read blogs and 2-7 percent have created them. This translates to between 2.4 and 8.4 million bloggers" (Gill, 2004).

Rebecca Blood, author of *We've Got Blog,* traces the root of the word weblog to Jorn Barger who first used the name in December, 1997. "The original weblogs were link-driven sites. Each was a mixture in unique proportions of links, commentary, and personal thoughts and essays. . . These weblogs provide a valuable filtering function for their readers. The Web has been, in effect, pre-surfed for them" (Blood, 2002). A list of blogs was compiled by Jesse James Garrett, editor of Infosift, and posted on Cameron Barrett's site CamWorld. "CamWorld began on June 11, 1997 - a few days before [he]

was to begin teaching a college class on HTML and new media design. [he] <u>initially used the site</u> to post links to web design articles [he] wanted [his] students to read for class" (Barrett, 2003). The site grew and was frequently referenced by other members of the community. Originally the community's members were individuals who had been using the web and comfortable authoring with HTML.

When blogging applications started being released in spring of 1999, the community grew from technically oriented individuals to anyone with the desire to create their own content. <u>Blogger</u>, is freely available and "gives you a way to automate and accelerate the publishing process without writing any code or worrying about installing any sort of server software or scripts (B. Stone, 2004). From the creation of this easy method of publication, "blogs [began to] vary greatly: personal (diaries, photos, poetry, post news about a hobby, keeping in touch with family members, gossip, celebrity fan mail), technical (project updates, develop ideas collaboratively, platforms for uncensored ideas) or news (breaking news stories, rumors)" (Young, 2003).

Blogs should be thought of as micro-content management systems instead of online diaries. Along with Blogger, the other applications have "come out of the traditional Web site content management space" (Searls & Sifry, 2003). "The software automatically formats and posts the entry. It also automatically archives older ones on separate pages. If categories are used in the creation of entries, the software can also create subject-specific archives based on keywords used" (Notess, 2002). The blog's design can also be customized to the author's needs. These features place blogging applications in the ranks with many content management systems. Boiko envisions a CMS that allows for a process of collecting, managing and publishing content in any

format.   Through various layers of metadata, content components are managed by

tracking authorship, version control, and the appearance of the output (Boiko, 2002).

Blogging applications may not be sophisticated to the extent that Bob Boiko describes in

his book *Content Management Bible* (2002), but increasingly, there's only a thin layer of

functionality separating blogware from low-end Content Management solutions such as

basic workflow, permissions, and update histories. Yet blogs do have the ability to

operate on multiple databases with a variety of users who can manage posts through draft

status and future postings (Hiler, 2002). No longer can blogs be viewed as only simple

pages containing links or ramblings from personal diaries. Blogging platforms now

appeal to a wide audience and add a multilayer of functionality to the content being

provided.

Blogging Communities

One application of the blog as a content management system is within the

corporate environment.   "Since the launch of Movable Type in October 2001, [they have

had] probably more business users than any other Weblog tool, and [they have] done

things like About.com [replacing] their entire proprietary publishing system with a

customized version of Movable Type" (Jones, 2003). "Macromedia is the largest tech

company to have official blogs on their corporate Web site, and Fox had a production

assistant's Weblog on the promotional Web site for their series Firefly" (Tepper, 2003).

Corporations are also using blogs to share news throughout the business and to

track business process on projects. "Many companies are creating team and project blogs

for internal use. These blogs serve as centralized locations for knowledge management

and project coordination" (Tepper, 2003). For example, Community Connect, a company

in New York that operates a variety of online communities, while interviewing

candidates for a position posted comments about each candidate on a password protected

weblog to help track their status (O'Shea, 2003). "Verizon Communications uses a

weblog to collect news and intelligence about the industry and competitors" (O'Shea,

2003). They were spending lots of time and server space by emailing articles and

information about companies to interested individuals without a system to track who

received what. Now that information is store in a series of topic-specific blogs. As these

examples show, corporations have adopted blogs for a variety of usages from the

complex to the simple.

Information specialists are another community that has adopted blogging for

communicating information on the Web. "Librarians are great filters of information"

making the platform of a blog a remarkable method for relaying that information

(Schwartz, 2003). One of the eminent librarian blogs is Gary Price's site Resource Shelf.

Price is seen as the "indefatigable finder of reference sources on the invisible web"

(Block, 2001). By posting the information he finds on a weblog his readers have instant

access without suffering a clogged email box. Jenny Levine was one of the first to

provide access to valuable resources and to give librarians a reason to go on the Web

back in 1995 (Block, 2001). Her weblog is The Shifted Librarian. Other library weblogs

of note are Library Stuff and Librarian.net.

Blake Carver has taken his blog to the next level of community building by

creating a collaborative blog called LISNews.com. By signing in with a user name and

password a variety of people can post information to a single site. LISNews is divided in

a variety of sections making it easy for readers to access the type of information they

seek. Carver has over twenty librarians contributing to his site from all over the United

States. LISNews "focuses more on traditional library news content" (Turner, 2003).

Blogs are not just representative of individual collaborations; non-corporate

institutions have begun to adopt the media as well. Several public and academic libraries

use blogs to broadcast recent information. "The UK's Gateshead Library has a blog with

topics ranging from the top albums of the 1980's, to Spider Man and the value of

XreferPlus" and "the Waterboro Public Library, ME, has run a prolific blog with content

added daily" (Carver, 2003). Other examples of blogs found in the academic world

include one maintained by Rowland Institute at Harvard and another by R. B. House

Undergraduate Library at University of North Carolina – Chapel Hill.

In academia blogs have not just been picked up by the library community, but

also by scholars. Ray Schroeder uses a blog to continue communication with his

students. He began by emailing students in the early 1990s and quickly converted to

producing listservs. Soon he realized with the listservs that students were not removing

their names after graduation but still received the updates and links he was providing. He

also began to have inquires from other individuals on the campus who were not in his

classes to be added to the listservs. "The listservs, at their high point, directly reached a

few hundred students, former students, and colleagues. The Online Learning Update

blog, on the other hand, collects thousands of visits each month" (Schroeder, 2003). By

changing the platform Schroeder presented the information his readership increased ten-

fold. Schroeder's experience is telling of how blogs can change and influence the flow of

information.

Professors are not only using blogs in the classroom they are also using them as a way to publish ideas and research. "In their skeptical moments, academic bloggers worry that the medium smells faddish ephemeral. But they also make a strong case for blogging's virtues" (Glenn, 2003). With the freedom and informality of the web scholars are allowed to explore topics and expound to any desired length and not having to limit the content to sound bites (Glenn, 2003). Henry Farrell, an assistant professor of political science at the University of Toronto at Scarborough, maintains a list of over ninety-three scholarly blogs most of which are less than a year old (Glenn, 2003). Disciplines listed on Farrell's blog include political science, economics, sociology, law, and history just to name a few.

Jason Griffey, a recent graduate of the School of Information and Library Science at the University of North Carolina at Chapel Hill, posted his Master's paper on his blog and submitted a link to Boing Boing: A Directory of Wonderful Things. Boing Boing is a collaborative site that provides links to a variety of interesting information on the Web. In the week following the posting of his paper, Griffey had received over 1500 unique hits to his blog and had received thirty-two comments on his original post made on April 7, 2004 (2004). Many of the comments were coming from scholars in the field. He received 720 hits alone on the day that the link appeared in Boing Boing (Griffey, 2004). This is a vivid illustration of how the blogging community can be used to disseminate information much more quickly than the typical channels a scholar would have to use to have his work published. Though questions may arise about the credibility of information that has not gone through the traditional channels of peer review or an

editorial board, a formal process is not necessary to ascertain credible information. Credibility can also be established through informal channels.

Defining Credibility on the Web

Nicholas Burbules defines credibility though a variety of methods. Judgments of credibility are assessed by what is useful, relevant, or interesting. Further assessment occurs when looking at the timeliness and comprehensiveness of information. "The standard criteria for judging credibility online are frustrated by the characteristic conditions of the World Wide Web and of the larger Internet" (Burbules, 2001). Assessing credibility on the Web is a difficult task. "The markers of institutional credibility and authority, the lines of tradition that allow viewers to judge media sources or publishers, for example, have not been settled yet" (Burbules, 2001). There are other perfunctory methods used to determine the credibility of a Web site such as visual quality and design, URL domain name, date of material, and personal judgment if the source appears to be authoritative (Burbules, 2001). Unfortunately these methods are not foolproof and can easily be taken advantage of by someone intentionally producing false information.

A recent study conducted by Stanford University Persuasive Technology Lab (SUPTL), "found that when people assessed a real Web site's credibility they did not use rigorous criteria, nearly half of all consumers (or 46.1%) in the study assessed the credibility of sites based in part on the appeal of the overall visual design of a site, including layout, typography, font size and color schemes" (Fogg et al., 2002). Consumer WebWatch, an affiliate of SUPTL, suggests five general guidelines to follow when determining credibility on the Web: Identity, Advertising and Sponsorships,

Customer Service, Corrections, and Privacy. Unfortunately each of these factors

captured less than 10% of the participants' attention within the Stanford study.

Overwhelmingly, participants evaluated credibility based on visual design. "This result

indicates that Consumer WebWatch, along with librarians and information professionals,

must increase efforts to educate online consumers so they evaluate the Web sites they

visit more carefully and make better educated decisions" (Fogg et al., 2002).

In the Stanford study subjects looked at a variety of content categories – ten in all

ranging from e-commerce, health, news, and search engines. An interesting factor was

that name recognition did tend to have a positive effect on Web sites credibility. "This

occurred more frequently in the e-commerce (25.9%), finance (21.8%), and news

(19.1%) categories" (Fogg et al., 2002). These findings do help support that name

recognition can influence how credibility is perceived on a Web site.

This is similar to the research which shows that identity, in the form of name

recognition, plays a strong role in determining credibility within online communities.

David Millen and John Patterson (2003) examined the effects of an identity policy for a

community network outside of Boston, Massachusetts. They concluded that the "identity

policy: bridged and enriched online and face-to-face interactions, promoted

accountability in support of local commerce, and fostered a social norm of polite

conversation" (Millen & Patterson, 2003). By having a name identity, users were more

accountable for their words (Millen & Patterson, 2003). It has also been shown that trust

in online environments is decreased when it is difficult to asses the motives of a user

(Marx, 1999). Without having to establish a policy of identifying oneself which would

govern the community as the one in the study, the blogging community has established

the protocol of identifying oneself because blogs are not only Web sites of content but also members of a community.

Burbules offers another concept of how to assess credibility on Web. He suggests that by linking Web sites together and collectively screening the addition of new material, [online communities] pool their intelligence and expertise to make credibility judgments and to cross-check one another. . . One might term this an instance of 'distributed credibility' in that it displaces an individual judgment with a collective intelligence (Burbules, 2001). By applying Burbules idea to the blogging community, this process of networking prevents one blog as standing as the one definitive authority. Instead the collective intelligence creates authority. It is from this distributed credibility that the blogging community provides a means of selection.

Applying Credibility to the Blogging Community

Unlike other online communities that may have developed based on anonymity such as MOO and MUDs, bloggers more often than not want their identity known. Most of the blogs that act as content filters clearly identify who is authoring and selecting the information on the site. If there is not a link directly provided to a description of the author, or a name visible on the site, often an email address is provided. There are a number of individuals who are famous and well received within their professional communities who have also gained fame within blogging communities and the blogosphere. In these instances the individual identity lends to the credibility of the blog, similar to Millen & Patterson's research. Of course, the identity of a blog's author does not have to be apparent on the blog, but can come from outside the blogosphere either through other media sources or by virtue of the social network.

The social network is another way that credibility emerges from a blog. This is because of the nature of the blogosphere to link between blogs. As more people link to a blog the author gains social capital. "Social Capital refers to features of social organization such as networks, norms, and social trust that facilitate coordination and cooperation for mutual benefit" (Robert, 2000). This capital has value in and out of the blogosphere. Wil Wheaton, a child actor for the television series Star Trek: The Next Generation, has made a name for himself inside and outside the blogosphere. "The blog has become Wheaton's portal into a new career as writer" (Gillmor, 2003). Links to Wheaton's blog can be found on a number of respected blogs within the blogging community such as Doc Searls, editor of Linux Journal. Also since this network exists inside and outside of the blogosphere, an anonymous blog may be added to a network because in the real world the author is known to the members of the network. The social network of a blog is visible through the use of a system called blogrolling, which is a list of semi-permanent links created on a blog. Often these links are divided into categories by topic or by how often they are read by the blogger creating the link. Most often the list appears in a sidebar next to the content of the blog. The blogroll becomes a salient marker of what other blogs are commonly read and linked to by the currently viewed blog.

From the blogroll a pattern emerges. The most popular and authoritative blogs list each other. "These new nodes on the net are perhaps more analogous to year-round conferences" (Block, 2001). Leaders in the field quickly rise to the top of the blogroll and are linked to most frequently. The network that is built though professional organizations and conferences is maintained through the blog community and reflects the

credibility of that community. The other way the community is maintained is through

commenting on individual blogs.   Authors can add a feature to their blog that allows

readers to comment about their most recent posts. "Credibility, authority, and

accountability take the form of feedback and linking" (Carver, 2003). Heated debates

and lengthy discussions have taken place in the comment section of blogs. If someone

posts erroneous, information the author's mistake will quickly be brought to his/her

attention, therefore keeping the information provided within the blog accurate and

credible.

     This feature of commenting within the blogs also establishes credibility

throughout the community through a form of informal peer review.   "In an essay in the

May issue of Reason magazine, Mr. Sanchez [a staff writer at the Cato Institute] noted

that blogging permitted the investigation of John R. Lott Jr. [a gun researcher who has

been accused of inventing the results of a telephone survey] to proceed much more

quickly than a controversy two years earlier about the former Emory University historian

Michael A. Bellesiles, whose book *Arming America: The Origins of a National Gun

Culture* (Alfred A. Knopf) was ultimately exposed as error-ridden" (Glenn, 2003). The

network of links and feedback worked to prevent academic misconduct.

     The network of bloggers can also capture and correct information that is not

covered by mainstream news. One of the most noted instances of bloggers capturing

egregious behavior is the comments Trent Lott made at Strom Thurmond's 100th

birthday. "Most major media outlets ignored the remark but online journalists, especially

Webloggers such as Josh Marshall, Andrew Sullivan, and David Frum posted scathing

attacks on Lott (with the latter two being conservatives)" (Glaser, 2002). Blogs kept

Lott's story fresh though linking and posting and even researched the subject further.

"Atrios [an anonymous blogger] (www.atrios.blogspot.com), found a 1948 Thurmond

campaign document telling voters that electing his rival, Harry Truman, would mean

[that:] 'anti-lynching and anti-segregation proposals will become the law of the land and

our way of life in the South will be gone forever'" (Burkeman, 2002). All of the

information gathered in the blog community led to the main stream media finally picking

up the story and Lott's eventual resignation as Senate Majority Leader. In this way the

network of blogs was able to enforce the ethic of credibility outside the blogosphere.

Blog Aggregators

The first of the three aggregators that will be examined is Blogdex. Blogdex is

the brainchild of Cameron Marlow and is a research project in the [MIT Media

Laboratory](MIT Media Laboratory). It was one of the first of the weblog search engines and was brought on line

in 2001. At that time Blogdex had about 9,000 weblogs in its database (Kahney, 2001).

The websites collected by Blogdex were originally culled from lists of weblogs available

at the time of its creation, but it has since grown to over 30,000 weblogs. This is because

of Marlow's work, and because of an opt-in service for any weblog who wishes to

participate. "Blogdex uses the links made by webloggers as a proxy to the things they are

talking about" (Marlow, 2004). Blogdex uses a similar technique to Google by ranking

each link so that as more bloggers link to the same thing, the link bubbles to the top of the

list, and is displayed on the homepage (Kahney, 2001).

The second aggregator used in this study is Daypop which has a database of over

59,000 sources. It extends into news and RSS feeds, but still acquires the bulk of its

information from the blogosphere. In an article written in 2003, when the database

contained 35,000, sources only 1,000 of them were news sources (Price, 2003). Daypop was launched in 2001 by its creator and sole maintainer Dan Chan. For the first year, Chan hand-picked the blogs that were put into his database. When he found a well written blog, he put it into Daypop. From the beginning Chan wanted to provide a way to look at the activity occurring in the blogosphere though links. "Link Analysis started with the creation of the Top 40 page shortly after Daypop launched. . . The Top 40 gives more weight to links that have recently been created. This means only fresh newly discovered links make it to the Top 40" (Price, 2003). It is from this list that the analysis for this study will be taken.

The third and final aggregator being used for this study is one of the newest BlogPulse, which was created by Intelliseek in February 2003. A seed list of 22,000+ weblogs from the Blogstreet directory was used to begin the accumulation of weblogs. In June when the list reached a total of about 100,000, the process of actively collecting stopped, "because [they] were reaching an upper bound on the number of weblogs that [they] could politely crawl within 12 hours on one server. In addition, given that [the] list includes the most oft-cited blogs, [they] felt that the set of 100, 000 represented a suitably representative cross-section of the discussion occurring in the blogosphere" (Glance, 2004). They still offer individuals to add their blogs to the database and the database is frequently purged of weblogs which have no post since the last purge. Currently the information on the BlogPulse site states, "BlogPulse locates content from more than 1 million blogs and indexes them on a regular basis" (2004). This is a much larger database than the other two aggregators. BlogPulse also has a Top Links feature which "are the most cited or most popular links appearing in blog entries daily. Top Links can

give you an idea of sources, stories and themes that have occupied the attention of

bloggers on any given day" (2004).

Each of these aggregators offer a feature that selects the most linked to content for

the day. The structure of each aggregator differs slightly and the method that the links

are collected invariably differs though exact information about which algorithm the sites

use is not available. From these lists the content will then be compare to archival criteria

to see if the content selected by the blogging community is of archival quality, and if so

which of these aggregators may be accomplishing a better job of selection.

Archival Criteria

Tibbo (2001) states that "appraisal theory and practice, along with life cycle of

records, can facilitate the retention of materials of enduring value. While archivists are

known as great savers, in reality, they are highly skilled selectors, generally retaining no

more than 5% of the original bulk of any collection" (Tibbo, 2001).   Yet, how archivists

make these choices is not easy and rely as much on theory as art. "Archivists engage in

heated debates about appraisal criteria and methodologies" (Eastwood et al., 2000). One

method is to examine the records for continued value such as "their usefulness for legal

purposes, their value as evidence of the functioning and organization of their creator, or

their potential for research" (Eastwood et al., 2000). These themes are common when

reviewing appraisal policies across archives, as archives are usually part of a larger

institution such as the state or a university. This relationship often guides the collection

policy of the archive limiting the scope of the collection. Looking at this statement in

terms of the Web, only one proponent of usefulness applies to the broad area of Web

documents. Not often are they created solely for legal purposes. Determining which

documents have value for research is too difficult a task with such a broad medium.

Primarily Web documents serve as evidence of the functioning and organization of their

creator, whether it is an individual or a larger organization. This evidence is reflected in

the national efforts to preserve content from the Internet as mentioned earlier. These

efforts have limited the information they are preserving to that produced by their nation

state because they are preserving evidence of their nation. Unfortunately, the Appraisal

Task Force stopped short of providing general guidelines for selection because those

decisions are so deeply governed by the preservation institution.

One set of criteria Abby Smith discusses from the University of Michigan's

policy. This policy, "aims to fit digitization into the context of traditional collection

development."

- Is the content original and of substantial intellectual quality?
- Is it useful in the short and/or long term for research and instruction?
- Does it match campus programmatic priorities and library collecting interests?
- Is the cost in line with the anticipated value?
- Does the format match the research styles of anticipated users?
- Does it advance the development of a meaningful organic collection? (Smith, 2001).

Again it is difficult to determine the intellectual quality of information produced

on the Internet, and it is difficult to determine if it is of substantial research value. Unlike

scholarly information produced in the university context which is held to a much higher

standard, information can be produced on the Web by anyone who has access. This is at

the same time the greatest opportunity of the Web, and is also its greatest hindrance when

working from the Archival perspective.

Brewster Kahle once said that the Internet was a medium for artifacts that are

ephemeral (Edwards, 2004). Richard Stone suggests that "perhaps the Internet represents

the Ultimate Ephemera, the Ultimate Junk Mail, as it displays characteristics of print

ephemera to an intense and heightened degree" (R. Stone, 1997). The classic definition

of ephemera comes from Maurice Rickards – "The minor transient documents of

everyday life" (2000). If we begin to view Web documents as we would ephemera,

parallels can then be drawn which assist with making assessments about the long term

value of information produced in both media. Ephemera are viewed as being transitory,

meeting a need which passes quickly, or it is simple disposable. Ephemera can be seen as

purpose driven such as public education, dissemination of a policy, advertising or event

based such as a crisis conceived hastily called protest meeting. Ephemera is especially

vulnerable, because an item which initially is widely available quickly becomes fugitive,

and ephemera is pervasive (R. Stone, 1997). Information on the Web is much the same

being produced quickly and removed often without warning. It can be used to promote a

certain idea or provide specific information. It also can be used as a resource for

information about a company or institution. It is often seen as the most accessible source

despite its tendency to change and the dependency on a computer to have access.

The National Library of Australia policy on collecting ephemera from 1997 is

based on four basic principles. These principles are the item should contain "(1) a

significant amount of factual or descriptive information; significant visual elements such

as design, portraiture, (2) material has to be generally on a level which has wider

applicability; it should have a resonance beyond the local source, and (3) the material

[should be] an exemplar of its type for reasons of design, language, topic, or origin"(R.

Stone, 1997). It is through this combination of content and design that the National

Library of Australia collected ephemera. Web information also contains this combination

of design and content which is why it is easy to compare these principles to content provided on the Web. Ephemera should "convey the spirit of an occasion or period evocatively though their content, language, and graphic style" (Beaumont, 2003). This could also be said of the evolving content on the Web which should be preserved for posterity's sake (Beaumont, 2003).

After examining several policies and criteria applying to digital collection, general collection and ephemeral collections, I developed five main criteria which will be used in comparison to the data collected.

Criterion A, "Is the Information Original or Unique?"

This criterion would be applied in terms of the source of the link. Is the link coming from a source that is independent or unique, not from an entity that has content which is similar to a variety of other sources? The underlying concern would be if the content is produced by an "independent" source then that content could be more vulnerable to loss because of the lack of institutional support.

Criterion B, "Is the Information Documenting Issues of Current Social or Political Interest?"

This criterion would be applied in terms of the date assigned to the content of the link and current would be defined as having occurred in the last thirty days from the time of the link being registered. An example would be a link referring to an article in the news that was a week older than the day the link registered on the aggregator. The content would still be considered current.

Criterion C, Does the Information have a Wider Application, a Resonance beyond the Local Community?

The information should have appeal to individuals outside the blogging community. The information still may be considered to have a limited interest such as only for computer programmers or science fiction fans, but the community of interest should exist outside of the blogosphere.

Criterion D, Is the Information Exemplar of Its Type for Reasons of Design, Language, Topic, or Origin?

This criterion will be applied in terms of the content of the link such as whether the link is of literary quality, artistic in its design, or is in depth about a topic.

Criterion E, Does the Information Advance the Development of a Meaningful Organic Collection?

This criterion will be applied in terms of the links contribution to the collection as a whole. Content is considered part of a meaningful organic collection not only from the information contained in the individual item but also for the relationship that exists between the item to the collection as a whole. Therefore, this criterion reflects back to the other criteria in that if a link fulfills some of the other criteria then the final inspection is to whether it would contribute to the collection as a whole.

Methods

The data for this study was collected from June 14 to June 21. Each aggregator was accessed between the time of 9:30 and 10:00 p.m., and the list which indicated the most linked to content for that day was saved on the researcher's computer. It was discovered during this time that BlogPulse could not be saved in this manner, though this proved to be acceptable because BlogPulse provides a listing of the past top links for the previous month on their Web site. From these lists only the top 20 links were analyzed.

The links were then accessed a week later to allow for the possibility of link degradation. The research compared each link to the archival criteria in the previous section to determine if the link referred to content that met the values of the criterion as determined by the researcher. A tally was kept for each aggregator to determine which one contained the greatest amount of content that fulfilled each criterion. A binary code was used to populate the tally – one count for yes and zero count for no. Links that were part of large sites that were to the same content but consisted of different path addresses were considered the same link and the content was not counted twice. Links that were broken and didn't produce any results were marked as such, and links that appeared to be to spam were also denoted. Spam in this study was defined as any link that was to a site that did not seem to be intentional selected by another person but rather machine-generated. A record was also kept of which links appeared in more than one aggregator, and how often that link appeared in the same aggregator.

Repeated links have an interesting effect in this study. While they do not provide new content for analysis, they do offer another aspect of analysis which is burst activity. The blogosphere is noted as a "network of small but active micro-communities" which "exibit[s] striking temporal characteristics" (Kumar, Raghavan, Novak, & Tomkins, 2003). "Within a community of interacting bloggers, a given topic may become the subject of intense debate for a period of time, and then fade away. These bursts of activity are typified by heightened hyperlinking amongst the blogs involved" (Kumar et al., 2003). Kumar and his colleagues' research was based on the algorithm developed by Jon Kleinberg. Kleinberg's earlier research "focus[ed] on the use of links for analyzing the collection of pages relevant to a broad search topic, and for discovering the most

"authoritative" pages on such topics" (Kleinberg, 1999). Kleinberg developed an algorithm designed to identify hub and authoritative pages when searching. He based his research on hyperlinks because, "hyperlinks encode a considerable amount of latent human judgment, and we claim that this type of judgment is precisely what is needed to formulate a notion of authority" (Kleinberg, 1999). There is simply no way of systematically measuring or properly assessing a pages authority (Kleinberg, 1999). In a later project, Kleinberg looked for "the appearance of a topic in a document stream [which] is signaled by a 'burst of activity' with certain features rising sharply in frequency as the topic emerges" (Kleinberg, 2002). As a link is repeated creating a burst it could be considered an indication of authority or quality of content and therefore be selected for preservation from this method as well. When the algorithmic application is applied to the blogosphere as Kumar and his colleagues were able to observe, a combination of human and machine generated authority occurs. The content that appears on a blog has been through a process of selection by human judgment. Once the content is posted the distributed network the blog will select the most relevant and credible information. The increased number of links to that content will cause it to be selected by the aggregators and the higher quality and more useful links will appear over a period of time. Since access to the algorithms being used by each aggregator is not available, it can not determine which, if any, are using one similar to the one developed by Kleinberg. A rough estimate will be made from counting the number of links repeated within each aggregator and across the three aggregators.

Limitations

Due to limited time and resources, this study did not allow for a test of intercoder reliability. Therefore bias maybe inherent in the data. The archival criteria were selected from a variety of separate policies and institutions though the choice of one criterion over another may have been influenced by the researchers' preconceived notion of the collection being analyzed. This bias is further compounded by the fact the results were produced from only the researcher's interpretation of the criteria in comparison to the links collected.

Also, the Blogosphere is limited by its users which may be of only a limited demographic. It is difficult determine how representative the blogosphere is of the Web because there is no clear demographic information about who is blogging. Inherently the blogosphere is distributive in nature, and a large number of communities are represented. I hope the wide variety of topics and interests shared by the blogging community will off set the lack of representativeness of the blogging community.

Results

Table 1. Percentage of content selected by the aggregators that fulfilled the criteria from total number of links:

Archival Criteria	Criteria A: Unique	Criteria B: Current	Criteria C: Wider Community	Criteria D: Exemplar	Criteria E: Meaningful, Organic
Blogdex	38%	69%	58%	7%	61%
Daypop	36%	60%	59%	9%	51%
BlogPulse	24%	56%	64%	3%	58%

Table 1 represents the criteria and the percentage from the links of the aggregators for the eight days of the study that fulfilled the criteria for each aggregator. As expected, all of the aggregators were successful at selecting content that documented issues of current social or political interest, and those topics which had an application to an audience wider than the blogging community. The ability of the aggregators to do this prompted at least half of all the content to be selected as part of a meaningful and organic collection. Some of the content selected from all the aggregators was a story about President Bush proposing a mandatory mental health assessment for everyone in the United States from the British Medical Journal (bmj.bmjjournals.com), several references to the first privately funded space flight which took place on June 21 (www.space.com), an article about Michael Moore's latest movie release Fahrenheit 9/11 (www.foxnews.com), Gmail (gmail.google.com), Google's answer to email had a presence in each of the aggregators which influenced the exemplar content found in Daypop.

Daypop had the highest percentage of exemplar content at 9% which was significantly higher than BlogPulse, at only 3%. This was surprising considering the low amount of content that was finally considered to be meaningful for building the collection on Daypop. Of course this could be a reflection of the hand-picked factor for the Daypop database. Less meaningful information, but what is selected is of higher quality. Some examples of what was found to be exemplar were programs or scripts that were written to openly available on the Web sites that were linked. Two that were picked up by Daypop were www.marklyon.org, a program so that email from other clients could be imported to Gmail, and torrez.us, another program so that users would be alerted when new mail had

arrived to their Gmail inbox without logging in each time. These pieces of content are examples of the artistry of computer programming, and are examples of how truly open the environment of the Web can be.

BlogPulse selected the highest percentage of content that applied to a wider community with 64%, with Daypop and Blogdex having about an equal share at just less than 60%. This is not surprising since BlogPulse is casting the widest net with the largest database being spidered each day for link activity. With a larger pool to select from, it would be expected that more topics of interest would bubble to the top of the lists. However, BlogPulse did poorly on content being of exemplar or even unique quality, ranking in the lowest percentage for both criteria.

Blogdex had the highest level of current content with 69%, which was significantly higher than BlogPulse at 56%. Blogdex seemed to be picking up more content that was news and current event related, even though Daypop uses RRS feeds and other news sources. BlogPulse had the lowest percentage of currency but this was affected by many of the links pointing to a site that did not have frequent updates where the date of the content could not be determined. If the study were repeated, the concept of "current" could be broadened or related to the time the referencing link was made.

**Table 2. Number of usable links for content analysis:**

Aggregators	Normal Links	Repeated Links	Broken/Spam Links	Percentage of usable Links
Blogdex	123	21	16	77%
Daypop	113	36	11	71%
BlogPulse	112	48	0	70%

There was evidence from all of the aggregators of their ability to select content that was current and had meaning beyond the blogging community. Over 70% of links were chosen from each aggregator. Given that 20-30% of the links on each aggregator were either repeated, broken or to spam content, the selection of over 50% of the links as part of a meaningful and organic collection reinforces the idea that the blogging community is able to select items of archival quality which would build a representative collection.

The repetition of links was tracked not only to prevent content from being assessed twice but also to ascertain if any pattern existed. Links often repeated within the aggregators themselves over several days and those same links would appear across the aggregators. Surprisingly, of the 480 links analyzed only 59 links were repeated in more than one aggregator and of those only 22 were repeated in all three. This is a repetition rate of only about 11% and only consistent repetition to all the aggregators, 5% of the time. This provides evidence there is great diversity in the content being selected across aggregators.

**Figure 1. Number of Repeated Links
Across Aggregators**
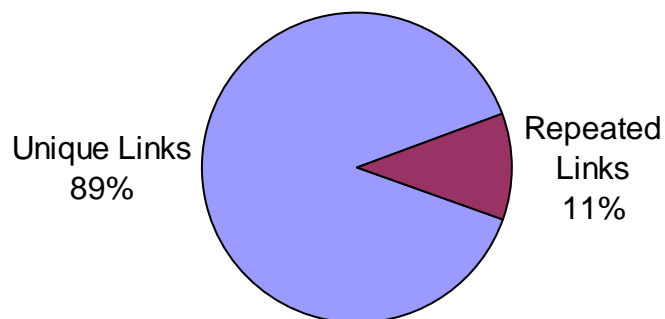
Discussion

During the time of the study, Blogdex suffered a number of broken links from the shutdown in Weblogs.com and subsequent server change by Dave Winer. Winer, who had offered free hosting to bloggers for the past four years, cited equipment trouble, financial costs, and personal stress as the reasons for the shut down of the service (Delio, 2004). It is unclear how many of these blogs were part of the Blogdex database, but for the first few days of data collection the links created from the shut down occupied the most of the top rankings. There was also a number of links that were about the shut down which appeared in the list. This story did not have as much prevalence in the other two aggregators. It was surprising to find that this discussion did not hinder Blogdex's ability to select content that was on par with the other two aggregators and even excelling in the realm of currency and uniqueness. In an overall assessment Blogdex seems to the best though the removal of spam links would increase its performance.

Daypop also seemed to suffer from spam and link exploitation. On a few instances there were links to business advertising services that looked more like ads than information that were being pointed to for interest. On another occasion there was a rash of links from Yahoo news which had expired causing broken links though all of these links were cited from the same five blogs. When these blogs were researched none of them had dates in their post list corresponding to the date that had appeared on the list. It became easy to recognize these problem links which were few and when found were discarded.

Both, Daypop and Blogdex have been around longer so it is probably inevitable that the problems arise from spam. Daypop also had difficulty updating its site as

frequently as expected. It was four days into collecting data before I was realized that the time stamp was intended to be a twenty-four hour stamp. The pages for the study were being saved after 9 p.m. every night. However, only two pages out of the eight were saved after 18:00, none of the pages registered a time stamp of 21:00 or later. Daypop states that the weblogs are crawled and updated every twelve hours and the news sites every three hours. Daypop did offer other feedback on the site which indicated that all of the links used for the analysis came from blog sources which was a concern in the beginning because Daypop did include new sources and RSS feeds within its database.

BlogPulse yielded a different sense of the Web than the other two aggregators. BlogPulse cast a wider net and indeed seemed to capture a wider variety of content which often did not fall under Criteria B., pertaining to current events. Often it was a site of popular social interest of which the content could not be easily dated to determine currency. BlogPulse also had the most repeated links across the aggregators with 30% compared to only 23% in Daypop and 13% in Blogdex. This was surprising since BlogPulse was drawing from a much larger pool. One might have an expectation that the smaller more established aggregators would have a greater number or repeated links because their communities represented in their database would more likely be linking back to each other. Finding that the larger database provided the greater number of repeated links, reinforces the ideas that blogging communities are spreading information outside their boundaries and that the blogosphere is becoming a representative community.

Further Study and Conclusions

This study looked at three aggregators available on the Web to determine if the most linked to content within their databases could be considered for preservation. Each of the aggregators was able to indicate a high number of links that were of archival quality particularly in the quality of currency and with the appeal to a wider audience. An improvement to the aggregators would be to address the ability to select a greater amount of content that is unique and of exemplar quality. This improvement could be made by giving more weight to the community and to individuals in the community.

The content selected did not frequently overlap in each aggregator giving evidence that the blogosphere is a diverse group of individuals who are selecting content from the Web in mass. A further study of the citations or even closer examination of the database used by the aggregators would give some indication as to whether the same blogs are occurring in each of the aggregators and how much overlap may occur at this time at this time. For the reason that some of those individuals come from expert communities and hold a high level of credibility or influence in the community a spider giving greater "weight" to the content they link to could improve the level that which is exemplar and unique. The ability to create a spider which could crawl more of the blogosphere for links while also weighting links that came from blogs who have expert knowledge or greater influence could help create a richer collection of content while still keeping at an amount which is manageable for preservation.

Currently, there is not a very strong system of selection in place when looking at the Web as a whole. The proposed solution of using the blogging community is a radical departure from the centralized method of selection that normally takes place within the

Archival community. Yet, it may be the best suited solution when applied to a medium like the Web and digital objects. Blogs are native to the Web and the community of blogs is growing to represent society as it exists within the digital realm.   A decentralized and democratic method of selection may be the only way to manage the glut of information being produced digitally.

With the Internet Archive growing at a pace of 12 terabytes a month, will it economically feasible to continue saving everything? Archives have had a strong tradition of selection throughout the ages. Simply because digital records occupy less physical space does not mean they should not go through the same process. The value of a preserved record is in its usefulness and accessibility not in sheer volume.

It is said that there is strength in numbers and so it is with blogs. Individually blogs may not hold much value; it is the sum of the community which presents value. It is when the community as a whole is observed that the medium presents value. Blogs are native to the Web. They perform the task of filtering and selecting content found on the Web and because of their nature of community building this selection becomes representative of the social culture on the Web. The ephemeral and significant are captured through the democratic process. An archive of the blogging community therefore serves society as a whole, and would represent the institutional knowledge of the Web.

References

Baker, K. (2003). *Nobody's doll*. Retrieved December 3, 2003, from

    http://www.nobodysdoll.com

Barrett, C. (2003). *About*. Retrieved December, 10, 2003, from

    http://www.camworld.com/about/

Beaumont, S. (2003). *Ephemera: the stuff of history*: Chartered Institute of Library and

    Information Professionals.

Berners-Lee, T. (1998). *A One-page personal history of the web*. Retrieved December, 7,

    2003, from http://www.w3.org/People/Berners-Lee/

Block, M. (2001). Communicating off the page. *Library Journal, 126*, 50-53.

Blood, R. (2002). *We've got blog*. Cambridge, MA: Perseus Books Group.

Boiko, B. (2002). *Content Management Bible*. New York, NY: Hungry Minds Inc.

Burbules, N. C. (2001). Paradoxes of the Web: the ethical dimensions of credibility.

    Library Trends, 49(3), 441-453.

Burkeman, O. (2002). Bloggers catch what Washington Post missed. *The Guardian*.

Carver, B. (2003). Is it time to get blogging? *Library Journal, 128*, 30-33.

Conway, P. (1996). *Preservation in the digital world* (No. Pub 62). Washington, DC:

    Council on Library and Information Resources.

Conway, P. (1999). *The relevance of reservation in a digital world* (No. Section 5,

    Leaflet 5). Andover, MA: Northeast Document Conservation Center.

Delio, M. (2004). *Thousands of blogs fall silent*. Retrieved July 10, 2004, from

http://www.wired.com/news/culture/0,1284,63856,00.html

Eastwood, T., Craig, B., Eppard, P., Gigliola, F., Normand, F., Giguere, M., et al. (2000).

Appraisal Task force report. Vancouver: InterPARAES.

Edwards, E. (2004). Ephemeral to enduring: the Internet Archive and its role in

perserving digital media. *Information Techology and Libraries, 23*(1), 3-8.

Fogg, B. J., Soohoo, C., Danielson, D., Marable, L., Stanford, J., & Tauber, E. (2002).

How do people evaluate a Web site's credibility? Palo Alto, CA: Persuasive

Technology Lab Stanford University.

Gill, K. E. (2004). *How can we measure the influence of the blogosphere?* Paper

presented at the WWW2004, New York, NY.

Gillmor, D. (2003, October 8). Blog has become former actor's portal into new career.

San Jose Mercury News.

Glance, N., Hurst, M., and T., Takashi. (2004, May 17-22). *Blogpulse: automated trend

discovery for weblogs*. Paper presented at the WWW2004, New York, NY.

Glaser, M. (2002, December 17). Weblogs credited for Lott brouhaha. *Online Journalism

Review*.

Glenn, D. (2003). Scholars who blog. *Chronicle of Higher Education, 49*, 14-17.

Griffey, J. (2004). *Pattern recognition*. Retrieved June 28, 2004, from

http://www.ibiblio.org/griffey/wp/index.php?cat=8

Hiler, J. (2002). *Blogs as disruptive tech*. Retrieved December 6, 2003, from

http://www.webcrimson.com/ourstories/blogsdisruptivetech.htm

Intelliseek's BlogPulse. (2004). Retrieved July 1, 2004, from http://blogpulse.com/

Jones, M. (2003). *Interview: Six Apart's degree of weblog integration; blog tools vendor positions itself for enterprise growth.* Retrieved November 9, 2003, from http://www.InfoWorld.com

Kahney, L. (2001). *Tracking bloggers with Blogdex*. Retrieved July 2, 2004, from http://www.wired.com/news/culture/0,1284,45546,00.html

Kleinberg, J. (1999). Authoritative sources in a hyperlinked enviroment. *Journal of the ACM, 46*(5), 604-632.

Kleinberg, J. (2002, July 23 - 26). *Bursty and hierarchical structure in streams.* Paper presented at the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada.

Kumar, R., Raghavan, P., Novak, J., & Tomkins, A. (2003, May 20 - 24). *On the bursty evolution of blogspace.* Paper presented at the International World Wide Web Conference, Budapest, Hungary.

Marlow, C. (2004). *About*. Retrieved July 2, 2004, from http://blogdex.net/about.asp

Marx, G. (1999). What's in a name? *The Information Society, 15*(2).

Millen, D. R., & Patterson, J. F. (2003, April 5-10). *Identity disclosure and the creation of social capital.* Paper presented at the CHI 2003: New Horizons, Ft. Lauderdale, Florida.

Notess, G. (2002). The blog realm: New sources, searching with Daypop, and content management. *Online, September/October*, 70-72.

O'Shea, W. (2003, July 7). The online journals known as Web logs are finding favor as an efficient way to communicate within the workspace. *The New York Times,* p. 3.

Perrone, J. (2004). *What is a weblog?* Retrieved July 5, 2004, from

    http://www.guardian.co.uk/weblogarticle/0,6799,394059,00.html

Phillips, M. (2003). *Online Austrailian publications: selection guidelines for archiving*

    and preservation by the National Library of Australia. Retrieved June, 17, 2004,

    from http://pandora.nla.gov.au/selectionguidelines.html

Price, G. (2003). *Behind the scenes at the Daypop search engine*. Retrieved June 2, 2004,

    from http://www.searchenginewatch.com/searchday/article.php/2209031

Quick, W. (2001). *DailyPundit*. Retrieved July 12, 2004, from

    http://www.iw3p.com/DailyPundit/2001_12_30_dailypundit_archive.php#831512

    0

Rickards, M. (2000). *The encyclopedia of ephemera: a guide to the fragmentary*

    documents of everyday life for the collector, curator and historian. New York:

    Routledge.

Robert, P. (2000). *Bowling alone: the collapse and revival of American community*. New

    York: Simon & Schuster.

Schroeder, R. (2003). *One path to the blog*. Retrieved November 25, 2003, from

    http://www.elearnmag.org/subpage/sub_page.cfm?section=3&list_item=14&page

    =1

Schwartz, G. (2003). *Blogs for libraries*. Retrieved November 9, 2003, from

    http://webjunction.org/do/DisplayContent;jsessionid=25756A4172E7D9DE9C77

    93638F013B5D?id=767

Searls, D., & Sifry, D. (2003). Building with blogs. *Linux Journal, 2003*(107), 4.

Smith, A. (2001). *Stategies for building digitized collections* (No. 101). Washington:

Council on Library and Information Resources.

Stone, B. (2004, July 12, 2004). *Knowledge*. Retrieved July 14, 2004, from

http://www.blogger.com/knowledge/2004/07/amazing-web-site-machine.pyra

Stone, R. (1997). *Junk as heritage: the collecting of printed ephemera on a national

scale*. Canberra: National Library of Australia.

Tepper, M. (2003). The rise of social software. *netWorker, September,* 19-23.

Tibbo, H. (2001). Archival persepectives on the emerging digital library.

Communications of the ACM, 44(5), 69-70.

Turner, S. (2003). Favorite library news and blog sites. *Mississippi Libraries, 67*(4), 122-

123.

Winer, D. (2002, May 17). *The history of weblogs*. Retrieved July 17, 2004, from

http://newhome.weblogs.com/historyOfWeblogs

Young, T. (2003). Blogs: is the new online culture a fad or the future? *Knowledge Quest,

31*, 50-51.