Blogs as a Means of Preservation Selection for the World Wide Web

Marianne K. Gouge

School of Information and Library Science, University of North Carolina at Chapel Hill, CB#3360.100 Manning Hall, Chapel Hill, NC 27599. mgouge@email.unc.edu

Currently, there is not a strong system of selection in place when looking at preserving content on the Web. 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 these aggregators to determine if any content which is linked to 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.

Introduction

For the past several years archivists and preservationists have been struggling with the preservation of digital records. The digital environment has a unique set of problems pertaining to preservation. Digital objects are more ephemeral than printed media and they are part of a networked community of shared information. "A hypertext link can point to anything, be it personal, local or global, be it draft or highly polished" is Tim Berners-Lee's vision of the Web (1998). How then do you capture this diverse and distributed collection of information? One method is to try and capture it all and save a replica of the Internet for future generations, which is the current method being used by the Internet Archive. The Internet Archive contains over 300 terabytes of data and is currently growing at a rate of 12 terabytes per month (Edwards, 2004). Another method is through selection of specific materials, 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. The process of selection adds value to materials because "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). It would take an enormous staff of these professionals to review the documents created on the Web to determine their preservation value.

However, there is a community which exists on the Web which contains professionals from a variety of disciplines already sorting though the information provided on the Web. This is the community of webloggers. 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 (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). The collection of weblogs is known as 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). Blogs offer the service of choice or selection because of their hypertextal and distributed nature. Blogs make value judgments by selecting content and linking to it. Currently the content from these communities is being harvested by a few sites that aggregate the most linked-to content. Are blog aggregators selecting content which is of archival quality?

Purpose of Study

The purpose of this study is to compare blog aggregators for their ability to collect digital content which is of archival quality. The Web is vastly unfixed collection of digital objects. Researchers from Hungry and the United States found that access to digital objects on regularly updated sites significantly decays after 36 hours of posting (Almaas et al., 2005). This creates the need to capture objects for preservation on a daily basis. To eliminate the need to store an entire copy of the Web each day the method of selection will need to also take place on a daily basis. The three aggregators selected for this study were chosen because each aggregator collects the most linked-to content for the day from the portion of the blogosphere sampled by each aggregator. This content is then ranked by the common method of link counting and displayed in a list that can be captured for analysis. It should be noted that the method and sources used to determine this content varies from aggregator to aggregator. The three aggregators selected were Daypop, Blogdex, and BlogPulse.

All three aggregators were selected because they publish a list of the most linked-to content for the day, compiled from the sites' databases. .Daypop and Blogdex are two established sites that collect this list of most linked to information in the blogosphere. The other established blog sites such as Technorati, do not offer the same feature. Technorati categorizes the content - by news or current event, for example - instead of by a straight system of ranking. BlogPulse is a newer less established site, but has a body of research to support the work the site is doing which was presented at WWW2004 in New York City (Glance, Hurst, & Takashi, 2004). Cameron Marlow, Blogdex's creator, also presented at this conference (Ceglowski & Marlow, 2004).

This study will analyze the content selected by the aggregators to determine what among this content is of archival quality. Archival quality will be determined according to criteria assembled from a variety of archival policies and principles. The selected content 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 any particular strengths or weaknesses 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 her blog as part of 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 be collected by the aggregators from their sample of the blogosphere.

Blog Aggregators Used

The first of the three aggregators that will be examined is Blogdex. Blogdex (http://blogdex.net/) is the brainchild of Cameron Marlow and is a research project in the MIT Media Laboratory. It was one of the first of the weblog aggregators and was brought on line in 2001. Currently Blogdex "crawls more than 30,000" blogs (Terdiman, 2004). This is because of Marlow adding blogs to Blogdex's database and because of an opt-in service for any weblog that wishes to be indexed. "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 content, the link bubbles to the top of the list, and is displayed on Blogdex's homepage (Kahney, 2001).

The second aggregator used in this study is Daypop (http://www.daypop.com/) which has a database of over 59,000 different sources, which include blogs, news websites, and RSS feeds. In an article written in 2003, when the database contained 35,000 sources, only 3% (1,000) of them were news sources (Price, 2003). Daypop was launched in 2001 by its creator and still is its sole maintainer Dan Chan. For the first year, Chan hand picked the blogs that were entered into Daypop's database. When Chan found a well written blog, he entered 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 (http://www.blogpulse.com/), which was created by Intelliseek in February 2003. A seed list of 22,000+ weblogs from the Blogstreet directory was used to begin BlogPulse's accumulation of weblogs. In June 2004 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 database] 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 et al., 2004). BlogPulse still offers individuals the ability to add their blogs to the database and the database is frequently purged of weblogs which have added no new posts 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 offers a feature that selects the most linked-to content for the day. The functionality of each aggregator differs slightly and the method by which links are collected invariably differs, and unfortunately it is not possible to know what method for collection is used because exact information about the algorithm each sites uses is not available. The lists of the most linked-to content from these three aggregators' will be compared to see if the content selected by the blogging community is of archival quality, and if so which of these aggregators may be performing better the task of archival selection.

Archival Material

Helen 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 the 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 (Eastwood, et al., 2000).

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" (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 (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" (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 is applied to the source of the content. Is the content from an independent or unique source, or from an entity that has content which is similar to a variety of other sources? The underlying concern is that if the content is produced by an "independent" source then that content could be more vulnerable to loss due to a lack of institutional support.

Criterion B, Does the Information Document Issues of Current Social or Political Interest?

This criterion is applied to the date assigned to the content. Currency is defined as the content having been created in the thirty days prior to the date of the link being registered with the aggregator. An example would be a link referring to an article in the news that was published a week prior to the link being registered. As it was state before digital information tends to have a very short lifespan possibly lasting only 36 hours.

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

The content should have appeal to individuals outside the blogging community such as not being a personal diary entry or an entry about the blogging community itself. The content still may be considered to have a limited interest such as only for computer programmers or science fiction fans, but the community of interest will exist outside of the blogosphere.

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

This criterion is applied to the style of the content such as whether the content is of literary quality, artistic or creative 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 is applied to the content's contribution to the collection as a whole. Content is considered part of a meaningful organic collection taking into account both the information contained in the individual item, and the relationship that exists between the item to the collection as a whole. Therefore, this criterion is dependent on the other criteria in that if a link fulfills some of the other criteria then the final collection development decision is whether it would contribute to the collection as a whole.

Methods

The data for this study was collected during a one week period: June 14, 2004 to June 21, 2004. Each aggregator was accessed between 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 that BlogPulse could not be saved in this manner due to the structure of the Website prevented downloading of the pages needed, though this proved to be acceptable because BlogPulse provides a listing of the past top links for every day of the previous month on their Web site. From these lists only the top 20 links were analyzed to normalize the length of the lists.

The links were then accessed a week later to allow for the possibility of link degradation known as link rot. The research compared each piece of linked-to content to the archival criteria discussed in the previous section to determine if the content met these archival criteria. 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 collected by the same aggregator to the same content on the same site but which had different URLs were considered the same link and the content was not counted twice. Links that were broken and content that appeared to be to spam were denoted. Spam in this study was defined as any link that was to content that did not seem to be intentionally selected by the blog author. A record was also kept of which links appeared in more than one aggregator, and how often that link appeared in the same aggregator.

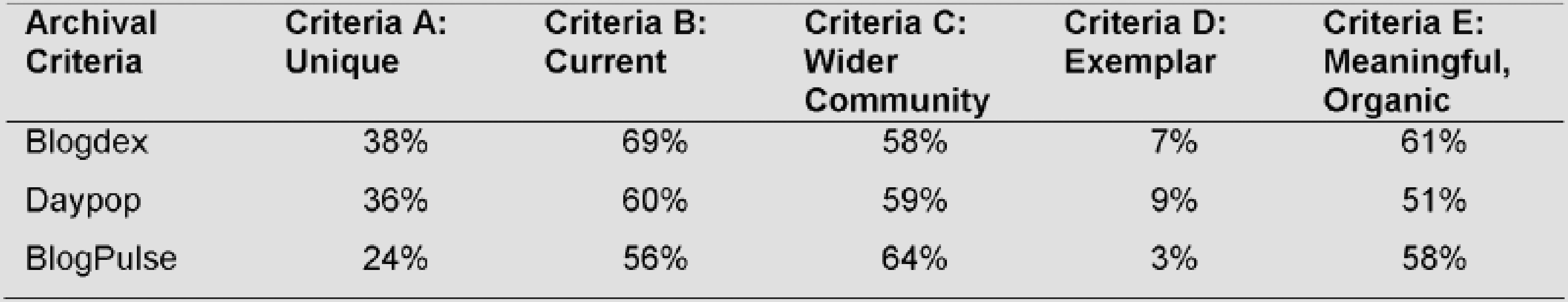
Limitations

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 blogosphere. Further, due to limited time and resources, this study did not allow for a test of intercoder reliability, and so interpretation of the criteria in comparing the content collected was the researcher's alone.

Also, the Blogosphere is limited by its authors who may be of only a limited demographic. It is difficult determine how representative the blogosphere is of the population of Web users because there is no clear demographic information about who is blogging. Inherently the blogosphere is distributed in nature, and a large number of communities are represented (Gouge, 2004). The wide variety of topics and interests shared by the blogging community to a certain extent compensates for the lack of representativeness of the blogging community.

Findings

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%

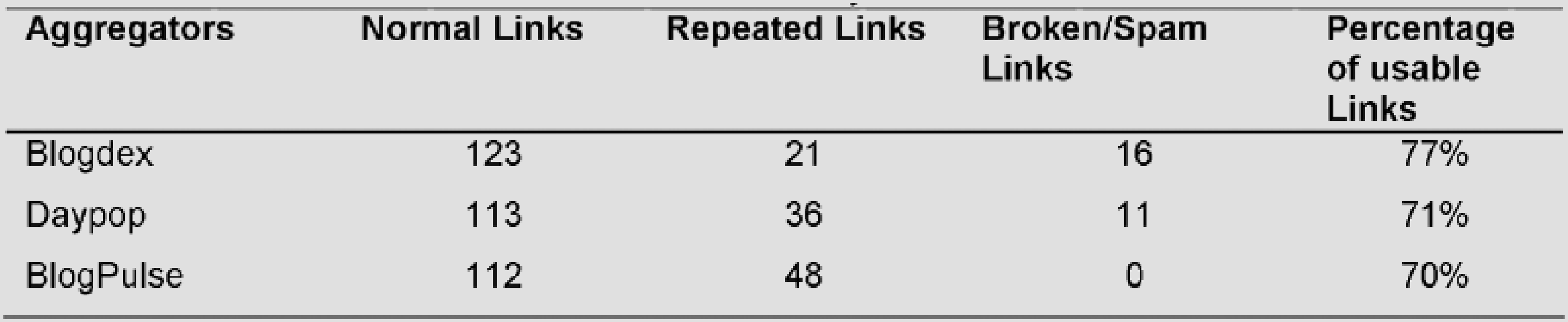
Table 1 represents the archival criteria and the percentage of the content collected by the aggregators for the eight days of the study that fulfilled these criteria. As expected, all of the aggregators were successful at selecting content that documented issues of current social or political interest, and those topics which had an application to an audience wider than the blogging community. Some examples of the content selected from all the aggregators was a story from the British Medical Journal (bmj.bmjjournals.com) about President Bush proposing a mandatory mental health assessment for everyone in the United States, 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), and a story about Gmail (gmail.google.com). Each of these stories had a presence in each of the aggregators.

Daypop had the highest percentage of exemplary content (criterion D) at 9% which was significantly higher than BlogPulse, at only 3%. This was surprising considering the small amount of content collected by Daypop that was considered to be meaningful for building the collection (criterion E). Of course this could be a reflection of the fact that the Daypop database is hand-picked, and may therefore contain less meaningful information, but what is selected is of higher quality. Some examples of what was found to be exemplary were programs or scripts that were written to be openly available on the Web sites that were linked to. Two that were picked up were http://www.marklyon.org, a program to import email from other clients to Gmail, and torrez.us, a program to alert users when new mail had arrived to their Gmail inbox without logging in each time. These pieces of content are examples of the artistry and utility 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 exemplary 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 news and current-event content, even though Daypop spiders RSS feeds and other news sources. BlogPulse had the lowest percentage of currency but this was affected by many of the links pointing to several sites that did not have information about when the content was created. 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 both current and meaningful beyond the blogging community. Over 70% of links collected by each aggregator were to valid content. The combined amount of repeated, broken and spam content links was between 37-48% across the aggregators. Of the remaining links to valid content, 50% were selected to be part of a meaningful and organic collection. This demonstrates that the blogging community is able to select items of archival quality for a representative collection.

The repetition of links was tracked not only to prevent content from being assessed twice but also to ascertain reasons why links were duplicated, if any duplicates existed. Links often repeated both across aggregators and within the same aggregator over several days. Surprisingly, of the 480 links analyzed only 59 links (12%) were repeated in more than one aggregator and of those only 22 (4.6%) were repeated in all three. This provides evidence there is great diversity in the content being selected in the blogosphere and therefore across aggregators.

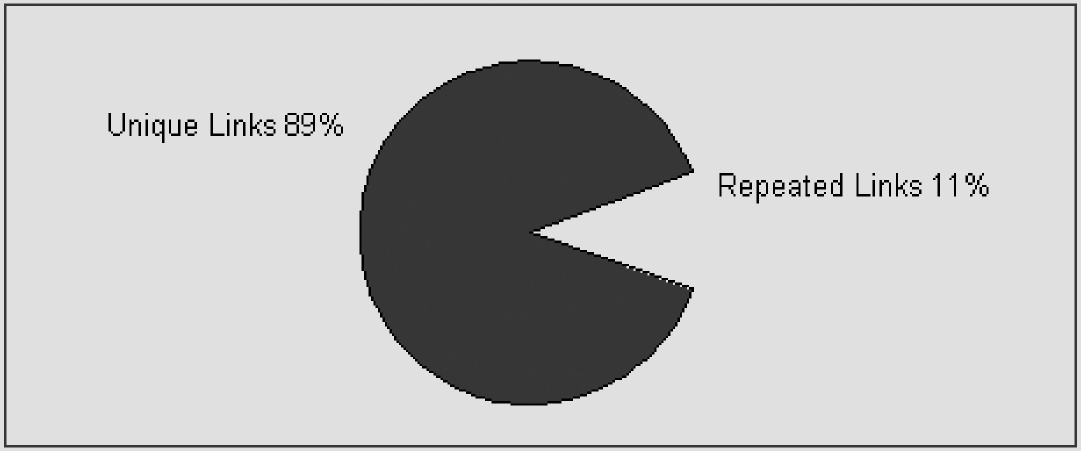


Figure 1: Number of repeated links across aggregators

Repeated links have an interesting effect in this study. While they do not provide new content for analysis, they do allow for the analysis of burst activity. The blogosphere is noted as a "network of small but active microcommunities" which "exhibit[s] striking temporal characteristics" (Kumar, Raghavan, Novak, & Tomkins, 2003). Kumar et al. state that "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 and colleagues' research was based on an 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 authority 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" (p. 606). There is currently no way of systematically measuring or properly assessing a page's authority (p. 606). 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 on these grounds 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 is linked to on a blog has been through a process of selection by human judgment. Once content is posted to a blog, the distributed network of bloggers will select the most relevant and credible information. The greater the number of links to that content, the more likely it will be selected by aggregators, and thus high quality and useful content will be collected over a period of time. Since access to the algorithms being used by each aggregator is not available, it can not be determined which, if any, are using an algorithm similar to the one developed by Kleinberg. A rough estimate may be made from counting the number of links repeated within each aggregator and across the three aggregators.

Discussion

During the time of the study, Blogdex suffered a number of broken links from the shutdown of Weblogs.com and subsequent server change by Dave Winer. Winer, who had been offering 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 Weblogs.com 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, content relating to the shutdown occupied the most of the top rankings. There were also a number of links that were about the shut down which appeared in the list. This story can be seen as an example of burstiness because one story or event came to occupy the center of the ranking for a short time period. 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. (It should be noted that the links about the Weblong.com shutdown would not have been chosen according to certain criteria because of the meta nature of the story.) In an overall assessment Blogdex seems to be the best across all criteria, though the removal of spam links would increase its performance.

Daypop also seemed to suffer from spam and link exploitation. In a few instances links were to business advertising services which held very little information and looked more like ads than information. On another occasion there were a number of stories from Yahoo news which had expired causing broken links. Since all of these links were cited from the same five blogs, they were investigated. None of the blogs had dates in the archives corresponding to the date that had appeared in the aggregator. These links were recognized as problem or false links and were discarded; this task could easily be automated.

Both Daypop and Blogdex have been around longer than BlogPulse so it is probably inevitable that they would have more problems arising from spam. Daypop also had difficulty updating its site as frequently as expected. On Daypop's site information about the aggregator states that the weblogs are crawled and updated every twelve hours and the news sites every three hours. However, only two pages were actually crawled by Daypop after 6 p.m. despite all of the pages being collected after 9 p.m.; none of the pages registered a time stamp of 9 p.m. or later. Daypop did offer feedback on its site which indicated that most of the links used for the analysis came from blog sources. This was a concern in the beginning because Daypop does include new sources and RSS feeds within its database. This might also account for the differentiation in the time stamp.

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 Criterion B, pertaining to current events. Often BlogPulse collected content of popular social interest which could not be easily dated to determine currency. BlogPulse also had the highest percentage of repeated links, 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 the 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 idea that blogging communities are spreading information outside their boundaries and that the blogosphere is becoming a community representative of the web as a whole.

Further Study and Conclusions

This study looked at three blog aggregators available on the Web to determine if the most linked-to content within their databases could be considered for preservation according to the archival criteria chosen. Each of the aggregators collected a large number of links to content that was of archival quality, particularly in the currency of the content and its appeal to a wider audience. An improvement to the aggregators would be to address the technical problems such as broken links or links to spam sights. A simple solution for broken links is to add a link checking feature to the blogging application. On the aggregator application side, improvement to algorithms will need to be made to prevent spam sights from entering the ranking. This is factor is similar to the problems that search engines faced in their beginning versions. Now that Google has purchased Blogger the possibility of a highly sophisticated aggregator could be on the horizon.

The content selected did not frequently overlap in each aggregator, suggesting that the blogosphere is a diverse group of individuals who are selecting content from the Web en masse. Further study of the links or of the database used by the aggregators would give some indication as to whether the same blogs are registered to each of the aggregators and how much overlap occurs between aggregators. The need to collect a greater amount of content which is unique or exemplary should be addressed as well. Some of the individuals who create blogs are experts in their community and have a higher level of credibility or influence in their own community, "weighting" the content these experts link to could improve the collection of content which is exemplary 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 representing expert knowledge or greater influence could help create a richer collection of content while still keeping the size of the collection manageable for preservation.

Another interesting feature to see would be adding tools to the blogging application which would capture blogrolls within the users' blogs, analysis could be done on the development of the social network within the community. Metadata could be added to track versioning as well. Currently the number of blogging applications is relatively small and the blogosphere relatively new, making this an opportune time to add functionality to applications which would be concerned with the preservation of the content.

Currently, there is not a strong system of selection in place when looking at the Web as a whole. The proposed solution of using the blogging community to perform this selection is a radical departure from the centralized method of selection that normally takes place within the archival community. Yet, it may be the best solution when applied to a medium like the Web and digital objects. Blogs are native to the Web not only because they utilize hyperlinking as Tim Berners-Lee envisioned the Web but also because the individuals that choose to communicate through blogs have accepted the Web as a part of daily life. The community of blogs is growing from only those who are technical experts to representing 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. Who better to participate in this democracy than individuals who are personally invested in the development, growth, and analysis of digital information.

With the Internet Archive growing at a pace of 12 terabytes a month, will it be 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 rigorous selection 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 a reflection of 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. Both the ephemeral and significant are captured through a democratic process. An archive of the blogging community therefore serves society as a whole, and would represent the institutional knowledge of the Web.

References

Almaas, E., Barabas, A-L., Dezso, Z., Lukacs, A.., Racz, B., &Szakadat,I. (2005, May, 12). *Fifteen Minutes of Fame: The Dynamics of Information Access on the Web*. Retrieved May 31, 2005, from http://xxx.arxiv.org/abs/physics/0505087

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/

Ceglowski, M. &Marlow, C. Upflux. *An open index for weblogs*. Retrieved May 30, 2004, from http://www.blogpulse.com/papers/upfluxwww2004. pdf

Conway, P. (1996). *Preservation in the digital world*(No. Pub 62). Washington, DC: Council on Library and Information Resources.

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

Glance, N., Hurst, M., and Takashi, T. (2004, May 17-22). *Blogpulse: automated trend discovery for weblogs*. Paper presented at the WWW2004, New York, NY.

Gouge, M. (2004) *Blogs as a Means of Preservation Selection for the World Wide Web*. Retrieved January 15, 2005, from http://etd.ils.unc.edu/dspace/handle/1901/108

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

Perrone, J. (2004). *What is a weblog?*Retrieved July 5, 2004, from http://www.guardian.co.uk/weblogarticle/0,6799,394059, 00.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

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.

Smith, A. (2001). *Stategies for building digitized collections*(No. 101). Washington: Council on Library and Information Resources.

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

Quick, W. (2001). *DailyPundit*. Retrieved July 12, 2004, from http://www.iw3p.com/DailyPundit/2001_12_30_dailypundit_archive.php#8315120

Terdiman, D. (2004, May 12). *Read This, Jump Into Blog Fray*. Retrieved July 2, 2004, from http://www.wired.com/news/politics/0,1283,63331,00.ht ml?tw=wn_story_related

Tibbo, H. (2001). Archival persepectives on the emerging digital library. *Communications of the ACM, 44*(5), 69-70.

Winer, D. (2002, May 17). *The history of weblogs*. Retrieved July 17, 2004, from http://newhome.weblogs.com/historyOfWeblogs