ICHIM
PARIS 21-23 SEPT. 05
www.ichim.org

Digital Culture & Heritage
Patrimoine & Culture Numérique

Bibliothèque nationale de France, PARIS
Sept. 21st - 23rd, 2005
21 - 23 septembre 2005

WEBLOG ARCHIVES: ACHIEVING THE RECORDNESS OF WEB ARCHIVING

Paul WU Horng Jyh and Theng Yin Leng
Nanyang Technological University, Singapore
http://www.ntu.edu.sg/sci/dis

Published with the sponsorship of the French Ministry of Culture and Communication

Actes publiés avec le soutien de la Mission de la Recherche et de la Technologie du Ministère de la Culture et de la Communication, France

Interprétation simultanée du colloque et traduction des actes réalisées avec le soutien de l'Agence Intergouvernementale de la Francophonie

Abstract (EN)

Several national libraries have extended their digital collection to Web publication. Notable Web archiving projects include the Pandora project of the National Library of Australia and the Minerva project of the Library of Congress. However, close examination of their collection policies reveals that little is required of archiving Weblogs. This is despite the fact that volume of Weblog publication is fast growing at a rate estimated to be 12,000 new Weblogs a day. According to the Pew Internet and American Life project, 27% of the entire internet users are also Weblog users. This trend suggests that the existing collection policies need to be re-examined. More significantly, contrary to obstacles faced by Website and E-mail archives, we discover Weblog to be an e medium that is conducive to record keeping principles. Specifically, Weblog allows its ***context***, ***structure***, and ***content*** to be easily definable. This implies that it is more straightforward to appraise the Weblog and establish its authenticity. We shall elaborate on it in the following: first, Weblog is anchored and moderated by one, or a small group, of actors. The actors' profiles are normally detailed within the Weblog itself. Weblog allows commentators to post their comments; this re-enforces a gate-keeping mechanism where the moderator and commentators can easily engage in a dialogic discourse, if necessary, that clarify the information. Weblog also automatically stores its current, as well as historical, postings in chronological order; as such each posting is "encaputalated" in the context when it is created. Second, unlike Emails and newsgroups, the structure of the posting is easily traceable, as it is implemented solely via hyperlink mechanism. The chronological storage feature also allows the evolution of the hyperlinks to be recorded over time, which is almost impossible for Website archives. As a result, there exists a wealth of research in the ACM community on "Weblogging Ecosystem: Aggregation, Analysis and Dynamics." It demonstrates how automatic analysis of the hyperlink structure done by computers can reveal, among other things, the clustering of micro-communities. Last, but not the least, from content's point of view, Weblog documents the contemporary thoughts and actions of the persons in a culture. In a study by Nardi and collaborators at Stanford, it listed reasons why people blogs, including to document one's personal thoughts, life and through the process, to engage others in a community forum. It is also worth noting that research of a discourse as serious as presidential election can be facilitated by analyzing Weblogs; as it is reported by BlogPulse senior researcher Natalie Glance in her publication: The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. To sum

up, the objective of this paper is to demonstrate the record-ness of Weblog as described above, using examples from our project on Asian Tsunami Web Sphere (http://tsunami.archive.org). At the end, we argue that Weblog can be established as a *primary*, *organic*, as well as *trust-worthy*, source of information, given that due archival inspection is applied, which in turn, is facilitated by its record properties.

Keywords: Weblog Archives, Recordness, Tsunami Web Archives

I. The Unprecedented Growth of Weblog Content

Gradually but surely, our cultural heritage is taking digital form. Traditional forms of culture artifact are now produced on the World Wide Web; in many ways the emergence of Web can be modeled after the emergence of the print technology. They are both a publishing technology that presents a society with an unprecedented way of culture expression that empowers all levels of the society, from the culture elite to the general public, as well as from the institutional to the individual.

We have seen waves of transition from off-line to online publication. In the early days, scientists aided by NSFNet started to share scholarly publications via emails and early internet information systems, such as WAIS and Gopher. This is followed by newsgroups and Bulletin Board System (BBS) in the USENet. With the birth of World Wide Web in the mid 90's, online publishing is no longer a tool exclusive to the academics, commercial and public sectors have started major eGovernemnt and eCommerce projects to publish their content online. As mass media organization such as newspapers adopt online publication, online communities are also becoming more popular, with examples such as GeoCity, Tripod and WELL.com. Since the first instance of Web Log by David Winer in 1996 (David Winer, 2000), we are now witnessing the most recent wave of the transition by the phenomenal adoption of Weblog as a form of online publication for anyone who is online and is willing to express their views. In fact, the volume of Weblog publication is fast growing. According to the Pew Internet and American Life project (Lenhart, etc., 2004), 27% of the entire internet users are also Weblog users. This development impacts theinformation professionals who are concerned about keeping records and preserving the memory of the contemporary culture. Thus, it also implies that the existing collection policies for any library and archives organizations need to be re-examined with respect to the growth of Weblog content, much of which is born digital and online at the same time. Specific challenges include: What will be the consequences of these changes in the nature of the medium for creating and preserving our cultural heritage? How would we be able to apply intellectual control over the vast amount of digital artifacts? For this purpose, we shall take a look on the issues related to the content and structure of weblog.

II. The Content, Structure, Context and Discourse of Weblog

Weblogs (also known as blogs) are normally defined as websites, which contain frequent, periodic, reverse chronologically ordered posts on a common webpage. Technorati, a weblog search-engine and measurement firm has reported that 12,000 new weblogs will be created everyday.

1. The Content of the Weblog

According to (Wikipedia, 2005), weblogs can be classified into the following 8 types: (1) Personal, (2) Topical, (3) Issues, (3) Collaborative, (4) Educational, (5) Directory, (6) Business, (7) Advice, and (8) Format type. Each of these types can be subdivided into subtypes, for instance, Topical can be divided into Health, Literary, and Travel, and Issues can be further divided into News, Political, Legal, Media, and Religious. Similarly, under the Business type, there are subtypes such as Entrepreneurial, Corporate, and Small Business. As can be demonstrated, this classification is similar to a YellowPages, which attempts to categorize according to the informational value of the weblog content. From a records management's point of view, it is even more crucial to ask the context and structure of weblogs are captured in the metadata associated with these records.

2. The Context and Structure of Weblog and Self-Documentation

Important aspects of context are the author characteristics, such as their gender and geographical location, and other demographic information. Other contextual attributes are temporal measurements, such as when and how frequent messages are being posted on weblogs. These contextual aspects can be derived from the packaging of a weblog easily.

In terms of the structure of weblogs, attributes such as (automatic) archives, badges, images, comments allowed, link to email author, ads, search function, calendar and guest book. Bloging softwares adopted are important structural features as they will impose boundary as how the structure of the weblogs will be (such as whether trackback is allowed or not). The linking practice and the features of the message itself, such as whether Header contains date, title, time, author's name, as well as Footer, whether time, author's name, internal links, comments, date are included or not.

In the message body itself, whether images and links are used, and if links are used, what kind of links they are: to websites by others, to news sites, to other weblogs, to internal to weblog, to websites created by or about self.

As demonstrated above, most of the contextual and structural information are contained in the format of weblog postings or websites themselves. That is, the context and structure metadata is explicitly encoded. Thus, it can be argued that weblog possesses the quality of self-documenting.

3. The Discourse of Weblog and the Record Continuum

Given the content, structure and context of a weblog posting cam be identified, it will be also important to understand why these records are produced - the why and the process of the record creating process. According to (Herring, etc., 2004; Vuorinen, K, 2005), weblog embodies many traditional forms of publication and communication such as a news filter, a column, a diary, to even a community by itself. The following table summarizes the discourses where weblogs are being used; the corresponding discourses in the physical world are also highlighted.

			CMC/ Asynchronous	**Physical/ Synchronous and Asynchronous**	**Records Continuum/ Identity**
Symmetric	**Exchange**		Blogsphere	Town Meeting	Institution Meeting
Asymmetric	**Exchange**		Community Blog	Call In Talk Show	Organization
	Broadcast	**Topical**	Filter Blog	Column	Unit
		Personal	Diary Blog	Diary	Actor

Table. 1: The discourses where weblogs are applied

As shown in the above table, the spectrum of weblog discourse includes all granularities of identities in a society. Following the records continuum thinking, the identities range from by an individual (a diary), a unit (a column), an organization (community) to the entire society/community (Upward, F., 1996; McKemmish, S, 1997) . From this perspective, weblogs are rich in the identity contexts they exist and the functions they perform in the society. As a result of this diversity, weblogs can be used to capture the culture facets from all levels of the society and at different levels of details, from the traces of a single individual's thoughts to the rituals and rules in a society.

A continuum-based approach suggests integrated time-space dimensions. Records are 'fixed' in time and space from the moment of their creation, but recordkeeping regimes carry them forward and enable their use for multiple purposes by delivering them to people living in different times and spaces. A weblog system is indeed such a system as the traces of human behaviors and thoughts are recorded, and they start to aggregate (e.g. via blogroll) and interact (e.g., via hyperlinking) with actors in different discourses. The interesting aspect of weblog is that the discourse of blogrolling and hyperlinking is then recorded over time for further inspection. At the end, a blogsphere evolved like a cyber-society with its snapshots of the social activities being recorded along its history.

Thus, weblog is indeed an important medium that should be included in the web archives project. Not just for its content, the discourse/narrative as they document the process and purpose of their very clearly, as their discourse down to the most minute traces are recorded. However, many of the national web archives project has largely ignored the collection of weblogs as their culture heritage.

4. The Discourse of Weblog and Domains of Recordkeeping

Furthermore, we observe that the recordkeeping principle of demarcating clear boundaries of domains is facilitated by weblog as well. Weblogs can records personal thoughts in a personal domain. New weblogs can also be created for concerned parties to engage each other in society issues - the society domain, as evidenced by many loggers in response to the Tsunami disaster (See the section on Tsunami Web Archives). The boundaries of personal and society domains are clearly defined, along structures such as hyperlinks and blogroll. The situation in a blogsphere can be quite simlar to what is advocated by the UK Public Record Office (1999, 24), which states "personal work space contains ‘early work in draft, team space contains ‘early formal drafts and discussion docuemnts’ and corporate space contains ‘finished docuemtns and formal records’". As the blogger community begins to emerge, further structure may be impose on the bloging discourse, to regulate content and hyperlinking practice among the blogger. As a result, the blogsphere will become more or more similar to a physical organization with clear responsibility and accountability for the records being created.

III. Re-Examining the Selection Policy for National Web Archives

Despite the fact that Weblogs, as a electronic medium, possess good archival quality, the quality of the Weblog content has often been called into quesiton. This is because the phenomenal development of the Weblog has been largely chaotic, like its preceding phases of transition from offline to online publishing. While weblogs contain much that would definitely be regarded as continuing value (e.g., the publication of scholarly and scientific research, the Web sites of government agencies, etc.), there is much content that is of low-quality (or worse). As a result, although many national libraries have implemented web archives proejcts, these projects may not treat weblog archives as one of their priorities.

National web archives project typically re-enforce the themes that web archives is to select only websites from influential and established organizations and individuals in the society. The content reflects the status quote in the national culture, history, politics, economics and social conditions – information that shapes the dominant view of national identity. National web archives projects tend to formulate their guiding principles for the selection of websites along the following measures of quality in their content, quality such as:

- Relevance
- Authoritativeness and accuracy
- Accessibility
- Being a primary source
- Of national significance

In terms of intrinsic value, some websites may be deemed innovative and unique eg. award winning sites for their application of new media technologies, for which usually only snapshots should be taken.

With the above background, it is not surprising to find weblogs are typically included in the "Exclusion List" in the national web archives selection policy, along with other similar records:

- Blogs
- CAMS (websites employing a web camera that uploads digital images for broadcast)
- Discussion lists, chat rooms, bulletin boards, news groups

However, it is not difficult to find counter example among the diverse democratic and creative processes exhibit in the blogsphere. For instance, in the article "Divided they blog," Adamic and Glance have shown clearly how weblog is an intrinsic social institution in the deliberation of important national issues; it mirrors any other physical and traditional social institution such as

the publication of books (Adamic, L. and Glance, N, 2005). The same phenomenon can be observed in the general public's re-action to Tsunami disasters, upon which we will examine in details later.

1. Re-considering and Adapting Web Archives Selection Policies

Two approaches to resolve the above dilemma have been considered. One is to place the selection process at the hands of the creators of the records and documents themselves; the other, at those of an automatic process; namely, these are the Deposit and Whole Domain web archiving policies, respectively.

Deposit Appraoch

Deposit approach is adopted by the Demark Naitonal Project. Denmark's history of legal deposit extends all the way back to 1697 when the first regulation was passed by royal order. The idea then was to facilitate exchange of printed work with royalty from other countries. This was revised extensively in 1902 at the advent of the Danish Industrial Revolution when the amount of printed matters increased tremendously. In 1997, the legal deposit extended requirements to cover published works 'regardless of medium' with the objective of preserving Denmark's' cultural heritage in published works. Based on this the Royal Library has been selectively collecting web resources since 1998.

The intent is to collect comprehensively rather than selectively and the current strategy for accomplishing this is via legal deposit and snapshot archiving. Currently this leaves large gaps in the collection since the deposit laws are difficult to enforce and there is much work to be done on awareness of the law.

Whole Domain Model

The Internet Archive (IA) is a non-profit commercial venture based in San Francisco since 1996. IA is comprehensively harvesting and preserving publicly available materials from the global Web, and has amassed the largest collection of public Web pages in the world since it began its web archiving activities. Most of the IA's collections are received from Alexa Internet, a commercial crawler and research company which donates collected data to IA immediately after

each crawl period. Broad snapshot crawls are made every two months, and the content of the archive was made publicly accessible in 2001 via the 'Wayback Machine'.

This whole domain approach is based on periodic harvesting, in which snapshots of Web sites of all levels of quality and subject scope are collected and stored for future reference. As many sites as possible are collected, but not all of these are complete. Additionally, some narrow crawls are made of selected sites to collect their entire content.

Central to the IA's approach to archive the whole web is their appraisal not just of the archival value of the web, but also of the possibility and complete feasibility of the effort. And whilst they seem to be keenly aware of the resource challenges facing such an objective (they admit they hardly receive donations from their invitations to fund their web archiving efforts on the internet), they are actively (and optimistically) seeking collaborative methods to overcome their weak areas for example access, funding, etc.

Summary

In "Information Politics on the Web," Rogers deliberated on the indexing issues for search engine (Rogers, 2004). Selection of web archives materials is quite similar to indexing. Rogers conclude that the current indexing model of Google as being in-volunteering and exclusive, as an opaque ranking algorithm is used to display search results. Google's indexing is exactly the same as a Whole-Domain selection approach in web archives, as it sends out crawler to collect the entire web indiscriminatively. If the Whole Domain model is considered as involuntary and exclusive, it seems that the Deposit model is the one to be adopted. However, the Deposit model may not work in the web archives domain as the roles of the publishers have been dis-intermediated, as the authors can publish their work directly online. In sum, the solution may lie in the very way we appraise web archives materials. We hope the following section on how weblogs have played an important role in documenting human history will prove the importance of Weblogs cannot be ignored anymore and national archives project need to rethink how to deploy a selection policy that is fitting with the development of the society. This being said, at the very minimal level, one may still need to consider preserving content due to the diversity of the society that exists (Bearman, 1994; Bowker, G. C, 2001).

IV. Tsunami Web Archives Project

The Tsunami that hit Southern Asia on December 26, 2004 was the worse natural disaster of current generation. More than a quarter of a million people, both locals and visitors from all over the world, lost their lives. Internet enables Web publishers to respond to a natural disaster such as this in a global scale. New Websites emerged, and some existing one took on a different role in a matter of a few hours. They used the Internet to provide information, features, news, services, reactions, as well as virtual memorials, to fill the large information void that existed right after the tragedy struck. Just a week into the disaster, Singapore Internet Research Center (SIRC) collaborated with WebArchivist.org and Internet Archives to begin collecting relevant Web Archives. Web materials were collected from as many sources as possible, in multiple languages, and from sites produced in over 40 different countries. Sites produced by nine types of actors were identified initially, such as NGOs, religious groups and individual citizen sites. Systematic searches for relevant URLs produced by these actors were conducted, and documented. In addition to this, links from the original URLs identified were followed to identify other URLs with relevant content. To ensure that we had comprehensive coverage of the relevant Websites, we identified sites in 13 different languages including English, Chinese (both traditional and modern), Hindi, Indonesian, Malay, Tamil and Thai. Our initial efforts identified over 1,599 different "sites" over the initial four week period. By capturing sites in their hyperlinked context, the archiving tools preserved not just the collection of Web pages, but an interlinked Tsunami Web sphere. The final Web Archives can be found at the following URL: tsunami.archive.org.
Aside from documenting the Tsunami Web sphere, we shall also report on the research findings in the following multi-disciplinary areas: (1) How citizens and informal web publishers, as opposed to official ones, make use of the internet in response to the Tsunami? Particularly, how weblogs have contributed to the communication and expression of the disaster. (2) Are there any patterns of communication that emerge that were not present in the previous studies? How does the information “travel” among the different publishers.

1. Types of Websites

Different types of formal and informal tsunami sites exist in the Internet. Formal websites such as tsunami news websites serve to provide information about the events while websites requesting

for aid serve a different purposes. Blogs and Forums are examples of informal websites and could be a means for people to share their views and feelings about the tragic and emotional event that has caught the attention of the world. Hence, analysis of the web environment of the Asian tsunami community requires the study to include both scenarios.

Many of the weblog sites are actively participating in the various activities of Tsunami relief efforts. As shown in Table 2, out of the 1599 websites in the Tsunami Web Archives, around 10% (or 175 sites) are weblogs or forums that are operated by citizens informally. This is compared with the 30%, highest ranking, of websites (or 495 sites) which is are portals or news websites. On the other hand, weblogs are as actively involved in Tsunami activities as other types of actors such as, corporates, NGO's and Government, which are having around 13% (or around 210 sites).

Types of Websites	# of Websites	% (N = 1599)
Religious Groups	100	6.25%
Corporate websites	211	13.20%
NGOs	210	13.13%
Government	218	13.63%
Schools	75	4.69%
Blogs Sites and Forums	175	10.94%
Portal and Press	495	30.96%
Others	115	7.19%

Table. 2: The distribution of different types of websites in Tsunami Web Archives

2. Linking Practice of Weblogs

Another important dimension as how websites function together is the kind of linking practice that is exhibited. We have selected the three sites from each 6 types of websites and study in more details how they use hyperlinks to engage other types of websites in the websphere. Following the typology of Knoke and Kuklinski (1982), the distribution of various types of linking practice in summarized in Table 3. As shown in Table 3, despite the concentration of communication linking practice, we have found that Blogs and Forums, which are quite similar among themselves (Vuorinen, K., 2005), share similar communication behavior as other, more formal types of websites, as they are involved in all four types of communicative actions, including transaction (such as donations and survey), communication (such as contact,

photograph), instrumental (such as Tsunami resources and information about tsunami), and Kinship (such as friend's links).

Genre	Transaction (%)	Communication (%)	Instrumental (%)	Kinship (%)
Religious Groups	4 (33.33)	4 (33.33)	1 (8.33)	3 (25.00)
Corporate	60 (33.33)	60 (33.33)	60 (33.33)	0 (0.00)
NGOs	21 (42.00)	25 (50.00)	2 (4.00)	2 (4.00)
Govern-ment	5 (5.26)	41 (43.16)	29 (30.53)	20 (21.05)
Schools	2 (4.26)	24 (51.06)	18 (38.30)	3 (6.38)
Blogs and Forums	21 (10.71)	149 (76.02)	21 (10.71)	5 (2.55)

Table. 3: The distribution of different types of websites in Tsunami Web Archives

3. Communication Patterns of Weblogs

A subset of the weblogs in the Tsunami web archives is selected to analyze their communication frequency in details. The selection of weblogs was based on the criteria of popularity: the weblogs were ranked according to the number of comments posted instead of citations, including those that were deleted. The popularity of the websites is measured by combing search results form Blogpulse and Google. The selection period for the weblogs was from 26 December 2004 to 31 March 2005, for a period of 119 days. A total of some 4800 comments are posted and distributed, as shown in Table 4. This implies that the average numbers of comments for the top 10 weblogs during the Tsunami period is 41 comments. This number is contrasted with the average frequency of one posting every 5 days that is reported in (Herring, etc, 2004). This dramatic increase in weblog comments strongly suggested that the websphere is mobilized in response to the extraordinary needs over the Tsunami disasters.

Ranking	Name of the blogs	# of comments
1	South-East Asia Earthquake and Tsunami	1809
2	Tsunami Help Needed	1060
3	Jeff Ooi	684
4	World Changing	408
5	Cheese and Crackers: Tsunami Videos	238
6	Waxy.org	189
7	C***S***F	188
8	Tsunami Survivor Stories	140
9	Suman Kumar	138
10	Brand New Malaysian	88

Table 4: The Top 10 Tsunami Weblog

Between 26 December, 2004 and March 31, 2005

4. Photo-Journalism of Weblog

Not surprisingly, it is found fairly large amount of multi-media documents, such as video and photo materials, are being posted on Tsunami-related weblogs, as shown in Table 5. Many of the images especially during the Tsunami period are captured in the personal blogs or forums. During the Tsunami period itself, almost all photos and video materials are captured in the blogs and personal websites (93 out of a total 94, or 99%). This makes the weblog sites as the major sources of images and video for the rest of the world, including professional news agencies. It is also noted that these sources are also more popularly cited by the Internet community as shown in the following Table 6. Among those images and video clips, a total of 13 of them, that appear more than once on a websites, the more iconic images of Tsunami, 10 of them (or 77%) are made by amateur journalists who publish their work first on their weblogs or forums. Our data does support the observation that independent journalism is emerging in the internet through weblogs.

Name of Media	No. of Professional images/video clips		No. of Amateur images/video clips	
	During	Aftermath	During	Aftermath
Asian tsunami videos	-	-	6	-
Waves of Destruction	-	13	47	6
Photoduck	-	-	9	39
propictureshot	-	-	5	-
Issuespotter	-	-	1	-
Jeff Hock.com	-	-	12	13
Gettyimages	-	3	-	-
Webshots	-	-	-	41
Ogish.com	-	-	1	-
Blogfikiran	-	-	4	-
Tsunami.maldiveisle	-	-	-	2
candc.mirror.unitedemailsystems.com/	-	-	3	-
Surfervillage	1	-	5	-
Sub-Total	1	16	93	101
Total	17		194	

Table 5. Distribution of images and video clips across Personal websites and Blogs

Identifier_personnel	No. of websites image appears		No. of websites video clips	
	2	3	2	3
Professional	3	-	-	-
Amateur	4	4	2	-
Total	7	4	2	-

Table 6. Images and Video clips that appear in more than one website

V. Conclusion

Weblog is the fast becoming the most popular online publication medium, as it is easy to use and as many Internet users mature to this new form of self-expression. On the other hand, most national web archives project have excluded weblogs as one of the sources of web archives collection. However, as the world responded to a un-expected nature disaster, the Asian Tsunami, it unwittingly provided a realization that weblogs are worthy of being the *organic,* as weblog provides rich representation of discourse and metadata, and *primary sources* of recorded information, as the adoption of many photos and video clips captured first-hand during the Tsunami disaster. With these, it becomes evident that national web archives need to include weblog archives.

It is also realized that weblog has an in-built strength of record-ness quality. Its packaging of web pages in its individual postings with many structure and format data allow internet researchers to analyze their structure with great precision. With the automatic archives function, the self-disposing/archiving function can actually be implemented based on any records retention schedule, as each documents is fixed at a particular time-space. With its Ping and RSS communication features, it has also achieved, though un-intentionally, one of the most challenging issue of record registry. Due to these unique features of weblog, a global records registry for weblogs is being established along with the creation of each new weblogs and with each new weblog messages subsequently posted. As this registry mechanism becomes part of the document creation process, we are ensured of a self-documenting, self-migration and self-migrating ways of publishing (Bearman, 1996).

Currently, a project is underway to study how the self-authentication can be proved for weblog's automatic archiving function. The experiment includes comparison between the archives generated by self-archiving capability against the ones collected intentionally by Tsunami Web archives. Our assumption is that if the historical facts supported by the Tsunami web archives can be similarly supported by the archives in the current live sites, then this consistency of historical facts will prove that Weblog indeed keeps the authenticity on its own records, thus achieving the self-authentication of a virtual archives.

References

Adamic L and Glance N (2005) The Political Blogosphere and the 2004 U.S. Election: Divided They Blog, BlogPulse Newswire. Retrieved Online September 6 2006 at (http://www.blogpulse.com/papers/2005/AdamicGlanceBlogWWW.pdf)

Bearman, D (1994) Keynote speech on: Virtual Electronic Junkyard or Cultural Treasure Trove? Useful Electronic Space or Virtual Junkheap? Library of Congress (www.loc.gov/catdir/semdigdocs/bearman.html)

Bearman, D. (1996) Virtual Archives, ICA Meeting, Beijing.

Bowker, G. C, (2001) Biodiversity Datadiversity, Social Studies of Science, 30(5), 643-684.

Nardi, B. Schiano, D. Gumbrecht, M. and Swartz L. (2001) I'm Blogging This A Closer Look at Why People Blog. Unpublished Manuscript

Herring, S. Scheidt, L. A., Bonus, S and Wright, E. (2004) Bridging the Gap: A Genre Analysis of Weblogs. Hawaii International Conference on Systems Science HICSS-37.

Knoke D. and Kuklinski J. (1982). Network analysis, Beverly Hills, CA: Sage.

Lyman, P and Kahle, B. (1998) Archiving digital cultural artefacts, *D-Lib Magazine,* July/August

Lenhart, A, Horrigan, J, and Fallows, D. (2004) Content Creation Online Research Report, *Pew Internet & American Life Project*. *(*http://www.pewinternet.org/PPF/r/113/report_display.asp)

McKemmish, S. (1997). Yesterday, Today and Tomorrow: A Continuum of Responsibility. *Proceedings of the Records Management Association Australia 14th National Convention*. 15-17 September, Perth. (www.sims.monash.edu.au/research/rcrg/publications/recordscontinuum/smckp2.html)

Public Record Office (1999) Management, Appraisal and Preservation of Electronic Records, Vol 2 [UK] Public Records Office, Also Available at (www.pro.gov.uk/recordsmanagement/eros/guidelines)

Rogers, Richard (2004) Information Politics on the Web, The MIT Press, Cambridge and London

Ross, S (2002) Cyberculture, cultural asset management and ethnohistory, A*rchivi* & C*omputer, XII.1*, 43-60

Upward, F. (1996) Structuring the Records Continuum - Part One: Postcustodial principles and properties. *Archives and Manuscripts*

Vuorinen, K. (2005) Using weblogs for discussion, Masters Thesis, University of Tampere

Wikipedia. (2005). Weblog. (http://en.wikipedia.org/wiki/Weblog)

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