Wu, P. H. J., & Heok, A. K. H. (2006). Is web archives a misnomer – how web archives can become digital archives? In C. Khoo, D. Singh & A.S. Chaudhry (Eds.), *Proceedings of the Asia-Pacific Conference on Library & Information Education & Practice 2006 (A-LIEP 2006), Singapore, 3-6 April 2006* (pp. 298-305). Singapore: School of Communication & Information, Nanyang Technological University.

IS WEB ARCHIVES A MISNOMER – HOW WEB ARCHIVES CAN BECOME DIGITAL ARCHIVES?

PAUL HORNG JYH WU & ADRIAN KAY HENG HEOK

Division of Information Studies
School of Communication & Information,
Nanyang Technological University
31 Nanyang Link, Singapore 637718

Abstract. Digital archives are not meant to be mere collections of digital artifacts organized for reference. It ought to be a record, providing evidence for dynamic digital cultural activities, because these activities form an organic and primary source of cultural information and artifacts. This paper investigates two national web archiving projects, argues for the need for greater attention to be paid to archival principles and shows how contextual elements can be retained in a collection. We will also demonstrate how some of these principles have been applied in the Asian Tsunami Web Archives via a web archival method called Web Sphere Analysis. In conclusion, we present a framework where existing and future web archives projects can incorporate Web Sphere analysis to transform their collection into a genuine Digital Archives and become a richer resource for future research.

Introduction

National initiatives by libraries to harvest and preserve their fast disappearing digital contemporaneous cultural heritage have lead to a collection that mimics an online reference library rather than actual digital archives. Although many countries refer to their efforts as web archives projects, many if not all of them, do not manage their collection according to archival principles. These forms of organization are pivotal in the preservation of the rich contextual information that surrounds a collection (Besser, 2000, Dublin Core, 2006).

The current situation appears symptomatic of the lack of proven strategies to deal with such ephemeral materials, revealing a desperate attempt, almost clutching at straws, in face of the millions of websites that sprout up each day as citizens turn publishers in their own right, creating content round the clock easily with the pushing of buttons. With many libraries saddled with missions to preserve a collection of knowledge for future generations, the challenge to provide access to such a vast wealth of materials, that seem to explode from the Internet revolution, can present itself as insurmountable (Lyman & Kahle, 1998).

To complicate matters, such wide scale publishing also makes it difficult to distinguish materials with value (Bearman, 1994). Without the traditional seal of authority, the familiar elements that comprise printed publications, the dizzyingly messy links of a web document also blurs the boundaries of what would otherwise have simply been a document's content (Pearce-Moses, 2005). But how does one identify the value or establish the boundaries of websites effectively? The answer lies in the archival principles of provenance and macro-appraisal. Provenance[1] defines the boundaries as the extent to which an authority can exercise power over the publication of the contents on the websites. Macro-appraisal[2], through functional analysis, will help define what evidential and informational

[1] Provenance. n. (provenancial, adj.) ~ 1. The origin or source of something. – 2. Information regarding the origins, custody, and ownership of an item or collection.
A Glossary of Archival and Records Terminology. Accessed on 4 Jan 06 at
http://www.archivists.org/glossary/term_details.asp?DefinitionKey=196

[2] Macro Appraisal. n. ~ A theory of appraisal that assesses the value of records based on the role of the record creators, placing priority on why the records were created (function), where they were created (structure), and how they were created, rather than content (informational value).
A Glossary of Archival and Records Terminology. Accessed on 4 Jan 06 at
http://www.archivists.org/glossary/term_details.asp?DefinitionKey=224

values a website may contain. These two principles are the intrinsic elements of archival science that are relevant in resolving the difficulties mentioned above.

The principles of a reference library, where materials are organized according to subjects to facilitate users' search, differs from those of an archives which places emphasis on evidence and the preservation of the context in which the materials exists. A look at two national web archiving projects will be conducted to examine how their collection development criteria can be improved to reveal relationships among the materials collected rather than just producing a list of items. Apart from the mere collection of web publications, what current project owners have to take note is that preserving materials without a context cannot make the collection one that can be called archival.

The contents of the various web objects must relate to each other in an identifiable context and these all arranged in a manner where the course of events is apparent. Care must also be taken to ensure that the collected materials are kept together and in the order that was established by their creator and not just rearranged according to subject matter, themes or chronologically to satisfy the need of any group of user. We shall demonstrate how some of these principles have been applied in the Asian Tsunami Web Archives via a web archival method called Web Sphere Analysis. In conclusion, we propose a framework where existing and future web archives projects can incorporate Web Sphere analysis to transform their collection into a genuine digital archive and become a richer resource for future research.

Two Examples of National Web Archiving Projects

Established national web archiving projects around Asia can be found in Japan, Australia, and New Zealand. Others include all the Scandinavian countries (Nordic Web Archives, 1999), the United States, United Kingdom, Austria, Czech Republic, France, Germany, Lithuania, and the Netherlands. Their varied missions and objectives span selective acquisition and preservation of niche areas of interest to whole domain web harvesting.

Two examples will be presented; one from Australia and another from Denmark, to see how their collections are lacking in contextual information which can reveal relationships among the materials collected. The former is chosen because it was one of the earliest countries to embark on a national web archiving project. The latter was picked because it has extended its legal deposit laws to cover materials published in all media, including web publications. An exemplar with such a long history and another with such a wide scope of influence would hopefully provide a representative sample of the other national efforts which are not as well established as the Australians or have backing of such comprehensive and overriding legislation to support its cause as the Danes.

PANDORA: Preserving and Accessing Networked Documentary Resources of Australia (http://pandora.nla.gov.au/index.html)

One of the pioneers in the area of web archiving, the National Library of Australia (NLA) houses an admirable collection of materials that are considered "of significance and to have long-term research value". Since 1996, this highly selective archive has emphasized quality, functionality, full catalogue of the national bibliography, accessibility and facilitation for analysis (Phillips, 2002).

The selection process is done by NLA staff and each of the PANDORA partners according to defined selection guidelines. The different partners focus on materials that have national, state or regional significance, relates to music and film, Australian military history and its Indigenous peoples.

Such careful selectivity and segmentation reveal concerns over limited resources and uncover an outlook that perhaps it might not be possible to archive everything. It could also be symptomatic of a legacy from the print journal and manuscripts projects that have extended into the web domain where the selection of materials for preservation is crucial in view of scarce resources.

In the case of PANDORA, it becomes almost peculiar to include on-line publications like journals, books and publications, which traditionally were perceived by archivists as not possessing sufficient unique qualities to be deemed important for archiving.

But in the eyes of the librarians, such online publications are unique because of their content and the way in which they are displayed. They can become a reflection or even record of the social, intellectual, cultural and technological context of the times.

Yet such a highly selective approach has made the Australian web archive highly segmented. It takes a resource out of context and often does not include other resources to which it is linked. Much contextual meaning is likely to be lost and this could have critical implications for some research

requirements. This probably explains why in 2005, NLA decided to supplement their efforts by undertaking periodic snap shots of the whole domain with the help of the Internet Archives.[3]

It acknowledges in its selection guidelines that dissemination of information online is inchoate and the way in which researchers will want to access, use and apply the potential of the web is in the process of developing. As much as they are confident of their selection criteria that is based on "sound professional experience and judgement", like the rest of us, they cannot truly know what will be important for future researchers.

Denmark (www.netarchive.dk)

The Danes have a long history of legal deposit which extends all the way back to 1697 when the first regulation was passed by royal order. The rationale then was to facilitate exchange of printed work with royalties from other countries. An extensive revision was made in 1902 just before the Danish Industrial Revolution hit the printing industry which greatly increased the amount of printed matter deposited. In 1997, the legal deposit was modernised and extended its reach to cover published works 'regardless of medium'. This was made in with the aim of preserving Denmark's' cultural heritage (Lundgaard, 2004).

The modernised law covers a selective collection and archiving of Internet material. Based on this the Royal Library has been collecting web materials since 1998. The clarifications on the laws through accompanying governmental instruction also gave rise to the concept of "dynamic - static" websites. The extent of collection is limited to static ones that include monographs and periodicals, but exclude the dynamic ones involving databases and homepages. Currently, only static documents are archived (Henriksen, 2001).

Although its intent is to collect comprehensively rather than selectively through legal deposit laws and snapshot archiving, large gaps are nevertheless found in the collection because knowledge of the legal deposit laws is low and difficult to enforce.

From an analysis of the Danish web archive collection, the project also reveals missing links and information from the lack of sufficient private sources. Two thirds came from public sources. Issues of boundary comes into play as it was found that the closer the collection comes to the individual and his private concerns, the less he or she is represented in the Danish internet archive (Henriksen, 2001).

The above merely proves that the deposit model is not easy to implement and a heavy reliance on it will render the collections patchy. Even though the intent of the law has recognized the value of digital records, collection from the web remains much harder to define due to its dynamic and complex nature.

As can be seen from the above, both being exemplars of most national web-archiving projects, there is a clear absence of the adoption of archival principles and capturing of crucial contextual information. All of them limit their collection to those within their national boundaries and ignore relevant linkages to information or materials outside their domain. Wu & Theng (2005) demonstrated how weblogs could be used to reveal the record-ness of web archives to aid in the understanding of contemporary culture. This is because weblogs not only publish the content of a message; it also records the context (e.g., time, hyperlinks to other weblog) and authors' identities, with some even providing extensive profiles of the creator than just a user id.

Archival Principles: The Importance of Contextual Information

Having discerned the lack of contextual information in the web collections of the two countries, we now elaborate on its importance. Governments in their urgency to capture their disappearing digital cultural heritage have rendered their collection as something that resembles more of a reference library rather than an actual archive. The principles of a reference library, where materials are organized according to subjects to facilitate users' search, as shown in the above two examples, differs from those

[3] "The Library acknowledges the disadvantages of the selective approach, including the subjective nature of the judgements about the value of resources to be included, and the fast changing information and research environment, which make it difficult to know how researchers will want to use information in the future. For this reason, in June and July 2005 the Library contracted the Internet Archive to undertake a whole domain harvest of the Australian web domain. In the absence of legal deposit legislation for online publications, it is likely that public access to the contents of the whole domain harvest will be limited." Para. 1.7 - Online Australian Publications: Selection Guidelines for Archiving and Preservation by the National Library of Australia. Revised August 2005 accessed at http://pandora.nla.gov.au/selectionguidelines.html

of an archive which places emphasis on evidence and the preservation of the context in which the materials exist.

Apart from the mere collection of web publications around, what current project owners have to take note is that preserving materials without a context cannot make the collection one that can be called archival. The contents of the various web objects must relate to each other in an identifiable context and these all arranged in a manner where the course of events is apparent. Care must also be taken to ensure that the collected materials are kept together as a whole and not just rearranged according to subject matter, themes or chronologically to satisfy the need of any group of user (The National Archives, UK, 2005).

Only when the complete collection is taken into account can its potential value be fully realised, one that highlights "the relations and patterns, means and artefacts of cultural production and exchange online". It is at this point that we introduce the concept of web sphere analysis, a technique that helps to highlight the links between the different websites and how these links develop and evolve over time. Theoretically, web sphere analysis is based on the concept of online actions, which actually subscribes to archival principles as it facilitates the organization of web collection as records created along the various coordinated or collaborative actions (Foot, 2005). This very record of the web sphere's growth actually shows up the relationship between the various web sites, lending a sense of provenance to the collaborative efforts of the various actors that is currently absent in all national web archiving projects.

Web Sphere Analysis: Revealing Collaborative Actions

Schneider & Foot (2005) presents the view of the World Wide Web as an evolving set of structures supporting online action. These structures manifest and enable the production, inscriptions and experience of cyberculture with social, political and cultural dimensions. The hyperlinked and multi-level nature of the web makes the identification and demarcation of units of analysis a critical but difficult task. The nature of the web is such that it is usually jointly produced by multiple actors, who create features and content.

They proposed a framework for web studies that enables the analysis of communicative actions and relations between web producers and users developmentally over time (Foot et al., 2003). A web sphere is defined as set of dynamically defined digital resources spanning multiple websites deemed relevant or related to a central event, concept or theme, and often connected by hyperlinks. As such the boundaries of the web sphere will change when new websites are found and added. Also the recursive nature of a web sphere makes it dynamic in the sense that pages could be referenced by other sites already included in the web sphere or pages that reference included sites may be added to the existing web sphere.

This goes to show that in a web sphere analysis, one cannot arbitrarily limit a collection to just those within a country's domain, making the boundary rigid would render the information incomplete and patchy. Relevant information to a central event, concept or theme must be included and the hyperlinks that bind the various websites should not be broken to preserve the integrity of the collection.

How then does one actually manage a collection of web materials? The following model provides a theoretical construct that allows the organisation of materials on a "collection" level rather than one that functions on the "item" level as what is currently being done in a library.

The Arizona Model: Managing a Web Collection

Pierce-Moses & Kaczmarek (2005) propose a model for curating a collection of web documents based on the assumption that archival principles of provenance and original order are useful to curate and provide access to documents in a collection.

They highlighted the similarity between an "archival" and a "web" collection, positing that both collections have common provenance, and both group related documents together. In an archival collection, the groups are call series and subseries while those on the web are called directories and subdirectories.

It distinguishes itself from a bibliographic approach to curating a collection in that it manages the materials as a hierarchy of aggregates. It does not dwell into the organising materials at the "item" level. It respects the two cardinal principles of archival science – in the case of provenance, in that it does not mix documents from one source with another and in the case of original order, it requires all documents to be kept in the order that the creator used to manage the materials. This is based on the

premise that the whole collection, by preserving the associations between items is actually greater than the sum of its parts.

However record series in traditional archival science are used for organization of materials within fonds[4]- where records are kept based on shared function, activity, form and use. Read in light of web sphere analysis, there appears to be a need for records series themselves be studied in relation to similar records series from another fond. Put simply in the context of web materials collection, the collaborative actions or communication between two actors or websites is the key that provides the bond that ties two separate fonds online together.

Recursively, archival principles also shed light on web sphere analysis by showing that linking patterns are not just convenient configurations but actually functions that reveal, and are driven by, a need to collaborate.

Another interesting development in archival theory that might help make a web collection archival, is the emerging idea of records continuum.

Records Continuum

To organize an archives based on "related records series between two actors" is consistent with the model of "records continuum" whereby archives can be constructed bottom up, not necessarily top down. What records continuum offers is the removal of a records life-cycle paradigm, and the recognition that records are always in the process of making.

This theory is itself informed by the Structuration theory of Anthony Giddens (1983) which focuses mainly on processes. It emphasized the need to continually re-evaluate and adjust the patterns for ordering our activities.

In that way, we need to constantly monitor the emergence and fading away of online websites and their relationships, this is only workable with continuous efforts to document the web sphere as it threads out in time and space. Taking periodic snap shots of the emerging web sphere can yield much information on how collaborative actions are established online and trends can be spotted – eg. which types of actors tend to work with another during a crisis, offering almost a predictive element in forecasting collaborative efforts in the future. This act of recording the process of growth is itself institutionalising our practice in creating documents, capturing records, organising memory and pluralizing memory (Upward, 1998), providing a dynamic record for researchers to delve into.

In current web sphere analysis this area of work is still largely missing because country's libraries have a mindset of looking only after their own domain's collection. They have not come to a stage where they see the need for greater collaboration to ensure that archival bond between websites can actually yield much richer contextual information for future researchers.

If we consider an archive being the aggregated record of all archival documents of an organisation, then archives, taken in the plural would logically contain the records of many organisations encompassing spatial spread and temporal transmission from each organisation to another. (Upward, 1996)

In a web environment, maintaining intellectual control or provenance over such a vast array of document becomes complex in face of the dynamic relationships between the contents of one web site to another. The solution lies in the incorporation of the "Series System" which advocates a need to document separately records description and administrative context. (Cunningham, 1998)

The "Series System" acts as a record keeping system in the web environment by facilitating every new addition to the collection by having a customised framework for the collection in a web sphere. This documentation of a separate records description and administration context can only be tailor made after some analysis of the materials in the websphere.

In a continuum-based approach, the integrated time-space dimensions allow for records to be 'fixed' in time and space from the moment of their creation, but record keeping regimes actually carry these information forward in time and enable their use for multiple purposes by delivering them to people living in different times and spaces (Wu & Theng, 2005).

To illustrate this, we would use the Asian Tsunami Web Archives (see Figure 1) as an example under inspection to see how the above archival principles: provenance, records series, record

[4] The entire body of records of an organization, family, or individual that have been created and accumulated as the result of an organic process reflecting the functions of the creator.
A Glossary of Archival and Records Terminology. Accessed on 4 Jan 06 at
http://www.archivists.org/glossary/term_details.asp?DefinitionKey=756

continuum and web sphere analysis can and are needed to create and organize a veritable digital archive.

The above web archive was created one week after the worst natural disaster in recent times struck South and South East Asia. As new websites emerge, some existing ones took on new roles in a matter of hours after the cataclysmic event occurred. The Internet was used to provide information, features, news, services, reactions, as well as virtual memorials to fill the large information void that existed just after the calamity. The Singapore Internet Research Center, together with Web.Archivist.org and the Internet Archives, began collecting materials from over 40 different countries, in 13 different languages 1,599 individual sites were identified in the initial four week period.

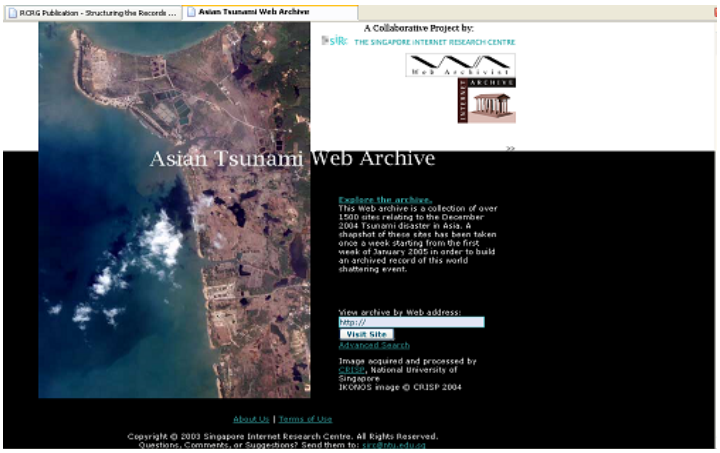

Figure 1. Asian Tsunami Web Archives (http://tsunami.archive.org/)

By capturing web sites regularly over a period of time, the archiving system preserved not just the static collection of Web pages, but an evolving Tsunami Web sphere. This mechanism allows user the ability to address the four dimensions of records continuum: create, capture, organize and pluralize over a span of time-space. Sites were captured, or the sub-sites within, when they were first published. Through a period of 3 months, the pages and hyperlinks of these websites were recorded. Thus, when the archives are analysed, the regulation process which these websites follow to reach equilibrium of collaborative patterns is revealed. From the viewpoints of records continuum, Tsunami Web Archives system serves as a "Locator System" that monitors the self-reflexive changes by the organizations. This can be taken as the foundation to build on where organizations can achieve self-archival responsibilities in a post-custodian society (Upward, 1998) and the self-management properties envisioned in virtual archives (Bearman, 1996).

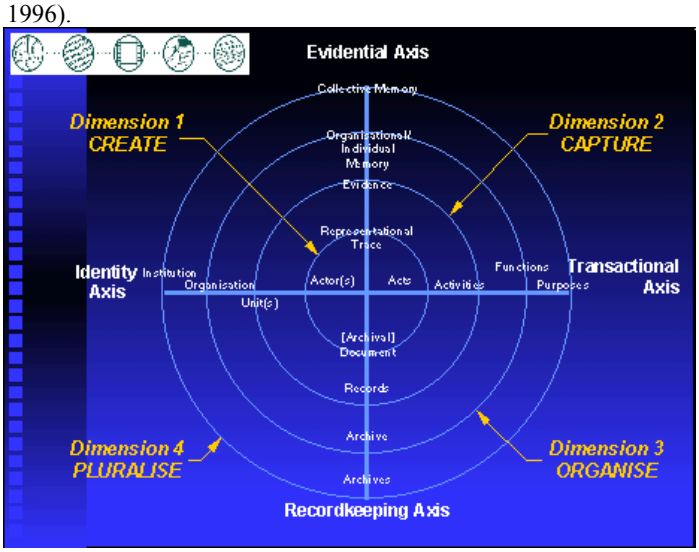

Figure 2. Records Continuum Diagram (Upward, 1998)

The Asian Tsunami Web Archives allows the research community to use the site as a reference resource. This is possible by having a separate records description and collaborative context made for the collection. Coming up with a framework (see Figure 3) to describe the collection (classifying the actors by producer types, countries and language) allow for new knowledge discovery, making patterns and trends more obvious and throwing up clues and ideas for further investigation. For example, looking at the types of actors involved in particular events, it was reported from the findings after analysis of the Asian Tsunami web archive that weblogs by individuals are just as involved in the participation of relief efforts. Others like examining the linking patterns, popularity of particular sites can also be gleaned from closer study. Another interesting finding was that during the Tsunami period itself, almost all (99%) the photos and video materials were captured from the blogs and personal websites. These then became resources to be used for the rest of the world, including professional news agencies, to paint the nightmarish pictures of the tragedy (Wu & Theng, 2005).

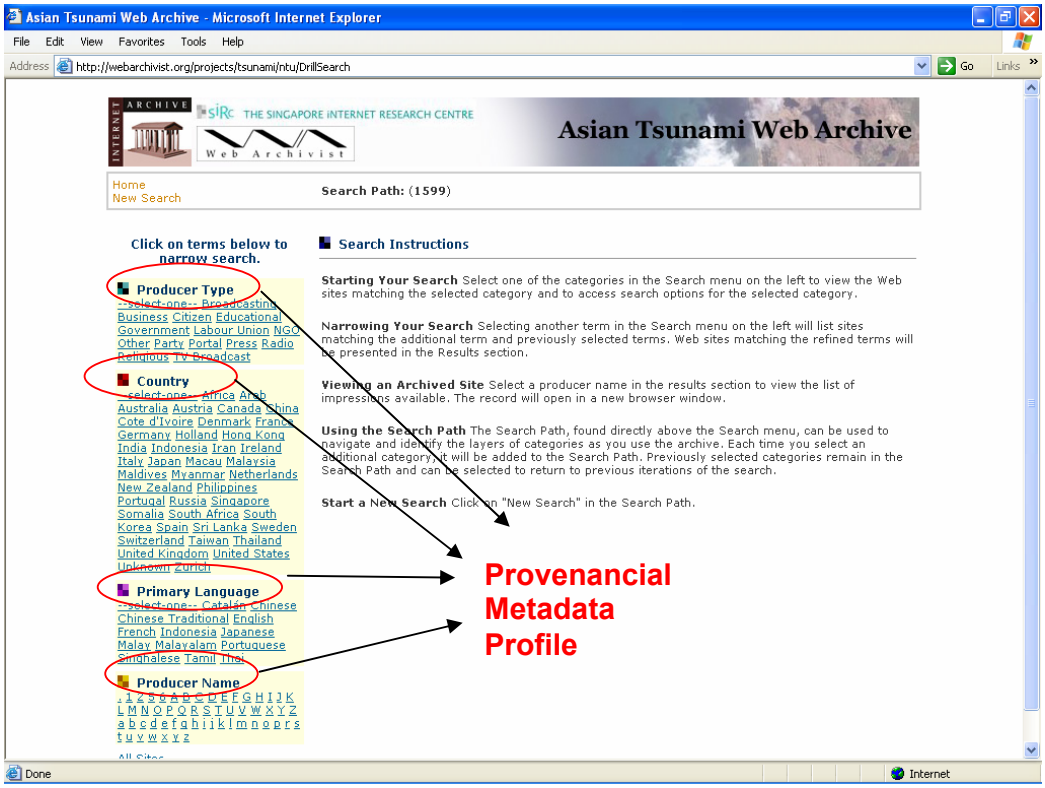

Figure 3. Framework to Describe the Collection

However, within this short span of time, the Tsunami Web Archives has not achieved all that is intended. There are two main outstanding tasks left. First, it should provide visualization tools that allow the user to see the links that are captured in each website, preserving the whole experience contextually when browsing. It also connects the users to all the relevant information the original websites intended to share with its audience then and presenting them with the ability to cross reference all related information on the topic that was captured.

Second, it should ensure that there is a comprehensive collection and preservation of cross-sites hyperlinking relationships. This capturing of the complete contextual information facilitates the various actors to cross verify the provenance of each other. Because collaborative actions offline tend to manifest themselves as announcements of their cooperation on the web or as links to the other organisation, the authenticity of the sites are mutually established. One can easily fake a website but it is hard to ensure the referenced site will do the same and point back to the originator. Being referenced by established sites, especially those of government, large multi national companies or esteemed NGOs

give credence to private or individual pages, validating the information from the smaller sites because of its connection.

Concluding Remarks

As can be seen, organising the materials in a web collection according to archival principles and using a record continuum perspective to structure the aggregates at the series level facilitates the research process so much more. Web archives are in the frontier of digital archives. Despite years of work, we are still in the process of discovering the best way to identify, select, acquire, describe, access and preserve these materials. Interdisciplinary and collaborative efforts (like the Asian Tsunami Web Archives) have produced new insights that integrate models in communication and information sciences. The next step is for national libraries to demonstrate the use of insights from such multi disciplinary experiments to manage their collection along with the contextual information that comes with it and proactively collaborate with all stakeholders concerned to ensure the preservation of such knowledge for posterity.

Bibliography

Bearman, D. (1994). Virtual Electronic Junkyard or Cultural Treasure Trove? Accessed on 3 Jan 06 at http://www.loc.gov/catdir/semdigdocs/bearman.html

Bearman, D. (1996) Virtual Archives. Paper presented at the International Congress of Archives meeting in Beijing, China, September 1996. Available from URL: http://www.ifla.org/documents/libraries/net/bearman.txt

Besser, H. (2000). "Digital Longevity". *Handbook for Digital Projects: A Management Tool for Preservation and Access*. Andover, Mass: Northeast Document Conservation Centre.

Christensen-Dalsgaard, B. (2004). *Web Archive Activities in Denmark.* Accessed on 3 Jan 06 at http://www.rlg.org/en/page.php?Page_ID=17661#article0

Cunningham, A. (1998). Australian Strategies for the Intellectual Control of Records and Recordkeeping System. Accessed on 3 Jan 06 at http://www.naa.gov.au/recordkeeping/control/strategies/default.htm

Dublin Core Metadata Initiative. (2006). Accessed on 2 Feb 2006 at http://dublincore.org/

Foot, K. (2005). Web Sphere Analysis and Cybercultural Studies. Accessed on 3 Jan 06 at http://faculty.washington.edu/kfoot/Publications/WSA-CybCultStudies-dist.pdf

Foot, K., Schneider, S., Dougherty, M., Xebos, M. and Larson, E. (2003). Analyzing Linking Practices: Candidate Sites in the 2002 U.S. Electoral Web Sphere. *Journal of Computer-Mediated Communication*. 8 (4). Accessed on 3 Jan 06 at http://www.ascusc.org/jcmc/vol8/issue4/foot.html

Giddens, A. (1983). Comments on the Theory of Structuration. *Journal for the Theory of Social Behaviour*. 13(1), 75-80.

Henriksen, B. (2001). Legal Deposit from the Internet in Denmark: Experiences with the Law from 1997 and the Need for Adjustments. Accessed on 3 Feb 06 at http://www.deflink.dk/upload/doc_filer/doc_alle/1023_BNH.doc

Lundgaard, E. (2004). Legal Deposit in Denmark. Accessed on 5 Feb 06 at http://www.statsbiblioteket.dk/engelsk/legal/legal.htm

Lyman, P. & Kahle, B. (1998). Archiving Digital Cultural Artifacts: Organizing an Agenda for Action. *D-Lib Magazine*. Accessed on 3 Jan 2006 at http://www.dlib.org/dlib/july98/07lyman.html

Pearce-Moses, R & Kaczmarek, J. (2005). An Arizona Model for Preservation and Access of Web Documents. *DttP: Documents to the People*. 33:1. p.17-24.

Phillips, M. (2002). Collecting Australian Online Publication. Accessed on 3 Jan 2006 at http://pandora.nla.gov.au/bsc49.doc

Schneider, S. & Foot, K. (2005). Websphere Analysis: An Approach to studying Online Action. *Virtual Methods: Issues in Social Science Research on the Internet*. Christine Hine (ed). Berg; Oxford.

The National Archives, UK. A2A: Basic Archival Principles for New Cataloguing Projects. Accessed on 3 Jan 2006 at http://www.nationalarchives.gov.uk/partnerprojects/a2a/pdf/basic_archival_principles.pdf

The Nordic Web Archives. (1999). Introduction. Accessed on 3 Jan 2006 at http://www.lib.helsinki.fi/tietolinja/0100/nwa.pdf

Upward, F. (1996). Postcustodial Principles and Properties. *Archives and Manuscripts*. 24(2)

Upward, F. (1998). Structuration Theory and Recordkeeping. Accessed on 5 Jan 06 at http://www.sims.monash.edu.au/research/rcrg/publications/recordscontinuum/fupp2.html

Wu, P. & Theng, Y.L. (2005), Weblog Archives: Achieving the Recordness of Web Archiving. Proceedings in the Ninth International Cultural Heritage Informatics Meeting September 21 – 23, ICHIM 05, Paris.