Corralling the Wild Web 2.0: An Annotated Bibliography Dealing with the Archival Organizing of Personal Records Born in the Online Environment

Marta Murvosh

Emporia State University

LI-804 - Spring 2010

L. Chase

Abstract

When it comes to preserving digital records in archives or other cultural institutions, many problems exist. Organizing digital records captured from Web 2.0, generally generated by laypeople who are not historians or archivists, remains problematic. These web-born records could be journals or diaries in long and short forms on blogs or social media sites. The web-born records might be knowledge repositories kept on wikis or user-group sites for clubs or social groups. The online environments where these web-born records "live" may be controlled by a company, rather than the person who created them. Once captured by an archives, the organization web-born records must be done to assure access to various user groups, including researchers accustomed to archival or library organizational systems or laypeople accustomed to retrieving information through Google-style search engines. This bibliography of 15 resources offers archivists guidance to organize web-born records and collections.

Keywords: Internet Archive, International Research on Permanent Authentic Records in Electronic Systems Project, Approach to Digital Archiving and Preservation Technology (ADAPT), social media, Web 2.0, metadata, *Describing Archives: A Content Standard*, *The Interactive Archivist,* Persistent Indexing Structure for Archives.

Corralling the Wild Web 2.0: An Annotated Bibliography Dealing with the Archival Organizing of Personal Records Born in the Online Environment

Archivists are under pressure on several fronts to move into the digital world, especially in the Web 2.0 online environment where users generate content. Members of the general public have a high expectation that they should be able to access to historic and cultural records via the internet. The individuals or organizations creating born-digital records using Web 2.0 tools expect preservation of documents with enduring value. Finally, a new generation of archivists and researchers believe that archives must address the issues of appraisal, preservation, and organization of digital-born records -- especially those created online using blogs, wikis, and social media applications. Those archivists believe that the preservation of web-born records of enduring value is necessary to help preserve history and culture. With increasing frequency, archivists, and librarians have begun to act to preserve web-born records before problems associated with the storage, format, and organization of digital records have been resolved.

Blog entries or photographs posted on Flickr do not sit tamely, neatly organized in acid free metal-edged Hollinger boxes, accessible via group and series numbers found printed on a finding aid. Web-born records live online to some extent as unfettered and as ephemeral and as in danger as mustangs. Like wild horses, the danger comes from their environment. Web-born records are threatened by the limitations of servers and the vagrancies of the internet environment, including the whims of their creators or the businesses that host the social media web sites they are posted on. Like many personal digital records, web-born records are rarely backed up by their layperson creators. Simply, web-born records could easy be lost or killed through deletion.

The preservation of web-born records begets the need to organize them. However, there are no easy answers to organizing electronic documents that have been captured from the web, often without their accompanying metadata, and stored on a server where they were not created. The Society of American Archivists has no manuals or guides specifically addressing the organization of web-born or born-digital records. It is a challenge that plagues the archival profession. In student Mary Samouelian's Theodore Calvin Pease Award-winning survey of archives using Web 2.0 applications (2009), 71 percent of respondents indicated that the biggest drawback to using web-born records was the "lack of consistency with descriptive standards" (Samouelian, 2009, p 64).

It is important to organize web-born records in such a way that they can be herded into collections that can be accessed by professional or amateur researchers. Determining how to organize these web-born documents is of primary importance for any archivist attempting to create a digital collection or incorporate web-born documents into an existing physical collection. The digital nature of the records of the records themselves will lead to the creation of digital metadata as a tool in organization.

There are efforts under way to develop methods to preserve and organize web-born records. A few include the nonprofit Internet Archive and the Library of Congress's collaboration with the Approach to Digital Archiving and Preservation Technology (ADAPT) in Maryland. Most of these groups are focused on trying to solve the problem of establishing standards for preservation and migration of digital records. Although these groups look at organization and how to improve

retrieval of web-born records, few of their efforts are directly focused on the organization of web-born digital documents, especially records created in a Web 2.0 environment.

This annotated bibliography cannot offer solutions to problems not yet resolved. However, it assumes that archivists will preserve web-born documents in their original format or as near to original as possible, rather than printing out copies and filing them in a traditional archival arrangement. This bibliography does offer a number of resources to guide archivists in making decisions and establishing policies to organize born-digital records, especially for archives that desire to establish collections of documents web-born within an existing repository or an existing collection.

This annotated bibliography of 15 resources includes annotations that describe the rationale for preserving and organizing web-born records, annotations that look at the possibility of using traditional archival arrangement tools in nontraditional ways, and annotations that detail the efforts of various librarians, archivists, and computer scientists who work to find solutions to the problems associated with the preserving and organizing web-born records.

Why We Need To Preserve and Organize Web-Born Records

The preservation of web-born personal records is a relatively new idea in the archival profession. It has only been the past six years that the literature has reflected the rationale for the preserving records created using social media applications and services. The following papers argue for the preservation of born-digital and web-born personal commentaries, diaries, and photographs, including those found in the Web 2.0 environment.

Cunningham, A. (1999). Waiting for the ghost train: Strategies for managing electronic personal records before it is too late. *Archival Issues 24*(1) 55-64. Retrieved from Wilson Web.

Adrian Cunningham, former secretary of the International Council on Archives Committee on Descriptive Standards, wrote one of the first papers to rationalize the preservation of electronic documents before they are deemed historically significant. In the 1990s, he proposed best practices to preserving personal electronic documents. Cunningham's experience with digital-born records is extensive. He has worked for National Library of Australia, Office for Government Information Technology, and National Archives of Australia, where he is now director of Strategic Relations and Personal Records. Waiting For The Ghost Train (1999) revisits many of the issues associated with the preservation of digital records in the 1990s that are applicable today. Cunningham advocates that archivists should intervene in the pre-custody stage of acquiring historically significant digital-born personal records. This practice will lead to improved organization and description once collections enter the control of an archives. Although Cunningham does not focus on web-born records, his suggestions lay the foundation for a change in archival attitudes toward the preservation and organization of digital-born government records. His advocacy of early archival intervention means archivists can encourage creators of web-born records to employ practices that would aid the creation of metadata.

O'Sullivan, C. M. (2005). Diaries, on-line diaries, and the future loss to archives; or, blogs and the blogging bloggers who blog them. *The American Archivist, 68*(1) 53-73. Retrieved

from http://archivists.metapress.com/home/main.mpx

Catherine M. O'Sullivan, currently a reference archivist at the National Anthropological Archives, received the Society of American Archivists 2004 Theodore Calvin Pease Award for her student paper *Diaries, On-line Diaries, and the Future Loss to Archives; or, Blogs and the Blogging Bloggers Who Blog Them*. Her paper was a call to arms, laying out the reasons for archivists to appraise, acquire, and preserve online journals (web logs or blogs). Although O'Sullivan details a strong rationale for archivist to establish best practices for appraising web-born records, O'Sullivan does not recommend policy or best practices. Still, O'Sullivan's paper is one of the first in the literature to examine the preservation of web-born diaries and journals. For best practices recommendations, see Cunningham (1999) above and Prom and Swain (2007) below. For case studies on Web 2.0 acquisitions, see Daines and Nimer (2009) below.

Prom, C.J. & Swain, E.D. (2007). From the College Democrats to the Falling Illini: Identifying, appraising, and capturing student organization websites. *The American Archivist, 70*(1) 344-363. Retrieved from http://archivists.metapress.com/home/main.mpx

Christopher J. Prom and Ellen D. Swain conducted a survey funded by the National Historical Publications and Records Commission from 2003 to 2004 to look at the ways archivists could document college students' lives using online resources. Both authors are archivists and professors at the University of Illinois, Urbana-Champaign. The authors found that the web sites of student organizations (clubs, etc.) are important but not the definitive source of student culture. They determined that preservation of these sites should be accompanied by preservation of other records, such as scrapbooks. Although the authors caution archivists to study their local groups before embarking on a preservation project, they make general recommendations that could be applied in most cases. These recommendations would help each archives to develop an organizational structure for web-born student-created documents once they have been captured by archivists. One recommendation is to use the university's administrative structure for student groups, including lists and description of student groups. These lists and descriptions could be used to organize the web-born collections and to create the metadata for the finding aids. The authors recommend using existing archival arrangements for organization and capturing each site as a record series, making it simpler to describe the web-born records.

Traditional Arrangement and Untraditional Recommendations

Archivists have been guided by principles of provenance and original order for almost 170 years. Archivists currently use these principles to organize collections and creating finding aids to retrieve digital records -- such as computer floppy discs, compact discs with files of audio or visual images, and DVDs of video. These principles and organizational systems can be applied to web-born documents.

Hunter, G.S. (2003). *Developing and Maintaining Practical Archives: A how-to-do-it manual, 2nd ed.* New York, NY: Neal-Schuman.

Gregory S. Hunter -- a professor in the Palmer School of Library and Information Science at Long Island University, a private consultant in information management services, and the first

president of the Academy of Certified Archivists -- lays out an easy-to-read explanation of basic archival principles. In places, this book is repetitive and overly simplistic, and Hunter doesn't specifically address the organizational challenges posed by the capture of web-born records. However, *Chapter 5: Arrangement* (p. 113-129) and *Chapter 6: Description* (p130-154) offer archivists and archival students a foundation to understand basic archival organization, often called arrangement. Like *DACS* and *EAD* (both annotated below), Hunter follows organizational principles that are built on original order and provenance, principles that date back almost 170 years. Unlike *DACS* and *EAD*, Hunter offers the history and theory behind archival principles that guide the organization of collections. Even though digital records are not specifically addressed in *Chapter 5* and *Chapter 6*, they still offer archivists basic guides to organize web-born records into existing collections.

Society of American Archivists, (2007). *Describing archives: A content standard*. Chicago, USA: Society of American Archivists.

Written by the Society of American Archivists members, *Describing Archives: A Content Standard* or *DACS* is the standard to describe and organize archival records. The corporate author is the largest professional group of archivists in North America and SAA sets the profession's standards. Like many cataloging and organizational tools, the *DACS* offers a controlled vocabulary and a system that archivists have used for years. In spite of publication after many institutions, including the Library of Congress, began to capture web-born, the *DACS* lacks specific information on dealing with digital records beyond a basic mention of digitization. There are no entries in its index for blogs, internet, social media, web, web logs, or wikis. The advice it offers for archivists on digital records is scant, a few pages. Still, the book is useful for archivists adding a series of web-born records to a collection that consists primarily of paper documents. *DACS* would give a digital or web archivist the controlled vocabulary used to organize most physical archival collections and to inform the description in their finding aids, thus creating better metadata.

Society of American Archivists & Library of Congress, (2003). *Encoded archival description, version 2002*. Chicago, USA: Society of American Archivists.

Written by the Society of American Archivists and the Library of Congress, the *Encoded Archival Description* or *EAD* was created to encode archival finding aids. The finding aids, which are metadata, were created using the standards established by *DACS* (annotated above). *EAD* is also compatible with the *General International Standard Archival Description*, the international standard for archival description and organization, which is not in this bibliography. Like *DACS*, *EAD* is a thorough resource for the creation of archival metadata, especially the creation of *EAD* headers containing meta-metadata. Unlike *DACS*, *EAD* is designed for archivists who plan to have machine-readable finding aids and online surrogates of finding aids. *EAD* could set standards for metadata of web-born documents captured by archives.

Dryden, J. (Ed.) (2007). *Respect for authority: Authority, control, context control, and archival description.* Binghamton, NY: The Hawthorn Information Press.

Jean Dryden, a professor at the University of Maryland's iSchool, edited *Respect for Authority*

while working on her doctorate at the University of Toronto in Canada. The volume, which was co-published as *Journal of Archival Organization, Volume 5, Numbers 1/2* (2007), includes three articles that describe the archival profession's shift from authority control (such as establishing headings) to context control. The volume also contains four case studies that look at access, finding aids, and the networking of contextual information at archives in four countries including the United States. This work supplements the *DACS*, which along with other standard archival cataloging resources does not include a definition or discussion of authority control. This issue is important to archivists organizing both physical and digital, including online or web-born, collections because authority control is used to provide subject access to collections. Dryden and her contributors hypothesize that contextual control provides additional access points that overlap to some extent with the *Anglo-American Cataloguing Rules*, which many library catalogers are accustomed to. Contextual context, Dryden argues, also helps improve the quality of metadata, leading to ease in information retrieval by patrons.

Light, M. and Hyry, T. (2002) Colophons and annotations: New directions for the finding aid. *The American Archivist*, *65*(2), 216-230. Retrieved from http://archivists.metapress.com/home/main.mpx

Michelle Light and Tom Hyry propose that archivists add colophons and annotations to each finding aid. Recently named director of special collections at University of California, Los Angles, Hyry, is known for known for streamlining processing collections and enhancing methods to process digital-born documents at Yale University's Beinecke Rare Book and Manuscript Library. Light is the archivist and acting head of Special Collections and Archives at the University of California, Irvine. The colophon would add transparency because the archivist would explain the choices made in preparing a collection for the archives, including explaining decisions made for its organization. Researchers or future archivists would add annotations, similar to internet tagging, turning the finding aid into a living document with constantly expanding metadata. For archivists working with records born of Web 2.0 -- which can be evolving documents -- the authors' proposed practice would increase access to web-born collections. Potentially, annotated or tagged finding aids would have another layer of metadata that would be accessed via an internet search, possibly aiding retrieval. The main weakness of this resource is the authors do not recommend software or policies to oversee tagging or annotation of finding aids.

Librarians, Archivists and Computer Scientists Working To Lasso The Web

Most resources related to archives and libraries moving into the Web 2.0 world focus on using social media as a marketing tool or to make archival collections more accessible through the use of online surrogate records. A few resources offer advice on organizing digital documents in a collection. Librarians and archivists must look to mathematicians, electrical engineers, and computer programmers to develop the means to organize web-born records and the information they contain. Already researchers in these fields have begun to supply solutions to the problem of organizing web-born records and increasing the efficiency of retrieval of those records. Ultimately, interdisciplinary relationships will provide the means for archivists and librarians to organize web-born documents through inventing better ways to create metadata and computer analogs that understand metadata more like humans do.

Kelly, B., Ashley, K., Guy, M., Pinsent, E., Davis, R., & Hatcher, J. (2008). Preservation of web Resources: The JISC PoWR project. Retrieved from http://www.ukoln.ac.uk/web-focus/papers/ipres-2008/Kelly_a14_word.doc

Brian Kelly, team leader of the web focus group at UKOLN, a research group funded by the British Museums, Libraries and Archives Council, and five colleagues produced *Preservation of Web Resources: The JISC PoWR Project.* The report summarizes challenges faced by the Preservation of Web Resources Project in England and lists the philosophical, legal, and practical issues that arise when trying to preserve web resources, including Web 2.0 documents. Kelly and his team, a research group based at the University of Bath, detail the various approaches that could be used in preservation. The report's main weakness is it deals with British laws, which are not applicable in the United States. However, the report provides information on challenges, strategies, and approaches that is valuable for archival mangers making decisions about the preservation and organization of web-born records.

Cunningham, A. (2008). Digital curation /digital archiving: A view from the National Archives of Australia. *The American Archivist 71*(2) 530-543.

Adrian Cunningham -- who has written extensively on the preservation and organization of electronic records, see Cunningham (1999) above -- advocates for archival intervention at all points in the record-making and preserving process. The author argues that the nature of digital-born records, including web-born documents, requires archivists to work with the creators of the documents to create descriptive metadata as soon as possible. For Australian government records, Cunningham advocated intervention at time of creation. For web-born records captured from the internet, metadata should be created at the time of capture and again before access is provided to users. Without these steps, archivists would be unable to manage their electronic collections. Cunningham's paper focuses on the success these practices have had with Australian government records, which has led to improvements in control, organization, and the ability to retrieve born-digital government records. The practices advocated by Cunningham could be applied to archival institutions that are acquiring web-born records, such as college student organizations or blogs or web sites relevant to local communities. These practices would aid organization of those records and certainly could become the basis of best practices for organizing web-born records.

Daines, J.G & Nimer, C.L. (2009). *The Interactive Archivist: Case studies in utilizing Web 2.0 to improve the archival experience.* Retrieved from http://lib.byu.edu/sites/interactivearchivist/

J. Gordon Daines, III is the archivist at Brigham Young University's L. Tom Perry Special Collections and known for integrating Web 2.0 in his institutions services. Cory L. Nimer is the manuscripts cataloger and metadata specialist at the same institution, and he has been part of the Web 2.0 initiatives at that repository. The Web site *The Interactive Archivist* was born of discussions about redesigning the delivery of BYU's finding aids and the discovery that BYU archival users wanted to interact with archival content using digital media. For a citation on researchers tagging finding aids see Light and Hyry (2002) above. At the Society of American

Archivists conference in 2008, Daines and Nimer and the SAA editorial committee determined that an online publication was needed to serve the literature needs about the impact of Web 2.0 on archival practices. *The Interactive Archivist* site launched in 2009 and includes case studies, a bibliography, and descriptions of each Web 2.0 technology and how they are used in archives. Although the site has little that specifically addresses organization and cataloging, the discussions of the technologies, their various uses, and the peer-reviewed research that illuminates real-world archival use of Web 2.0 tools is invaluable for problem solving and to aid archival decisions before acquiring and organizing a web-born collection. For a discussion on the various means to organize digital born records, see Whittaker and Thomas (2009) below.

Whittaker, B.M. and Thomas, L.M. (2009). Chapter 6: Access to collections: Catalogs, finding aids, and Web 2.0. *Special collections 2.0: New technologies for rare books, manuscripts, and archival collections*. (pp. 77-97). Santa Barbara, California: Libraries Unlimited.

Beth M. Whittaker is head of special collections cataloging at The Ohio State University Libraries. Lynne M. Tomas is head of rare books and special collections at Northern Illinois University. The authors delve into the pros and cons of various cataloging methods, including MARC, Encoded Archival Description, *DACS*, SGML, PastPerfect-Online, and Dublin Core, as well as how Web 2.0 tools could be used. They concluded that the system used by Open Archives Initiative Protocol for Meta Data Harvesting or OAI-PMH, (which is annotated below) offered the greatest promise for handling different vocabularies and different data types. According to the authors, OAI-PMH also shows the most promise for the many groups and individuals looking for software solutions that would allow for cross-collection searching and information retrieval. The authors approach the issue of how to organize archival collections from the perspective of library patrons and archives users -- a focus that tends to be lacking in much of the literature. They also discuss the likelihood that patrons will eventually all but require that archives to preserve and organize web-born records and information created in social media environments.

Open Archives Initiative (2008) Initiative Open Archives Initiative protocol for metadata harvesting. Retrieved from http://www.openarchives.org/

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is one of several standards and open software products created by the initiative, which is a coalition that develops and promotes standards for digital and web content interoperability to facilitate the transfer, especially dissemination, of information. The initiative is led and funded by Cornell University's Computing and Information Science and the Digital Library Research and Prototyping department at Los Alamos National Laboratory Research Library. Similar to library cataloging systems and metadata systems, such as Dublin Core or the in-development Resource Description Access (an update to the Anglo-American Cataloging Rules 2nd ed.), OAI-PMH would be easy for experienced library catalogers to use. However, it could be difficult to merge this system with an existing archival organizational systems and might be best applied for collections that are entirely digital or web-born.

Song, S. and JaJa, J. (2008) *Archiving temporal web information: Organization of web contents for fast access and compact storage*. Retrieved from

https://wiki.umiacs.umd.edu/adapt/images/8/89/Temporal-web-archiving-final-umiacs-tr-2008-08.pdf.

Graduate assistant Sangchul Song and professor Joseph JaJa, both work for the University of Maryland, College Park as part of Approach to Digital Archiving and Preservation Technology (ADAPT). ADAPT is working with the Library of Congress to improve the preservation and organization of web-born records. JaJa has published extensively on algorithms used for online programs. Their report describes the existing methods used to store and organize captured web pages. The authors also describe tests of their Persistent Indexing Structure for Archives (PISA) software. The main difference between PISA and older software is PISA determines whether a web page has changed. This determination prevents duplications. PISA also can eliminate duplicates from existing web archives. This report describes a significant breakthrough, overcoming one of the weaknesses of the existing software, potentially making web-born documents easier and faster to organize. PISA adds another layer to the metadata -- the date a page was updated. Existing metadata is the URL and date captured. Although this report is moderately technical, it gives archivists an understanding of the software challenges faced when attempting to organize and create metadata for web-born records. For a more extensive overview of the various technologies and their impact on the organization of web-born records, see Cruse and Sandore (2009) below.

Cruse, P. & Sandore, B. (eds.,) (2009). The Library of Congress National Digital Information Infrastructure and Preservation Program. *Library Trends 57*(3).

An entire issue of *Library Trends* was devoted to articles on the Library of Congress' National Digital Information Infrastructure and Preservation Program, which is a partnership with cultural heritage institutions and universities to develop a means to preserve and organize digital information. Although this issue is not solely devoted to web-born records, many articles are. There is literature detailing how metadata may be harvested from the web and the impacts of various programs -- such as Stanford University Libraries' LOCKSS (Lots of Copies Keep Stuff Safe) -- on metadata, which is used for archival organization and retrieval. Another article details the problems posed when crucial preservation metadata is left out and the impact on semantic computer structures. Finally, an article by Joseph JaJa and Sangchul Song, who developed PISA (2008) software described above, offer a report on developments in web archiving, describe advances in the Approach to Digital Archiving and Preservation Technology (ADAPT) and its impact, including on descriptive metadata for captured web-born records. Although some articles are technical, the information is very valuable for people who specialize in archival arrangement and metadata. This resource also provides an extensive overview for archivists to understand the latest developments in capturing, organizing, and retrieving web-born records.

References

Samouelian, M. (Spring/Summer 2009). Embracing Web 2.0: Archives and the newest generation of Web applications. *The American Archivist, 72*(1) 42-71.