3-2012

An Overview of Web Archiving

Authors: Jinfang Niu

This overview is a study of the methods used at a variety of universities, and international government libraries and archives, to select, acquire, describe and access web resources for their archives. Creating a web archive presents many challenges, and library and information schools should ensure that instruction in web archiving methods and skills is made part of their curricula, to help future practitioners meet those challenges. In preparation for developing a web archiving course, the author conducted a comprehensive literature review. The findings are reported in this paper, along with the author's views on some of the methods in use, such as how traditional archive management concepts and theories can be applied to the organization and description of archived web resources.

D-Lib Magazine

March/April 2012
Volume 18, Number 3/4

An Overview of Web Archiving

Jinfang Niu
University of South Florida
jinfang@usf.edu

doi:10.1045/march2012-niu1

Abstract

This overview is a study of the methods used at a variety of universities, and international government libraries and archives, to select, acquire, describe and access web resources for their archives. Creating a web archive presents many challenges, and library and information schools should ensure that instruction in web archiving methods and skills is made part of their curricula, to help future practitioners meet those challenges. In preparation for developing a web archiving course, the author conducted a comprehensive literature review. The findings are reported in this paper, along with the author's views on some of the methods in use, such as how traditional archive management concepts and theories can be applied to the organization and description of archived web resources.

Keywords: web archive, web archive methods, web resources

Introduction

Web archiving is the process of gathering up data that has been recorded on the World Wide Web, storing it, ensuring the data is preserved in an archive, and making the collected data available for future research. The Internet Archive and several national libraries initiated web archiving practices in 1996. The International Web Archiving Workshop (IWAW), begun in 2001, has provided a platform to share experiences and exchange ideas. The later founding of the International Internet Preservation Consortium (IIPC), in 2003, has greatly facilitated international collaboration in developing standards and open source tools for the creation of web archives. These developments, and the growing portion of human culture created and recorded on the web, combine to make it inevitable that more and more

libraries and archives will have to face the challenges of web archiving.

Library and information schools need to prepare students for these challenges. A survey of the course catalogs of the top 32 U.S. Library and Information schools in the fall of 2010 found that only one school, the University of Michigan, offered a half semester course on web archiving. The University of Indiana covered web archiving as a topic in its 'Content Analysis for the Web' course. UCLA covered web archiving as a topic in its 'Management of Digital Records' course (US News and World Report Weekly, 2009). Although many schools, for example the University of Illinois, offer courses on digital preservation, digital curation, and the impact of web 2.0 on archival theory and practice, it is not known to what extent any of them address web archiving issues. The author believes that web archiving requires enough unique knowledge and skills to require a separate course. In preparation for developing a web archiving course, the author conducted a comprehensive literature review, and evaluated the functionalities of a number of prominent web archives. This paper, and a second paper also published in *D-Lib Magazine*, "Functionalities of Web Archives", resulted from the course preparation research.

Like the management of many other kinds of information resources, the workflow of web archiving includes appraisal and selection, acquisition, organization and storage, description and access. This workflow is the core of web archiving. The following sections review methods used in each step of the workflow and present the author's viewpoints on some of these issues. Although digital preservation is definitely a major step in the overall process of web archiving, it is not unique to web archiving. The preservation of web resources is no different from the preservation of other digital resources. It can be covered in a digital library course, or an electronic records management course as well. Therefore, this review does not cover digital preservation.

Appraisal and Selection

The term appraisal is used in the archives community to refer to the process of evaluating the value of records, and deciding whether and how long records should be preserved. It is essentially a process of selection. In this paper, it is used synonymously with selection. All web archives select web resources to preserve based on one or more criteria. Although the Internet Archive tries to archive the whole web, it only grabs web pages on the surface of the web (Lecher, 2006). Web pages further down the hierarchy of websites are often not harvested by the Internet Archive.

Existing web archiving efforts use the following selection criteria to determine what to preserve: domain (such as .gov or .edu), topic or event, media type and genre. Many European countries archive the web in their country domain. The library of the NASA Goddard Space Flight Center (GSFC) captures pages in the Goddard domain (Senserini *et al.*, 2004). The Library of Congress has created various event-based web collections, such as the September 11, 2001 web archives, the election web archives and the Iraq War 2003 web archives (Library of Congress Archives, 2011). Media-type based selection includes or excludes certain media types. The Goddard library, for example, avoids crawling large video

files and software products (Senserini *et al.*, 2004). The web archiving project conducted by Chirag Shah and Gary Marchionini (2007), on the other hand, focused on preserving election videos on Youtube. Some web archives select based on genres such as blogs, newspapers, virtual worlds, etc. The National Library of France created a web collection of e-diaries (Lasfargues *et al.*, 2008). The Internet Archive has a software archive and an archive of videogame videos (Internet Archive, 2001a; Internet Archive, 2001b). The Preserving Virtual Worlds project conducts research specifically on archiving online virtual worlds (Preserving Virtual Worlds, 2008). Antonescu, *et al.* (2009) pointed out two different approaches to preserving online virtual worlds. One approach preserves the technical infrastructure — the objects and avatars existing in the virtual worlds — while the other approach preserves the interaction and life experiences of avatars in virtual worlds. Winget and Murray conducted research to preserve the records and artifacts created during the process of developing videogames (Winget and Murray, 2008).

Theoretically, selection based on objective criteria can be easily automated. On a technical level, it is easy for software to decide the media type (audio, video or textual) and domain (e.g., .gov, or .au) of web resources. Similarly, it should not be very difficult to differentiate between such genres as online journals or blogs, or notice the differences between blog posts and comments. High quality or popular web content may be identified based on the number of incoming links and visitors, number of viewers for online videos, and user ratings. The National Library of the Czech Republic automated the identification of the Czech Web outside the national domain, which includes Czech websites that are not in the .cz domain, but in .net, .com, .org, or .edu domains (Vlcek, 2008). A WebAnalyzer was created and integrated with the crawler. During crawling, the WebAnalyzer analyzes web pages and looks for some pre-defined properties that characterize the Czech web. Every time a pre-defined property is found, a certain amount of points are added to the URL. When a certain threshold is reached, the web page is considered a part of the Czech web and will be archived.

Selection based on topic or event, however, needs human judgment. Manual selection by information professionals is time-consuming and expensive, and therefore only used in small-scale web archives. To reduce the cost of manual selection, some web archives accept user recommended URLs, use existing registries of URLs or involve subject specialists in helping select web resources to archive. The Preserving and Accessing Networked Documentary Resources of Australia (PANDORA) and the National Taiwan University Library web archives accept user recommended websites (National Library of Australia, 2008; Chen *et al.*, 2008). The Digital Archive for Chinese studies (DACHS) invited scholars who are experts in Chinese studies to recommend related websites (Lecher, 2006). The UK government web archiving project selects websites using a registry of all UK central government websites; URLs in the registry are submitted and maintained by website managers (Spencer *et al.*, 2009).

Another way to speed up manual selection is to use the macro-appraisal theory in the archives management field. As explained in the Arizona Model for curating government web publications, macro appraisal entails appraising and selecting web resources based on

aggregates of web pages rather than individual web pages (Pearce-Moses and Kaczmarek, 2005). Appraising aggregates reduces the size of the problem and makes the appraisal process more efficient. The aggregates can be decided on different levels. The U.S. National Archives and Records Administration (NARA) used several units of analysis in its guidance for government agencies to conduct risk analysis for web records: group of websites, a whole website, a website minus one or two portions that exhibit substantially different characteristics, and clusters of web pages (NARA, 2005). These several units of analysis can also be applied in web archiving selection. For example, librarians or archivists can evaluate the value of a whole website instead of individual web pages to decide whether the website should be archived.

Selection criteria, such as domain or media type, can be associated with either a value-based selection or a representative sampling method. The web archive of the National Taiwan University gathers web resources that are valuable from historical, cultural, social, educational, or academic viewpoints (Chen *et al.*, 2008). Spam filtering is also a type of value-based selection method. Representative sampling, on the other hand, avoids the subjectivity and bias in value-based appraisal and tries to create a representative image of what is to be preserved. Lyle (2004) applied sampling strategy to web resources that have been downloaded by crawlers as a way to reduce the quantity of web resources to be archived. The National Library of France used the sampling strategy to decide the seed list and filtering criteria before crawling; the National Library believes that collections should "mirror the French society and culture in all its diversity regardless of the scientific value or popularity of the publications" (Lasfargues *et al.*, 2008). Due to this belief, "the web archive includes the 'best' (literature, scientific publishing) as well as the 'worst' (from advertisings to pornography). Small, medium and big got the same chance to be collected" (Lasfargues *et al.*, 2008).

Acquisition

Depending on the scale of the web archive, the relationship between the web archive and the website owners, and the nature of the archived web content, different acquisition methods may be used to obtain web content. Libraries and archives have a long tradition of accepting transfers from government agencies, donors and legal deposits from publishers. This method still applies to web archiving. For example, NARA asked every federal department to hand over a snapshot of their website to NARA by the end of President Clinton's term of office (Bellardo, 2001). Adrian Brown (2006) pointed out that database-driven dynamic websites are not suitable for direct transfer, because databases are often proprietary and hard to preserve long-term. An easier approach is to convert data from proprietary databases to an open standard format like XML, using a tool like DeepArc.

An acquisition method unique to web archiving is crawling. This method relies on crawlers to harvest content from web servers. Crawlers use a seed list to start downloading web content, and follow the hyperlinks to discover and download additional web content. Selection decisions are the basis for compiling a seed list and configuring crawler parameters. For example, the National Library of France decided to crawl every website in

the .fr and the .re Top Level Domains and any other domain that has been redirected from a .fr or a .re domain (Lasfargues *et al.*, 2008). This selection decision is configured into the crawlers as a filter. Only links that pass the filter will be archived. Crawling is replacing depositing in acquiring web publications in some libraries and archives. For example, the Arizona State Library has switched to crawling from waiting for submissions from state government agencies (Pearce-Moses and Kaczmarek, 2005). Some web resources need to be manually captured due to crawler limitations. For example, some crawlers cannot harvest GIS files, dynamic web content, or streaming media. NARA provides a guide for appropriate capture methods of specific web content record formats that cannot be captured by crawlers (NARA, 2004).

Repeated crawling of non-updated pages causes duplicates in the web archive, which will waste resources for management, storage and preservation. Fortunately, today's smart crawlers, such as the current version of Heritrix, have the functionality to reduce duplicates in downloading and storing web resources. Repeated crawling of large, frequently updated websites causes temporal incoherency. It may take several days or even longer to crawl a large website, during which time the web sites are undergoing changes. Suppose there are two web pages (p1 and p2) on a website. The crawler downloaded p1 at time t1. When the crawler reaches p2, p2 and p1 have been updated to p2-a and p1-a respectively. In this scenario, the original website includes p1 and p2, the updated website contains p1-a and p2-a, but the archived website includes p1 and p2-a. In other words, the crawler has harvested a website that never really existed. Research is being conducted to reduce temporal incoherency in web archives (Spaniol *et al.*, 2008).

In acquiring web resources, the decision of whether to seek permission from copyright owners depends on the legal environment of the web archive, the scale of the web archive, and the nature of archived content and the archiving organization. In a country where legal deposit covers web resources, such as New Zealand, the legal deposit library does not need to seek permission for archiving web publications produced in that country. Government archives that have the legal mandate to preserve public records, such as NARA and the UK National Archives, also do not need to seek permission from record producers. In the same legal environment, permission seeking is more likely to be conducted for small-scale than large-scale web archiving, because it is more manageable to seek permission from a relatively small number of copyright owners. Large-scale web archives, such as the Internet Archive tend to use the opt-out mechanism (obey robot exclusion and allow takedown upon requests). Hal Varian (2006) argued that the opt-out mechanism of Google's Library Project is a sensible choice, because the transaction cost of the opt-in model, in which permission is sought, is too high to be successful. This argument is also valid for web archiving, perhaps even more so as most web archives do not benefit financially from archiving web content, and it is difficult to identify the copyright owners for web content posted by anonymous users.

The scale of the web archiving effort also affects the decision to obey robot exclusion. According to a 2006 copyright law of France, the National Library of France can ignore robot exclusion while crawling the French web (Lasfargues *et al.*, 2008). In practice, the National

Library of France usually does not obey robot exclusion when performing small scale focused crawls, but it does obey robot exclusion in broad crawls because it is easier to manage the consequences (such as protests from website owners and associated crawler traps) of ignoring robot exclusion in small-scale web archiving than in large-scale web archiving (Lasfargues *et al.*, 2008). The nature of archived content also affects the decision to seek permission or not. The Library of Congress seeks permission to archive blogs and websites of news organizations, but only notifies most other types of websites that the library is archiving their websites (Grotke and Jones, 2010).

Organization and Storage

Web archives need to preserve the authenticity and integrity of archived web content. The requirements for authenticity and integrity vary with the purpose of the collection. In some scenarios, preserving only intellectual content is enough. In other scenarios, such as in preserving legal evidence, the structure and context of resources may also need to be preserved. In traditional archive management theory, the context of archival records includes provenance and the original order. Provenance includes information about the source of records, such as the record producers, the transactions that cause the records to be produced, and chain of custody. Original order is the order in which record producers or record managers originally arranged the records to demonstrate the relationships among records. Although many web archives preserve web content as information resources rather than as evidence, the concept of provenance still applies. For archived web resources, provenance includes the URL of a website, the content producers, and the business transaction or the purpose that caused the web resources to be produced. The URL is external metadata associated with the web resource. Other information about provenance is often embedded in the content of the web resource.

For web resources, the concept of original order can be combined with the concept of structure defined in traditional archives management theory. The original order is essentially the external structure of the archived web object. The structure defined in traditional archives management theory is essentially the internal structure of the archived web object. For example, for an archived website, its external structure shows how this website is arranged in relation to other websites, which can also be regarded as the original order of the website. Incoming links that come from outside of the website and outgoing links from this website to other websites are part of this external structure and thus the original order of this website. The internal hierarchical structure of the website shows how the components and sub-components of this website are arranged in relation to each other, which can be regarded as the structure defined in traditional archives management theory. This internal structure is defined by the hyperlinks within the website. For a lower level archived object, such as a webpage, the external structure shows how this web page is arranged in relation to other web pages. Outgoing links from this webpage, and incoming links from outside of this webpage define the external structure and the original order of this web page. The internal structure shows how the internal components of this web page, for example, how the textual content, images, audio, videos. etc., are arranged. In

repeated harvesting, historical context that shows how the web content evolved also exists. It includes the older and newer versions of the web pages.

According to Masanes (2006), current web archives mainly use three approaches to organize and store archived web content: local file systems, web-based archives and non-web-based archives. All three approaches preserve the intellectual content of web pages, but vary in the degree of preservation of context and structure.

In a web archive that uses a local file system, the browser can navigate the file system just like navigating the web (Masanes, 2006). Both the internal hierarchical structure of websites and the link relationships among different websites are preserved, except for those non-archived links that fall outside of the scope of the web archive. However, two contextual transformations have to be made to let web resources fit in file systems. First, the naming of URIs needs to be modified to conform to rules for local file systems. Second, absolute links have to be transformed to relative links to allow navigation within the file system as otherwise absolute links will point to live web pages instead of archived content.

In a web-based archive, web pages and associated metadata are grouped and stored in container files and the original URIs and links are preserved. Although the links also need to be redirected or transformed to point to the archive instead of the live web, the link redirection or transformation only happens when users access those links as opposed to having to be written into the archive. This second approach preserves authenticity to the largest degree.

The non-web-based archive approach extracts web documents from hypertext context and re-organizes them into a catalog-based access mode or transforms them into PDF files. This approach preserves authenticity and integrity to the least degree.

Description and Metadata

The metadata generation approach and the richness of metadata generated vary depending on the scale of the web archive and the resources available at the archiving organization. Very large web archives often rely on automatic metadata generation. Some metadata information, such as the timestamp generated when the web resource was harvested, the status code (e.g., 404 for not found or 303 for redirection), the size in bytes, the URI, or the MIME type (e.g., text/html), can be created or captured by crawlers. Metadata information can also be extracted from the meta tags of HTML pages, although some meta tags are not accurate due to Search Engine Optimization. The Greek Web Archive project automatically extracts keywords from web pages and anchor text, and then uses the keywords to classify web pages into clusters (Lampos *et al.*, 2004).

Small-scale web archives can afford to create metadata manually. The online campaign literature archive of University of California at Los Angeles uses the Dublin Core metadata standard, Library of Congress subject headings and locally defined authority lists. Its administrative metadata is derived from the detailed notes created by staff during the

capture and review process (Gray and Martin, 2007). The Digital Archive for Chinese Studies web archives invited scholars to contribute some descriptive metadata (Lecher, 2006). The National Taiwan University Web Archives created a three-level hierarchical classification scheme and cataloguing rules specially for web contents (Chen *et al.*, 2008). Metadata can also be created through user tagging, commenting, or rating. The Library of Congress automatically generates Metadata Object Description Schema (MODS) records based on metadata created by URL nominators, and then enhances the records by catalogers (Grotke and Jones, 2010).

Web archive collections have a multilevel hierarchical structure. A web archive collection may include a number of crawling sessions. In each crawling session, a number of websites are crawled. Each website includes many web pages. Each webpage may be composed of many files such as a text file, an image file and a video file. This hierarchical structure matches the hierarchical structure of an archive collection. The multilevel description methods used for archives can be applied to archived websites. The archives community uses a top-down approach: metadata is created for the higher levels first, then if resources are available, metadata for the lower level will be created; metadata created for higher levels can be inherited by lower levels; metadata is rarely created for item-level objects. This top-down approach and metadata inheritance mechanism can also be applied to web archives. In addition, some metadata for the item level objects, such as file format, size in bytes and date of modification, can be automatically extracted.

In the case where a web archive decides to use a bibliographic approach and create only a single level description, it should choose the unit of description based on the scale of the web archives and the resources available. A unit of description on a higher level such as a whole website means less detailed description and fewer metadata records will be created. The Library of Congress and the Harvard University web archive create one MARC record for a web archive collection that includes many websites. This MARC record is searchable through the library catalog. A unit of description on a lower level such as the page level results in more detailed description and more metadata records will be created. In addition to the MARC records for a web archive collection, the Library of Congress web archive also creates MODS records for websites (Library of Congress Web Archives, 2011). These MODS records are searchable in the web archive but not accessible through the library catalog. PANDORA also chooses a website and a part of a website as the units of description (Hallgrimsson, 2006).

Access and Use

The accessibility of web archives depends on the legal environment of the country in which the archive is hosted. The legal deposit legislation of New Zealand allows the National Library of New Zealand to preserve any publicly available pages of a New Zealand website and to provide access to the archived copy of the website (National Library of New Zealand, 2010). In the US, the Library of Congress makes the bibliographic records for all archived websites publicly accessible and can only provide public access to webpages whose producers have given permission (Grotke and Jones, 2010). Many web archives are dark archives or only

accessible onsite, such as the web archives of National Library of France and the Institut National de l'Audiovisuel (INA) of France, the Finnish web archive, Netarchive.dk, Web Archive Norway, the Webarchive of Slovenia, Web Archive Switzerland and Web Archive Austria (International Internet Preservation Consortium, 2011). Some publicly accessible web archives offer reduced functionality and delayed access to avoid competition with the website owners (Masanes, 2006). For example, there is a delay of at least three months between when a web site is harvested and when it will display in WAX (Harvard University Library, 2009). In the case of IA Wayback Machine, the delay is 6-12 months (Archive-it, 2011).

The search capability of different web archives depends on the richness of metadata and the search and indexing tools used. The Library of Congress web archive and the New Zealand web archive support search through authority controlled access points. This was made possible by the fact that these two web archives used subject headings in their metadata records for archived websites. Web archives based on the Wayback Machine, on the other hand, are only searchable by URL, whereas web archives based on the NutchWax search engine can also support full-text search. Some advanced access interfaces have been created. The UK web archive created two visualization interfaces for their web archives based on mining content, tag clouds and a 3D wall (UK Web Archive, 2011). Jatowt, *et al.* (2008) also experimented with several advanced methods to display the historical versions of web pages; they created a slide show and a two-dimensional graph to show how the content of a URL has evolved overtime.

Conclusion and the Next Step

Existing web archives demonstrate a variety of methods and approaches to selecting, acquiring, organizing, storing, describing and providing access. This variance is caused by external factors, such as the legal environment and the relationships between web resource producers and the web archive, as well as internal factors, such as the nature of archived web content, the nature of the archiving organization, the scale of the web archive, and the technical and financial capacity of the archiving organization.

This overview is based on a comprehensive review of literature that explains how web archiving is being done. However, none of the literature directly addresses the knowledge and skills required by the professionals in the field who perform the daily routine of selecting, acquiring and cataloging web archives. The author is planning a research project to fill this gap, for which librarians and archivists who perform these tasks will be interviewed. The viewpoints of the practitioners will provide valuable additional input for the design of the web archiving course that is being developed from the findings of this literature search, and from an evaluation of web archive functionalities.

Note

The possible uses of web archives depend on the functionalities of web archives. To get an

overview of the functionalities supported by current web archives, the author evaluated ten publicly accessible archives that are members of the International Internet Preservation Consortium (IIPC) that have an English interface. Detailed research methodology and findings are reported in the article, "Functionalities of Web Archives", also published in the March/April 2012 issue of *D-Lib Magazine*.

References

[1] Antonescu, M., Guttenbrunner, M., & Rauber, A. (2009). Documenting a Virtual World – A case study in preserving scenes from SecondLife. Proceedings from IWAW '09: *9th International Web Archiving Workshop*, Corfu (pp. 5-10).

[2] Archive-it. (2011). FAQ. *Archive-it*.

[3] Bellardo, L. J. (2001). *Memorandum to Chief Information Officers: Snapshot of public web sites*.

[4] Brown, A. (2006). *Archiving websites: A practical guide for information management professionals*. London: Facet Publishing.

[5] Chen K. H., Chen, Y. L., & Ting, P. F. (2008). Developing National Taiwan University Web Archiving System. Proceedings from IWWA '08: *8th International Workshop for Web Archiving*, Denmark (pp. 1-8).

[6] Gray, G., & Martin, S. (2007). The UCLA Online Campaign Literature Archive: A case study. Proceedings from IWAW '07: *7th International Web Archiving Workshop, Vancouver* (pp 1-5).

[7] Grotke, A., & Jones, G. (2010). DigiBoard: A tool to streamline complex web archiving activities at the Library of Congress. Proceedings from IWAW '10: *10th International Web Archiving Workshop*, Vienna.

[8] Hallgrimsson, T. 2006. Access and finding aids. In *Web Archiving* (pp.131-152). Berlin: Springer-Verlag.

[9] Harvard University Library. (2009). WAX Public Interface Help. Harvard University.

[10] Internet Archive. (2001). Software Archive. *Internet Archive*.

[11] Internet Archive. (2001). Videogame Video Archive. *Internet Archive*.

[12] International Internet Preservation Consortium. (2011). Member Archives. *International Internet Preservation Consortium*.

[13] International Internet Preservation Consortium Access Working Group. (2006). Use cases for access to Internet Archives. *International Internet Preservation Consortium*.

[14] Jatowt, A., Kawai, Y., & Tanaka, K. (2008). Using page histories for improving

browsing the Web. Proceedings from IWAW '08: *8th International Workshop for Web Archiving*, Denmark.

[15] Lampos, C., Eirinaki, M., Jevtuchova, D., & Vazirgiannis, M. (2004). Archiving the Greek Web.

[16] Lasfargues, F., Oury, C., & Wendland, B. (2008). Legal deposit of the French Web: Harvesting strategies for a national domain. Proceedings from IWAW '08: *8th International Workshop for Web Archiving*, Denmark.

[17] Lecher, Hanno E. (2006). Small scale academic web archiving: DACHS. In *Web Archiving* (pp. 213-226). Berlin: Springer-Verlag.

[18] Library of Congress Web Archives. (2011). Minerva. *Library of Congress*.

[19] Library of Congress Web Archives. (2011). Technical Information. *Library of Congress*.

[20] Lyle, J. (2004). Sampling the Umich.edu Domain. Proceedings from IWAW '04: *4th International Web Archiving Workshop*, Bath, UK.

[21] Masanes, J. (Ed.). (2006). *Web Archiving*. Berlin: Springer-Verlag.

[22] NARA. (2004). Expanding Acceptable Transfer Requirements: Transfer Instructions for Permanent Electronic Records: Web Content Records. *NARA*.

[23] NARA. (2005). NARA Guidance on Managing Web Records. *NARA*.

[24] National Library of Australia. (2008). Services to researchers. *National Library of Australia*.

[25] National Library of New Zealand. (2010). Frequently Asked Questions About Archives Websites. *New Zealand Web Archive*.

[26] Niu, J. (2012). Functionalities of web archives. *D-Lib Magazine*, Vol.18, No. 3/4. http://dx.doi.org/10.1045/march2012-niu2

[27] Pearce-Moses, R., & Kaczmarek, J. (2005). An Arizona model for preservation and access of Web documents.

[28] Preserving Virtual Worlds. (2008). *Preserving Virtual Worlds*.

[29] Senserini, A., Allen, R. B., Hodge, G., Anderson, N., & Smith, D. Jr. (2004). Archiving and accessing web pages: The Goddard library web capture project. *D-Lib Magazine*, 10 (11). http://dx.doi.org/10.1045/november2004-hodge

[30] Shah, C., & Marchionini, G. (2007). Preserving 2008 US Presidential Election Videos. Proceedings from IWAW '07: *7th International Web Archiving Workshop*, Vancouver, Canada.

[31] Spaniol, M., Denev, D., Mazeika, A., & Weikum, G. (2008). Catch me if you can:

temporal coherence of Web archives. Proceedings from IWAW '08: *8th International Workshop for Web Archiving*, Denmark.

[32] Spencer, A., O'Reilly, B., & Vasile, G. (2009). Past and present: Using the UK Government Web Archive to bridge the continuity gap. Proceedings from IWAW '09: *9th International Web Archiving Workshop*, Corfu. (pp. 38-43).

[33] UK Web Archive. (2011). Visualization. *UK Web Archive*.

[34] US News and World Report Weekly. (2009). Best Graduate Schools: Library and Information Science. *US News and World Report Weekly*.

[35] Varian, H. R. (2006). *The Google Library project*.

[36] Vlcek, I. (2008). Identification and archiving of the Czech Web outside the National Domain. Proceedings from IWAW '08: *8th International Workshop for Web Archiving*, Denmark. Retrieved from http://iwaw.europarchive.org/08/IWAW2008-Vlcek.pdf

[37] Winget , M.A. & Murray, C. (2008). *Collecting and preserving videogames and their related materials: A review of current practice, game-related archives and research projects*.

About the Author

Jinfang Niu is an assistant professor at the School of Information, University of South Florida. She received her Ph.D. from the University of Michigan, Ann Arbor. Prior to that, she worked as a librarian at the Tsinghua University Library for three years. Her current research focuses on electronic records and digital curation.