R01: Exploring web archiving concepts and frameworks

"Who is capturing web content as cultural heritage, how is it being managed and what challenges still remain."

Brief overview

The purpose of this research project is to explore the practice and theory of web archiving to identify gaps and opportunities in relation to selection, description and access to web archives.

The goal of the project is to identify the current conceptual and practical challenges facing archivists working with web content and digital archives.

- To identify the instruments and practices that support web archiving, including standards and other tools.
- To identify and explore the conceptual foundations that support web archiving practices.
- To identify opportunities for improved practice, including development of innovative technologies.
- To identify any gaps between practice and theory. Including, identification and analysis of links (if any) between selection, description and access management.
- Identification and clarification of the role of web archiving as cultural heritage.
- Identification and clarification of the role of the institution in cultural heritage formation.
- Exploration of the value of the Mediated Recordkeeping model as an analytical research instrument.

Research problem

Web archiving practices were developed almost 25 years ago with the Australian National Library leading the profession. Since this time web archiving has evolved into a major field of practice with leading work being done in national libraries and in independent organizations, including the Internet Archive (https://archive.org/), and the International Internet Preservation Consortium (IIPC) (http://www.netpreserve.org/). Web archiving raises several major and complex challenges for archives, including how to conceptualize web contents and formats, how to preserve content and make it accessible in a meaningful way, and how to design and deliver policy and processes to identify and support the creation, selection and management of a cultural heritage archive (Niu, 2012; Toyoda & Kitsuregawa, 2012).

The rise of web archiving has helped build and support increased awareness of the technical and practical issues inherent in the preservation, curation and stewardship of digital materials and how web collections contribute to the field of digital archives (Ashley et al., 2010; Cathro, Webb, & Whiting, 2009; Corrado & Moulaison, 2014; Costa, Gomes, Couto, & Silva, 2013; Day, 2006; Kalb, Lazaridou, Pinsent, & Trier, 2013; Kasioumis, Banos, & Kalb, 2014; Lyman, 2002;

Niu, 2012; Toyoda & Kitsuregawa, 2012). There has also been a growing discussion outside of the library and information discipline about what researchers need from web archives (and that web archives are not currently fulfilling) (Axel Bruns, 2011; Dougherty et al., 2010; Gomes & Costa, 2014). Other discussions include the exploration of linking web archives and other digital sources such as digital repositories (that house data sets as well as unstructured data) as part of the emerging fields of linked data and the semantic web (Bojārs, Breslin, Finn, & Decker, 2008; Heath & Bizer, 2011; Risse et al., 2012). Recently, the Internet Archive in conjunction with a consortium of US universities was awarded an IMLS grant to study and develop solutions related to web archiving technical infrastructure to support integration, access, and distributed processes in the support of preservation (Jefferson, 2015).

However, my previous research (Gibbons, 2009 & 2015) has shown there is a need to explore the conceptual assumptions underpinning policy and practice in the selection and description of web archives to fully support participatory, inclusive and contextual archives that support a multiplicity of future needs. Much of the work done has been on improving practice, but little study and reflection has been done on the assumptions underpinning the practice. Collections Policies, the common policy instrument supporting the capture and creation of web archives may not always be interpreted and applied in the way that was intended, or in a narrowly focused way that supports the practice instead of the practice supporting the policy goal. Conversations from web archive scholars over the last 5-10 years also highlight the problems with collecting content over context and the ethical dimensions of web archives (Phillips & Koerbin, 2006; Rauber, Kaiser, & Wachter, 2008; Wu & Heok, 2007). Further conversations in relation to community engagement and participatory archives also form part of a critical conversation about how to create and nurture digital archives that include web content is also relevant to this research project (Eveleigh, 2012; Gilliland & McKemmish, 2014; Labrador & Chilton, 2009).

The design of the research incorporates an understanding of the Mediated Recordkeeping Model (MRK), a conceptual model that identifies and maps conceptual, practical and societal processes and implications involved in creating, capturing, organizing, curating and pluralizing cultural heritage. The purpose of the research project is to map identified practice and theory from research participants onto the model to explore gaps and opportunities for web archiving and ultimately, digital archives. Gaps and opportunities may include conceptual and/or practical including use and design of supporting technologies, models for collaboration, and policy templates.

Selection, description and access management as processes sit within the pluralize dimension as a practice, but as a concept they are influenced by create and capture, as conceived, designed and implemented by archivists and other heritage professionals within institutions. Additionally, the impact of the conception, design and implementation impacts on dimensions of organization and curation within the institution and the ability to develop and manage frameworks that support a complex cultural heritage with multiple and potentially contested points of view.

Methodology

The research is positioned within an interpretivist and continuum informatics world view.

There will be three stages of data collection, the first being an anonymous survey sent to practitioners working in institutions, and the second will be interviews at specific and willing institutions to follow up on findings from the survey. The third stage includes disseminating research findings to participants of the second stage. It is anticipated that participants will contribute to the development of the research outcomes by providing feedback including comments and suggested changes. Analysis of the processes undertaken in the third stage will help to inform rigor, reflexivity and contribute in evolving the researcher's skills in developing and managing research partnerships.

Data will be analyzed using various instruments and processes including:

- Statistical analysis of survey results. (First stage)
- Content and thematic analysis to identify and examine key topics, themes and their related context. (Second stage)
- Use of the MRK model for mapping themes and contexts. (Second stage)
- Auto-ethnography (Third stage)

Anticipated outcomes

The goal of the project is to identify the current conceptual and practical challenges facing archivists working with web content and digital archives. With a focus on exploring decision-making in relation to policy, a likely outcome is the development of guidelines and/or a policy template. It is also anticipated that within the framework of decision-making a model of collaboration that addresses the need to include continuously evolving descriptive and access management contexts will also be developed. Additionally, it is anticipated that technology limitations will be identified as an issue, and so there is a possibility to identify requirements for the design of a system that supports current web archiving goals. Furthermore, there is potential to design a model digital archives and practical requirements to support the model.

Links to further research

This research project is part of a larger research project (R05) that looks to study:

- The applicability of the MRK in research with various and diverse communities.
- The development of an archival system that supports individual needs for memory-making that can also link to community, organizational and institutional systems.

References

Ashley, K., Davis, R., Guy, M., Kelly, B., Pinsent, E., & Farrell, S. (2010). A guide to web preservation: practical advice for web and records managers based on best practices from the JISC-funded PoWR project. (S. Farrell, Ed.). JISC. Retrieved from http://jiscpowr.jiscinvolve.org/wp/files/2010/06/Guide-2010-final.pdf

Axel Bruns. (2011, October). *Archiving the Immediate: How and Why Archives Should Approach Socia...* Technology. Retrieved from http://www.slideshare.net/Snurb/archiving-the-immediate-how-and-why-archives-should-approach-social-media

Bojārs, U., Breslin, J. G., Finn, A., & Decker, S. (2008). Using the Semantic Web for linking and reusing data across Web 2.0 communities. *Web Semantics: Science, Services and Agents on the World Wide Web*, *6*(1), 21–28.

Cathro, W., Webb, C., & Whiting, J. (2009). Archiving the Web: The PANDORA Archive at the National Library Australia. *National Library of Australia Staff Papers*. Retrieved from http://www.nla.gov.au/openpublish/index.php/nlasp/article/viewArticle/1314/1600

Corrado, E. M., & Moulaison, H. L. (2014). *Digital Preservation for Libraries, Archives, and Museums*. Rowman & Littlefield.

Costa, M., Gomes, D., Couto, F., & Silva, M. (2013). A survey of web archive search architectures. In *Proceedings of the 22nd international conference on World Wide Web companion* (pp. 1045–1050). International World Wide Web Conferences Steering Committee. Retrieved from http://dl.acm.org/citation.cfm?id=2488116

Day, M. (2006). The long-term preservation of Web content. *Web Archiving*, 177–199.

Dougherty, M., Meyer, E. T., Madsen, C. M., Heuvel, C. V. D., Thomas, A., & Wyatt, S. (2010). Researcher Engagement with Web Archives: State of the Art. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1714997

Eveleigh, A. (2012). Welcoming the world: an exploration of participatory archives. In *Int. Council on Archives (ICA) Conference (ICA 2012)*. Retrieved from http://www.gosbook.ru/system/files/documents/2012/11/13/ica12Final00128.pdf

Gibbons, L. (2015). Culture in the continuum: YouTube, small stories and memory-making. [PhD thesis]. http://arrow.monash.edu.au/vital/access/manager/Repository/monash:153826

Gibbons, L. (2009). Testing the continuum: user-generated cultural heritage on YouTube. *Archives and Manuscripts*, *37*(2), 89.

Gilliland, A. J., & McKemmish, S. (2014). The role of participatory archives in furthering human rights, reconciliation and recovery. *Atlanti: Review for Modern Archival Theory and Practice, 24*. Retrieved from http://escholarship.org/uc/item/346521tf.pdf

Gomes, D., & Costa, M. (2014). The importance of web archives for humanities. *International Journal of Humanities and Arts Computing*, *8*(1), 106–123.

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, *1*(1), 1–136. http://doi.org/10.2200/S00334ED1V01Y201102WBE001

Jefferson. (2015, October 8). IMLS National Digital Platform Grant Awarded to Advance Web Archiving | Internet Archive Blogs. Retrieved from http://blog.archive.org/2015/10/08/imls-national-digital-platform-grant-awarded-to-advance-web-archiving/

Kalb, H., Lazaridou, P., Pinsent, E., & Trier, M. (2013). *Interoperability of web archives and digital libraries: A Delphi study*. Retrieved from http://www.purl.pt/24107/1/iPres2013_PDF/Interoperability%20of%20web%20archives%20and%20digital%20libraries%20A%20Delphi%20study.pdf

Kasioumis, N., Banos, V., & Kalb, H. (2014). Towards building a blog preservation platform. *World Wide Web*, *17*(4), 799–825.

Labrador, A. M., & Chilton, E. S. (2009). Re-locating meaning in heritage archives: a call for participatory heritage databases. *Computer Applications to Archaeology 2009 Proceedings*. Retrieved from http://works.bepress.com/angela_labrador/5/

Lyman, P. (2002). Archiving the world wide web. *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving*, 38–51.

Niu, J. (2012). An Overview of Web Archiving. *D-Lib Magazine*, *18*(3/4). http://doi.org/10.1045/march2012-niu1

Phillips, M., & Koerbin, P. (2006). PANDORA, Australia's web archive: how much metadata is enough? *Journal of Internet Cataloging*, *7*(2), 19–33.

Rauber, A., Kaiser, M., & Wachter, B. (2008). Ethical Issues in Web Archive Creation and Usage-Towards a Research Agenda. In *8th International Web Archiving Workshop (IWAW08)*.

Risse, T., Dietze, S., Peters, W., Doka, K., Stavrakas, Y., & Senellart, P. (2012). Exploiting the social and semantic web for guided web archiving. In *Theory and Practice of Digital Libraries* (pp. 426–432). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-33290-6_47

Toyoda, M., & Kitsuregawa, M. (2012). The History of Web Archiving. *Proceedings of the IEEE*, *100*(Special Centennial Issue), 1441–1443. http://doi.org/10.1109/JPROC.2012.2189920

Wu, P. H. J., & Heok, A. K. H. (2007). Is web archives a misnomer - how web archives can become digital archives? In C. Khoo, D. Singh, & A. S. Chaundry (Eds.), *Asia-Pacific Conference on Library and Information Education and Practice 2006* (pp. 298–305). School of Communication & Information, Nanyang Technological University.