Submitted on: 10/06/2015

Growing a web archiving program: A case study for evolving an organization-management plan

Todd Suomela
Digital Initiatives, University of Alberta, Edmonton, Canada
todd.suomela@ualberta.ca

Abstract:

Web archiving presents a number of technical and organizational challenges for libraries. The University of Alberta Libraries has been using Archive-IT to manage a web archiving program for since 2009. This presentation will describe the history of web archiving at the University of Alberta and show the evolution of those services over time. Web archiving is not just technically challenging, it can also be organizationally challenging. Alberta has elected to use a distributed model for collection management by spreading the work for collection development and maintenance across subject librarians and library support staff. Some of the challenges of such a management plan include collection scoping, skill transfer, quality assurance, and metadata creation. The libraries also collaborate with regional and national consortia while working to expand services to researchers and casual users of the library. Attendees will takeaway lessons about collection management, collaboration, and research services for web archives.

Keywords:
Web archives
Library collections management
Distributed collaboration

Introduction

Web archiving presents a new organizational challenge for libraries. Should web archiving be a special collection for libraries or should it be integrated across the spectrum of collecting activities? Is it a function of a centralized department/person, or distributed across traditional library boundaries? How should library archives interact with researchers? Born digital materials intersect the needs of many different disciplines and departments within a contemporary research library. The presentation will present the virtues and drawbacks of a distributed and emergent management organization for web archives at the University of Alberta.

History of Web Archiving at the University of Alberta

In 2009 the Heritage Community Foundation approached the University of Alberta Libraries for help archiving 80 purpose-built websites that were created to celebrate the unique cultural heritage of Alberta. The libraries surveyed the available tools for web archiving and chose to subscribe to the Archive-IT service created by the Internet Archive.

At the same time a Born Digital Working Group was formed within the library to evaluate existing technologies for managing born digital materials, develop policy, and create procedures. The BDWG reported its results to the Collection Development Committee over the next few years. By 2014 the working group had served its purpose by creating a core web archive policy document modeled on policies used by the Internet Archive. A set of procedures for proposing, creating, and managing web archive collections had also been proposed and approved by the library leadership team.

Since starting in 2009 the web archiving program has grown to include 18 different collections. The current collections managed by the University of Alberta Libraries cover a variety of subject areas. A core area for collections is government and university documents which have migrated to solely digital publishing models. There are active collections in specific disciplinary areas such as education, health sciences, humanities computing, politics, and economics. The circumpolar collection is an example of an interdisciplinary collection that grew out of an existing print collection in support of numerous researchers on campus. Finally there are a number of event-driven collections that have been established to capture provincial and national events, such as the Alberta floods of 2013 and the Idle No More aboriginal social movement.

The libraries adopted a distributed management philosophy for developing web archive collections, relying on the experience of subject liaison librarians to propose new collections, configure web crawling parameters, perform quality assurance on crawls, provide basic metadata, and introduce the results to users. The workflow for a collection begins with the proposal. Collections can be proposed by anyone, including researchers, librarians, and the public. The proposal is sent to the Collection Development Committee, which reviews the proposal for the appropriate fit within the overall library collection policy. If approved the proposer works with the Archive-IT technical team at Alberta to get the collection setup by importing seeds and setting appropriate filters, running test crawls, and scheduling regular crawls. After the initial setup the responsibility for collection maintenance and monitoring devolves to the subject librarian who serves as the collection coordinator.

The University of Alberta Libraries also works with other libraries in Canada. A regional collaboration was established with the Council of Prairie and Pacific University Libraries in order to purchase ongoing licenses for Archive-IT, share best practices, and coordinate web archiving efforts in Western Canada since 2013. At a national level, the library contributes to discussions about web archiving through the activity of the Canadian Association of Research Libraries.

Many libraries and institutions involved with web archiving assign an individual to oversee the collection process. So far Alberta has relied on subject librarians to take on the responsibility of collection management, instead of centralizing the function in any particular office or person. The hiring of a web archiving coordinator has been discussed but the

position has yet to be permanently filled. Over the years a number of annual interns and postdocs have contributed to various aspects of web archive management through the development of policies, training materials, evaluation programs, and outreach to research communities. The University of Alberta Libraries recognize the importance of web archiving and have expended the resources to build an institutional team but still face a number of challenges.

Organizational Challenges for Web Archiving Management

There are specific organizational challenges related to web archiving at the University of Alberta which may be helpful lessons for building new, or improving existing, web archiving service at other libraries. A distributed management system which spreads the responsibility for web archiving across many different subject librarians can make coordination, training, policy making, metadata, stewardship, and quality assurance more difficult. The following are some of the challenges we have faced.

Determining the appropriate scope for web collections is a challenge for all libraries (Niu, 2014). The circumpolar collection, for example, spans many disciplines, which means multiple librarians are involved in its management and that multiple faculty may be impacted by particular decisions. The collection attempts to cover material related to the polar regions across many different institutions and geographies, including research groups at other universities, non-profits, and government institutions which may be in countries other than Canada. One recent challenge has been dealing with data files, such as map images and geographic information system files. These files can be quite large and have taken up significant space in the UAL Archive-IT data budget. They raise questions about whether the University of Alberta should be responsible for preserving data from other institutions, and whether the data is useful to current or future researchers. In order to answer this challenge the current collection guidelines are being reevaluated by consulting with affected subject librarians and research groups. But the issue of preserving data from institutions outside of Canada or the University of Alberta is more general and affects all libraries. At the University of Alberta Libraries a more general guideline, across all of the collections, about collecting data files is being considered.

The skill and time needed to manage a web archive collection is often underestimated. Many of the collections have annual, semi-annual, or quarterly scheduled crawls. The reports for these crawls are sent to the collection coordinators who may not have dealt with web archiving since the last report. They must review the report for quality and match with collection policies but may not recall the appropriate procedures or best practices for performing the evaluation. Some of this can be mitigated by more detailed documentation and further training, but skills erode over time regardless of how conscientious training and documentation efforts may be. Most subject librarians are responsible for a multitude of tasks and web archiving activity may be infrequent or a low priority. The libraries have addressed this problem by assigning personnel to a web archive technical team who are available to help subject librarians with collection management.

Quality assurance for web archive collections is a challenge for many institutions. Tools such as Archive-IT provide a useful interface for web archivists by offloading the task of maintaining the software infrastructure needed for web archiving. But the tools available for quality assurance are still manual and require significant time during the review of crawl reports. Improving quality assurance through the development of automated tools is one

response and some organizations are working toward that goal, but until such upgrades are made available there is limited time for quality assurance for the University of Alberta web archive collections.

Metadata is another challenge faced in the distributed web archiving organization at the University of Alberta. Archive-IT uses three levels of metadata: collection, seed, and document. Collection level data is relatively easy to create and can often be gathered during the collection proposal phase when the boundaries and selection criteria of the collection are being decided. Seed level metadata can also be added during the collection development phase by asking collection coordinators to enter the information while the collection is being setup. Adding metadata after a collection has been setup is harder because the work may be an interruption to the already ongoing tasks of collection management. In these cases the metadata work may be passed onto the bibliographic services group at the library. The most detailed document level metadata is the most difficult to collect simply because the volume of data gathered by web crawls is so large. In these cases the UAL has decided to do targeted cataloging based on file types, such as PDFs, and other criteria. Almost all of the efforts to add document level metadata are usually assigned to the bibliographic services group, which means coordinating with another organizational group. Given the amount of work needed for metadata description there does not seem to be any other way to accomplish the task except to divide the labor.

Transferring web archive collections between librarians is another challenge. So far there have been few cases when collections needed to be transferred but when they have arisen the most common problem has been defining the exact boundaries of the collection. In the initial enthusiasm to create collections in Archive-IT some very broad collections were created which are now becoming unwieldy. The most likely response will be to divide some of these collections into more focused subsets, but that work depends on the cooperation of multiple stakeholders. The lack of any technical tools within Archive-IT for transferring content between collections adds an additional challenge to changing existing collections or merging content between collections.

Looking beyond the context of a single institution, there are important challenges for web archiving across the country and internationally (Szydlowski, 2010). Library and Archives Canada has been archiving the web since the middle of the 2000s and is working on making more of the collections available to the public. In Western Canada, the Council of Prairie and Pacific University Libraries meet regularly to discuss issues around web archiving. The Canadian Government Information Digital Preservation Network is another collaboration across Canada to collect and preserve digital government materials. There is no doubt that collaboration is important because no single institution has the resources to collect everything. Libraries and Archives Canada, for example, does not have the resources to collect political material at the provincial or municipal level, so other institutions need to fill in the blanks. But not every institution has a web archiving program and each program is at a different development stage. The University of Alberta is lucky to have five years of experience in web archiving, even if there are things that can be improved. Sharing some of that experience with others within Canada and throughout the world is the purpose of this paper.

Conclusion

The University of Alberta Libraries began web archiving activity in 2009, since then the web archive program has grown to include 18 different collections and 9.4 terabytes of data. The distributed management structure of the web archiving program has presented challenges for managing collection scope, maintaining librarian skills, quality assurance, metadata, and collection stewardship. Web archiving is complicated at many levels and this paper presents the experiences at the University of Alberta in order to help other libraries understand some of the challenges of different management structures. No single library or institution can manage the process of archiving the web. Collaboration is necessary within an institution as well as across national and international boundaries.

Acknowledgments

Thanks to Geoff Harder and the University of Alberta Digital Initiative Team. Thanks also to the Council on Library and Information Resources – Digital Library Foundation Postdoctoral Fellowship Program.

References

Niu, J. (2014). Appraisal and Selection for Digital Curation. *International Journal of Digital Curation*, *9*(2), 65–82. http://doi.org/10.2218/ijdc.v9i2.272

Szydlowski, N. (2010). Archiving the Web: It's Going to Have to Be a Group Effort. *Serials Librarian*, *59*(1), 35–39.