



# Preserving Seeds of Knowledge: A Web Archiving Case Study

Jeremy M. Heil and Shan Jin, CRM, CIP

This case study provides details about how a cross-functional team has collaborated on a year-long pilot project to preserve website content and what it has learned at its mid-way point.

**A**fter nearly three years of planning and preparation, the Queen's University Archives in Kingston, Ontario, Canada, embarked on a one-year web archiving pilot project in June 2016 to:

1. Establish a baseline for ongoing preservation of the university's *web presence*, or the records it has made available through online protocols, such as hypertext transfer protocol (HTTP), file transfer protocol (FTP), and their variants
2. Determine what resources would be required to operationalize a web archiving program

Having reached the mid-point of this project, Queen's archives has developed a standard of practice and gathered enough information to set out policies and procedures governing web archiving at the university that could be useful for other organizations.

## Background Information

Queen's University is one of the oldest and most reputable universities in Canada. Founded in Kingston in 1841 by a royal charter issued by Queen Victoria, today it is a mid-sized university with an enrollment of about 22,000 students and 8,000 staff.

The university has always been proud of its long and rich history, and it has a lengthy archival tradition. Its archives started to form when the first university archivist was hired in 1960, and it has continued to evolve and expand its holdings. Today, the archives is part of the university library, and it houses approximately six miles (10 kilometers) of textual records, 2 million photographs, tens of thousands of architectural plans and drawings, and thousands of sound recordings and moving images.

The university archives, whose mandate includes the preservation of the university's records, was secure in handling the acquisition and preservation of the university's analog records, but the university business had moved into the digital. Many publications and records, such as course calendars, committee minutes, and general university information, can now be found only on the university website.

This can be problematic when websites change or when office functions are adjusted and absorbed into other areas. Without a web archiving program to preserve it, a massive amount of information could be lost, and the record of the university's daily life could be erased within a few years, creating an institution with an extremely short-term memory.

## Preserving the Online Record

Beginning in 2013, archivists made many attempts to garner support for a web archiving solution, finally receiving approval in 2016. The research and project submission focused on the need for stable and standard formats for long-term preservation and the ability to make captured websites available to researchers.

The international web archive preservation standard ISO 28500:2009 *Information and documentation – WARC file format* best suited the university's needs, allowing it to store both the content and representation information from HTTP; maintain data record integrity; ensure the capture of only new or altered content (deduplication); and maintain captured and added metadata in an open system.

## Selecting a Software Tool

Archivists at Queen's made some early attempts to capture websites using free software products that copy the entire open structure of a website, mirroring the website on a local hard drive. While effective, the archives was still faced with ongoing preservation concerns and the challenge of capturing ever-changing online content, especially for live websites with a long history.

After thorough research on available technologies, the archivists submitted a proposal to subscribe to the online Archive-It service through the Internet Archive. The WayBackMachine (<https://archive.org/web/>) demonstrates how the WARC format is more robust and stable in the longer term and is easily accessible to researchers.

The archives' decision, then, needed to weigh the benefits of outsourcing against installing software, storage, and search capabilities on a local server. It was determined that training to build this expertise in-house would be an unwise use of scarce resources, especially for a pilot project.

## Assembling a Team

In October 2015, the library leadership team approved the pilot project proposal. The team consisted of:

- The digital and private records archivist, who chaired the team
- A project manager
- Three archivists with responsibilities for university records and private manuscripts
- The head of the library and archives' discovery and technology services

- The university's records management and chief privacy officer

Each team member brought unique and diverse expertise to the project:

- The project manager was assigned by the library to assist in this project under a newly implemented project planning structure.
- Archivists took the lead on selecting content and evaluating the effectiveness of the preservation platform.

The project's purpose was to begin preserving important Queen's University web pages to ensure that the informational, evidential, and historical value they provide is maintained and remains authentic.

- A representative from the library's technical services department managed the contract and provided feedback on technical issues.
- The records management and privacy office advised on security and privacy issues related to the captured content, and it facilitated communications with university departments.

One concern faced by public bodies is the storage of private data across political boundaries. Because the chosen service operates outside of Canada, the university archives needed to comply with the university's "authorization to operate" process (see <http://queensu.ca/cio/information-security-office/authorization-operate>) and verify that all the captured data would be of a public nature. Since the pilot project scope did not include any password-protected websites – which is a good indicator of where private data resides – it could proceed.

## Establishing Project Objectives

The project's purpose was to begin preserving important Queen's University web pages to ensure that the informational, evidential, and historical value they provide is maintained and remains authentic.

The objectives of the pilot project were to:

- Connect with stakeholders to learn about their current practices for managing web content
- Work with stakeholders to promote an understanding of the importance of preserving web content as part of the university record and to build best practices for content management into website administration
- Establish user requirements for a web archiving service at Queen's

- Test a common software tool used for web archiving and monitor its day-to-day operational requirements
- Provide information to pilot user groups and potential user groups on the program's effectiveness
- Consider the policies, procedures, and ongoing costs of an operational service

- Provide recommendations to the university records committee about implementing an ongoing service

The objectives identified to operationalize an ongoing service would be to:

- Manage, preserve, conserve, and make accessible the university's information assets
- Support the university's teaching, research, service, and administration
- Fulfill the university's legal mandate under the Province of Ontario Freedom of Information and Protection of Privacy Act to ensure that reasonable measures respecting the records under control of the institution are developed, documented, and put into place to preserve the records
- Provide an excellent resource for history projects, text analysis and big data mining projects, experiential learning projects, and digital humanities projects, such as Web Archives for Historians (<https://webarchivehistorians.org/>), International Internet Preservation Consortium ([www.netpreserve.org/](http://www.netpreserve.org/)), and Web Archives for Historical Research (<https://uwaterloo.ca/web-archive-group/>)

## Planning Communications

Before the archives began the website capture process, it reached out to select stakeholders, including webmasters and domain administrators, to discuss the nature of the project and communicate the anticipated outcomes (see <http://archives.queensu.ca/search-our-collections/web-archives-project>). These meetings also served to identify whether other groups or individuals on campus were already pursuing their own web archiving strategy in any form.

Meetings with information technology services and the department of communications and marketing proved fruitful in many ways, including communicating the need for web archives, confirming the absence of alternate web archiving strategies on campus, and identifying further websites for capture in the pilot project.

The necessity of this communication strategy was borne out when conducting a test *crawl*, which is the active capture of a *seed* (a specific website URL identified for the crawler to capture), that is conducted over a specified length of time per user-inputted parameters.

After an initial crawl of the university student newspaper website did not capture the complete scope of the website, a second crawl was run over a longer period. The results of this crawl showed that a website administrator recognized the increased activity of the crawler

and subsequently blocked the web archiving service from continuing its work, resulting in an error. The project team needed to contact the student newspaper to ensure that the service would be unblocked and thus be allowed to crawl and capture the entirety of the website.

## Identifying Key Webpages

Archivists identified a core of four thematic areas relating to the university: administration, faculties, and

The ability to add metadata to reflect the contextual details of a website is key to good web archiving practice. The team adds Dublin Core metadata to each seed to clarify...and assist with the discovery of content.

services; university publications; faculty research; and student organizations and other affiliated groups. They chose to focus exclusively on the first two subject areas for the pilot, limiting the official university pages to what were considered the core sites related to the university mission.

While the archives actively selected the key seeds for archiving, the university units also played a part in this process. One of the key stakeholders of this project is the records management and privacy office; it works with archives to develop and approve records retention schedules to help university units manage their records, including official webpages.

A records retention schedule was created for web content management, covering the creation and maintenance of the content of the university's official web pages. It requires that departments or units on campus consult the archivist for final disposition of web content when a website is discontinued or when administrative units are merged, altered, or eliminated.

## Capturing Web Content

Working within a 500 GB space budget over one year, the project team had to be frugal with the scope of web content captured, paying close attention to what file types were being captured in each crawl.

### Test Crawls

The Archive-It service recommends that any initial crawl be conducted as a "test crawl" to let the archivists examine the *scope* (the rules followed during a crawl) of what was captured without applying unwanted content towards the space allocation budget. If the test crawl re-

sults in too much extraneous information or too little of the core content being captured, the team can adjust the crawl scope, either adding or omitting specific website elements or defining whether to include links to external websites.

This frugality also prompted a focus on how often web content changes within specific websites. Administrative websites tend to be updated with news items, but the main content remains static over an academic year or longer unless the university makes changes to the functional structure of certain offices. With this in mind, the team chose to capture these websites only once annually, with the knowledge that any news items of importance would typically remain on a website for at least one year.

By comparison, the Queen's *Journal* student newspaper changes weekly, but it maintains its own archives of past issues. So, here the team opted to conduct a deep capture of the entire website with its archives only once and schedule quarterly crawls to capture newly added content.

### Capture Limitations

Web capture in general can be limited by many factors, including the use of dynamic content on a website, special scripts, or other custom design elements. The university is legislated to adhere to certain web design standards relating to accessibility, and official pages created or significantly updated in the past two years have met these standards. As a result, the web crawls had little difficulty capturing the majority of the content, look, and feel of each website.

The project team acknowledges some acceptable loss when faced, for example, with a carousel of images on banner pages, where only one or two images would be captured as the crawler scanned that page. Examining test crawls reveals limitations like this and provides the chance to alter the scope of the crawl before committing the capture to storage.

## Archiving: Good Practice

The ability to add metadata to reflect the contextual details of a website is key to good web archiving practice. The team adds Dublin Core metadata to each seed to clarify anything that is not already in the website and assist with the discovery of content. This has been especially useful where websites omit common details like page titles or summaries. The added metadata also helps show the relationship between similar content in different websites, making the web archive more discoverable and researcher-friendly.

## Taking Next Steps

While this pilot project nears completion, the team is looking at the next steps toward operationalizing this

service. After these months of practice, the team can now draft policies and procedures for the ongoing use of a web archiving service and estimate annual operational costs.

Team members have already identified key webpages to be preserved and drafted plans to slowly increase the coverage over time. Next comes integrating web archiving practice into the university's web content management and records management practices. Finally, the archives will examine whether the web archiving service could be expanded to support researcher use at the university, with an outlook to capture content that falls outside the normal scope of university and university-related record preservation. **E**

## Project Acknowledgements

The Queen's University Web Archiving Project was made possible through the hard work of its team members, including Heather Home (public services archivist) and Deirdre Bryden (archivist, university records), both of whom were integral in planning the project, as well as Sandra Morden (head, discovery and technology services), Carolyn Heald (director, records management and privacy office), Paul Banfield (university archivist), and Joe Davis.



**About the Authors:** Jeremy Heil has been the digital and private records archivist at Queen's University Archives since 2001. Having earned a master's of archival studies degree from the University of British Columbia, he presents regularly on topics relating to digital records and teaches workshops on digital preservation. Most recently, he has been involved in developing archival accession standards and is serving as the managing editor of *Archivaria*. Heil can be contacted at [heilj@queensu.ca](mailto:heilj@queensu.ca).



Shan Jin, CRM, CIP, is a records analyst/archivist at Queen's University archive. She is a Certified Records Manager and a Certified Information Professional and has contributed to several ARMA technical reports. She earned a master's degree of library and information studies from Dalhousie University. Jin can be contacted at [jins@queensu.ca](mailto:jins@queensu.ca).



# Purchasing Power at Your Fingertips

Your Connection to RIM & IG  
Products and Services

## BUYER'S GUIDE ONLINE



**Advertise!**

Add your company's name to the online **Buyer's Guide**. [www.arma.org/buyersguide](http://www.arma.org/buyersguide)  
Contact Jennifer Millett at [jennifer.millett@armaintl.org](mailto:jennifer.millett@armaintl.org) today.