@Risk North 2

Digital Collections
Collections numériques en péril
2018.11.09 - MONTRÉAL

Report of the Open Forum

Published March 2019

Background / Introduction

Presented by the Canadian Association' of Research Library (CARL) in collaboration with the Canadian Research Knowledge Network (CRKN), Library and Archives Canada (LAC) and Bibliothèque et Archives nationales du Québec (BAnQ), this event was a key undertaking of CARL's multi-stakeholder Digital Preservation Working Group (DPWG). Formed in 2017, the DPWG's aim is to assess digital preservation capacity, available resources, and funding opportunities within the Canadian research community; to identify and promote relevant approaches, standards, practices, and technologies; and to liaise with relevant international efforts in order to cultivate an appropriate knowledge of the broader field of digital preservation within Canada.

The event was a planned complement to *@Risk North: Collections en péril*, a similar gathering one year earlier, which had sought to examine shared print collecting at the regional and national levels. (That event had in turn been inspired by the Center for Research Libraries' 2016 Collections Forum *@Risk: Stewardship: Due Diligence, and the Future of Print*.)

Prior to this full-day open forum, there had been few opportunities for those involved in digital preservation across Canadian memory institutions to meet and reflect on investments and capacity for digital preservation within the country's various memory institutions. Attendees and presenters at @Risk North 2 included representatives from libraries, archives, museums, and community organizations, whose presentations and discussions sought to increase understanding of the current state of digital preservation readiness, resources, and collaborative initiatives in Canada, with a view to identifying opportunities to strengthen our collective capacity for digital preservation in Canada.

The Big Picture

Before delving into the current state of practice within Canada, Clifford Lynch, Executive Director of the Coalition for Networked Information (CNI), engaged the

attendees in a broad reflection on evolving challenges related to digital preservation. During his talk, Lynch challenged attendees to "pivot from focusing on the work defining the mechanics of digital preservation to broad strategy."

Referring to the scholarly record, Lynch asserted that "It is increasingly understood to include not only the outputs of research, but also the evidence upon which the research is based, as well as everything else necessary, such as software, methodology (sometimes now recorded as video), analytic workflows. This is a reasonably bounded set of entities to preserve, although we are still trying to understand what we need to retain in datasets." Although the sheer amount to be preserved poses clear challenges, and although the content that is not hosted by the larger publishers is more at risk, he pointed out that it is the broader cultural record that is "much less bounded, more amorphous and fluid, and much less safe." He characterized this as a slowly unfolding disaster.

One reason he cited was that, whereas memory organizations were once able to buy and collect the cultural content that would eventually become the basis for future scholarship, the terms of service for web-based streaming services effectively cut out memory institutions. Whether for music, film, or ebooks, when individuals no longer own collections, they can no longer donate them to archives. Lynch also expressed concern for the preservability of news sites, social media, and the broader web. Whereas websites were once relatively static, practically all now include ubiquitous personalization, which makes it impossible to capture anything beyond a "curatorial approach to sampling." Lynch posited that new strategies will need to be explored to respond to these market shifts. National libraries may need to play a larger role; legislative responses may be needed; and, we should be reaching out to authors, publishers, and creators about the need to work together to address current and future preservation needs.

A Canadian Snapshot

A key component of the day was a presentation by Grant Hurley (Scholars Portal) of findings from the CARL Digital Preservation Working Group's *Survey on Digital Preservation Capacity and Needs at Canadian Memory Institutions*. The results were based on surveys completed by 26 research libraries (CARL member institutions) and 25 other memory institutions (a mix of smaller academic libraries, community/not-for-profit organizations, and government-based organizations).

Hurley painted a portrait of a small but active community tackling many collection types and formats on a broad scale, but lacking sufficient resources to achieve program maturity and sustainability. Results indicate that there is a broad commitment to digital preservation across organizations, though most cited that this commitment has not translated into resources for program development. A similar

pattern emerged in regard to policies and procedures. While 59% said that the development of documented procedures is in progress, 75% said they lacked time and resources to complete this work; 47% also responded that current procedures are ad-hoc or project specific. Adoption of tools for digital preservation processing and digital forensics was low overall. Platforms for access to digital content is a more mature area, especially among CARL institutions, though more consolidation/integration of tools may be needed. Respondents noted that there are privacy/intellectual property issues, and that scalability for existing tools is a barrier.

Respondents store their digital assets across a variety of different storage systems/media, and the transition to cloud or other replicated storage networks has been slow, especially outside of larger academic libraries. A relatively large proportion of 37% of assets on average are still being kept on fragile externa media. This is especially the case for smaller organizations, some of whom depend on hard drives or other removable media to store assets.

Low staffing for digital preservation was seen overall, and most staff involved in digital preservation do so as a small portion of their responsibilities. Whereas the average FTE devoted to digital preservation across all institutions was 1.11, a minority (8%) of institutions reported a much more significant staffing (5 full-time staff or more each). 65% of respondents had less than 1 FTE for digital preservation work in total across all applicable roles listed. Many respondents were hopeful about increasing staffing in this area, but often cited that they intend to reassign current staff to achieve this.

Half of respondents reported that funding for digital preservation is inadequate, and several stated that dependence on short-term money for long-term activities was a concern.

A panel commented on these findings, with Lisa Goddard (University of Victoria) noting that overall, the study points to some good momentum and coordination at the national and regional level. Building capacity at the local institutional level appears to be a concern, but could be an opportunity for collaborative work, possibly along the lines of the Portage networks of expertise. Geoff Harder (University of Alberta) suggested that in addition to this survey, much can be learned by documenting our 'epic failures' – lessons learned that may benefit the rest of the community. He recounted a case where the University of Alberta lost 11 e-theses due to a series of unfortunate events. They were able to recover the files, but only by going back to creators and admitting to their error. The issue taught them a lot and led to several operational improvements. Harder suggested that we need to share more openly about what is working and not working and be willing to learn from our mistakes. The survey suggests we do well to collaborate as resources are scarce, which only increases risk. Collaborative tools, platforms and services are part of the solution

to ensure safeguards are in place within and outside of our institutions. Mireille Laforce (Bibliothèque et Archives nationales du Québec) agreed that the community is making progress, but the lack of financial resources, people, expertise, and tools, is still striking. She suggested that it would be interesting to run this survey again periodically, perhaps in 5 years.

At Scale Efforts

A substantial part of the day's presentations dealt with 'at-scale' efforts. The discussion was introduced by Carole Urbain who spoke briefly about the role of the National Heritage Digitization Strategy. Beginning at the institutional level John Richan described the process of setting up a new digital preservation program at Concordia University, and Steve Marks discussed rethinking an established program at the University of Toronto. Moving then to the regional level, Corey Davis described the Council of Prairie and Pacific University Libraries (COPPUL)'s open source consortially shared LOCKSS-based WestVault system, while Kate Davis (OCUL Scholars Portal) discussed some of the challenges of ramping up the Scholars Portal platform to the national scale and of delivering the new Permafrost storage service. Pascale Montmartin from BAnQ and Faye Lemay from Library and Archives Canada spoke from the perspective of national libraries and archives, having to develop their own practices in the context of international practices, while also playing a leadership role for the nation's memory institutions and launching collaborations with these groups. Karin MacLeod (also from Library and Archives Canada) further described LAC's efforts to streamline digital preservation tasks by acquiring preservation-ready content in digital form whenever possible.

These presenters brought sage reflections and many open questions: Kate Davis suggested that governance rather than technical issues is the real challenge when ramping up a service to the national scale. John Richan asked, "How much processing is enough?" When explaining that the University of Toronto had moved from a monolithic all-in-one repository-and-access-point to multiple platforms with a new focus on tracking, Steve Marks cautioned against 'performative preservation' – preservation as a one-time rather than an ongoing activity. Corey Davis echoed Clifford Lynch's earlier comments about the importance of a community-owned and operated system that allows for decorrelated and independently administered content managed independently at several different locations while being up to the task of ensuring authenticity and security of the scholarly record.

Later in the day, an attendee suggested that in light of the ever-expanding digital record, the only way to possibly achieve the preservation of all of our country's digital cultural assets will be to make it a practise to regularly inform each other of the work we are doing, to avoid duplication of effort at all costs, share best practices, and develop trust in each other's work.

Specific Content Types

The day also included sessions describing the preservation of specific media or content types, and innovative approaches. Annie Murray (University of Calgary) spoke about the Herculean task of preserving the 40,000 audiovisual recordings that make up the EMI music collection hosted at the University of Calgary, while Mireille Nappert (Canadian Centre for Architecture) spoke about preserving the software necessary in order to make CAD files accessible and usable via emulation. Umar Qasim (University of Alberta) on behalf of Portage's Preservation Expert Group, spoke to the timely topic of preserving research data. Meanwhile, Lisa Goddard (University of Victoria) described the efforts of the Endings Project[1] to begin addressing the preservation of Digital Humanities projects (an area felt to have been underexplored to date) through a survey of digital humanists that will lead to a fixed capture of some projects' final site and their supporting database content. However, she noted that the initial question they asked their DH scholar interview subjects proved telling: In asking "How and when do DH projects conclude?" they discovered that often, they do not – which adds some unexpected complexity to the matter of preserving them.

Programming Solutions

Two projects described the use of in-house programming to simplify repetitive and time-consuming tasks. Jess Whyte (University of Toronto) aimed to minimize error and simplify a frequently occurring task when she created 'Floppycapture.py', a script that minimizes repetitive typing when processing floppy disks. Meanwhile, Tim Walsh at Concordia University set out to solve a much larger issue. Inspired by a system used by police to find personal information in large files, he is in the process of building 'Bulk Reviewer' (currently in alpha version), a system for identifying, reviewing, and removing sensitive files from archives.

Failures, Risk, and Trust

Enmeshed in discussions about software and practices were three recurring related themes: the need to admit and record failures, the difficult task of establishing and retaining the trust of those whose work is being preserved, and the need to recognize and mitigate risk.

Just as Geoff Harder urged us to admit failures earlier in the program, Clifford Lynch returned to this theme in his closing comments, adding that especially those stories with unhappy endings need to be documented, as they are essential for making the case for preservation work.

[1] https://onlineacademiccommunity.uvic.ca/endingsproject/

In her presentation about the work of the 'Indigitization'[2] project, Sarah Dupont (University of British Columbia) stressed the importance of establishing trust when working with projects centred on the preservation of Indigenous oral histories. It is for this reason that Indigitization's grants to community groups do not require that the digitized content be available publicly but rather leaves the access policy decisions to each group.

Cliff Lynch returned to the theme of risk at the end of the day, stressing the importance of taking great care when adopting machine processing (e.g. Bulk Reviewer) to reduce human processing requirements. He admitted that without such automated systems it would likely be impossible to preserve the digital world at scale, but suggested that before committing to this direction, we first "need to have a good understanding of where humans tend to fail, and where machines tend to fail." We then must strive to explicitly define what constitutes acceptable risk, and establish "best practices for risk for different classes of collections."

Next Steps

In addition to the presentations, participants engaged in table discussions where they were asked how to move forward in terms of building on successes to date and addressing current gaps, and to consider the role of organizations in helping with this work.

Key suggestions highlighted by the attendees were:

- More frequent communication about work at all levels (in terms of collections but also practices – and not waiting until the work is done to share information about projects with other institutions) in order to prevent duplication of effort and increase opportunity to learn from each other;
- Follow-up activities to this event to sustain the momentum; more opportunities to gather as a community, even if these are just regular virtual gatherings;
- Increase advocacy efforts to create better awareness of the issues, challenges, and necessity of digital preservation with members of high-level administration so that more resources can be devoted to this task and risk can be minimized;
- More discussions and efforts on how to ensure diverse representation across the content that gets preserved;
- More strategies for born-digital content and more formalized complementarity between national library efforts to preserve the published cultural record and those of research libraries;
- Development of strategy to address the human resource training needs on campuses identified in the DPWG's Digital Preservation Readiness Survey (e.g. application for SSHRC Connections Grant to bring training to various regions);

[2] http://www.indigitization.ca/

- Continue to include organizations across different types and sizes rather than limited to one association or institution type;
- Coordination of all the above.

CARL will ensure follow-up in terms of these suggestions as well as the gaps identified in the final report of the digital preservation capacity survey. This will be reflected in the next iteration of the Digital Preservation Working Group's work plan, and may include an eventual @RiskNorth 3, the scope and timing of which will be carefully considered and for which further community feedback will be sought.