Digital Preservation Initiatives in Ontario: Trusted Digital Repositories and Research Data Repositories

Wayne Johnston
Head, Research Enterprise and Scholarly Communication
University of Guelph Library
wajohnst@uoguelph.ca

Abstract

The first in a series of two articles dealing with digital preservation, this article discusses repositories, more specifically Trusted Digital Repositories (TDR) and Research Data Repositories. The focus will be on the TDRs at Scholars Portal and Library and Archives Canada (LAC), and the data repository at the University of Guelph.

Keywords

trusted digital repository; digital preservation; research data repository

Trusted Digital Repositories

On one level, a Trusted Digital Repository (TDR) is a set of metrics that are used to certify that a given repository is an appropriate custodian of a collection of digital assets. More than an array of abstract measures, however, a TDR represents a stable and sustainable organization, a set of policies and procedures for sound management of the digital objects, and a robust and secure technical platform.

To be certified as a TDR an organization must undergo a meticulous audit that ensures the proposed TDR meets all criteria of the ISO 16363 standard. The first category of criteria is organizational infrastructure. This includes issues like governance and organizational stability, procedural accountability and policy framework, and financial sustainability. The second category is digital object management which includes processes for ingest, preservation planning, information management and access. The final category is technical infrastructure and security risk management which includes appropriate technologies and security systems.

A key aspect of the TDR is development of preservation metadata. Preservation Metadata Implementation Strategies (PREMIS) was a working group that developed a data dictionary for digital preservation. The name "PREMIS" is now the de facto name for that data dictionary. It includes concepts such as provenance (Who has had custody/ownership of the digital object?), authenticity (Is the digital object what it purports to be?), preservation activity (What has been done to preserve the digital
object?), technical environment (What is needed to render and use the digital object?) and rights management (What intellectual property rights must be observed?)

**Scholars Portal’s TDR**

Scholars Portal (SP), an initiative of the Ontario Council of University Libraries, began the certification process late in 2010. After many months of documenting procedures and policies, their audit was initiated by the Center for Research Libraries (CRL). Steve Marks, the Digital Preservation Policy Librarian at SP, reports that in April of 2012 the CRL conducted a two-day site visit for an in-person review of the SP TDR. CRL met with all members of the SP team and saw demonstrations of the various systems involved. Following that site visit the SP TDR team received a report from CDR including items requiring follow-up. The SP TDR team members are working on a response which should wrap up this very rigorous certification process.

The SP TDR platform essentially builds preservation capabilities onto the MarkLogic environment already in use for hosting journal content. Among other things this meant adding preservation metadata including checksums to monitor bit rot. The system now includes, where possible, the full text of the journal articles in XML format, descriptive and discovery metadata, and preservation metadata using both PREMIS and a structural metadata format similar to METS. (METS is the Metadata Encoding and Transmission Standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library.) The system has been designed to be robust while also remaining streamlined enough to accommodate the vast amount of data to be processed. The strategy also incorporates a storage array that is shared with the University of Toronto, used for the non-XML content which is generally in PDF form. While MarkLogic has proven to be a very effective platform for SP both in terms of content delivery and long-term preservation, their strategy is not tied to MarkLogic should future needs necessitate a platform change.

While the journal content is the immediate high priority there are a number of other forms of content that are likely to be included in the TDR over time. eBooks make up a rapidly growing collection at SP, and they will probably fit quite readily into the existing journal platform. The eBooks platform already houses its content in MarkLogic so it won’t be a problem to incorporate them in the TDR. In fact, the eBook platform will be undergoing a major rewrite in the near future, and that would be an opportune time to make the collection compliant with the new preservation requirements. Data in the ODESI (Scholars Portal’s web-based data exploration, extraction and analysis tool) and GeoPortal (Scholars Portal’s geospatial data discovery tool) platforms will be more challenging to include. In the case of ODESI the proprietary format dictated by Nesstar (the software system for data publishing and online analysis) is not conducive to preservation.

SP’s new Dataverse platform for research data will be included in the TDR but probably using a different environment such as SafeArchive. SafeArchive is a policy-driven solution for archival storage and replication that oversees harvesting of content from
content stores such as Dataverse and DSpace. LOCKSS (Lots of Copies Keeps Stuff Safe) is a system for distributing digital content among peer institutions to ensure access in the event that the primary location becomes unavailable. It could be used for replication while SafeArchive would monitor that activity and provide a preservation audit. Another technology under consideration is iRODS. iRODS is adaptive middleware that performs data management functions through a set of rule-oriented microservices.

When asked about plans for including electronic theses from the member institutions, Steve Marks said it is something that will certainly factor into their plans, but there is no specific roadmap at this time. Rather than including electronic theses and dissertations (ETDs) on the eBooks platform it might take the form of a dark archive (i.e., available for disaster recovery but not for direct access by users) supplementing primary content delivery through each institution's repository.

**Library and Archives Canada's TDR**

Library and Archives Canada (LAC) has a major role to play in digital preservation in Canada.

> Preservation is a fundamental responsibility through which the National Archives of Canada ensures the continuing availability and authenticity of the archival records that it holds in trust for present and future generations. The NA recognizes that preservation is a pervasive function that is integral to all archives activities from acquisition through to access; every staff member plays a part. The preservation function is also implicit in NA’s responsibilities to facilitate the management of records of government institutions and ministerial records and to support the archival community (Library and Archives Canada, "Preservation Activities").

Faye Lemay, Manager of Digital Preservation at LAC, was asked to provide an update on LAC's TDR initiative. She explained that their work progressed to a certain point after a couple of years, at which time they conducted a strategic reassessment. The decision was made to move from a comprehensive solution developed in-house to a series of commercial solutions including Sharepoint and OpenText. This change in direction was based on the idea that a more agile approach composed of modular units would provide more flexibility in dealing with future evolutions in the technology. In a presentation at the annual conference of the International Council on Archives, Ronald Surette, Director General of Digital Preservation at LAC, asserted that "[t]he only durable solutions are those that can be changed/replaced easily, quickly, cheaply". The new strategy will feature multiple repositories behind a common integration layer to support control, discovery and access.

When asked about certification, Lemay said that LAC is not pursuing certification at this time, opting instead to focus on development of the technical solutions, but the ongoing investment in policies to support the emerging solutions will feed into certification when the time comes.
The following quote outlines the scope of materials that fall within LAC’s digital preservation programme.

Library and Archives Canada acquires a large scale and broad range of digital content including digital publications, selective websites, large web domains, blogs, electronic government records, digital photos and art, digital audiovisual, geomatics, electronic theses from Canadian universities, digital technical and architectural drawings, private textual electronic records, broadcast data, etc. As well, Library and Archives Canada generates considerable digital content with a large-scale digitization program (Library and Archives Canada, "Speech").

Lemay spoke about the "whole of society" approach to prioritizing targeted collections for the TDR. The LAC website provides an explanation of this concept:

A new approach to appraisal is being developed that takes into account how well a heritage item represents the 'whole' of Canadian society, and which organization should house it. This new approach is being designed both for LAC and for other organizations that acquire and manage heritage material (Library and Archives Canada, "Modernization").

The concept is also being used as the basis for appraisal throughout the government:

The departments that make up the Government of Canada are responsible for managing their own information resources. A new framework that reflects the 'whole' of government is being developed to guide departments on the best way to dispose of irrelevant information and preserve that which supports the government's accountability (Library and Archives Canada, "Modernization").

The whole of society model looks at a number of facets to determine the relevance of collections to Canadian society. These facets include people, organizations, people's roles within organizations, time, place and social domain. It entails additional metadata that complements and extends the metadata related to provenance and preservation.

Lemay explains that TDR collections will not evolve evenly. The current priorities focus on publications and government records. Each collection will be evaluated using the "whole of society" model, and then an appropriate platform will be identified. When asked where e-theses fit into their TDR plans, she said they are not yet at that level of detail.

**Research data preservation**

CARL published an excellent data management toolkit in 2009 titled *Research Data: Unseen Opportunities*. It contains a very useful summary of why researchers should care about research data management, including data preservation. The toolkit lists the following benefits of an effective data management strategy:
Accelerates scientific progress
The sound management of research data will allow researchers to access and understand others’ data and re-use them for their own scientific purposes, thereby speeding up the rate of new discoveries.

Increases the visibility and impact of research
[...] For example, a study of citation rates for cancer clinical trials publications found that clinical trials that shared their data were cited about 70% more frequently than clinical trials that did not.

Ensures compliance with funding agency policies
[...] Some publishers also require that the data connected to their publications are preserved.

Avoids duplication of research
When a dataset is publicly available it is much less likely to be recreated, avoiding expensive and needless data collection/production activities.

 Enables replication and verification of research results
When data are archived and shared, results are repeatable, and data can be used for reanalysis, backing up original research findings. They may also be used to expose errors or inconsistencies with original data analysis.

Enhances collaboration
Making research data discoverable and reusable encourages interdisciplinary collaboration as well as citizen science (Canadian Association of Research Libraries).

There are strong models on the international stage for what can be accomplished to support research data management through national strategies, policies and infrastructure. Particularly noteworthy are developments in the UK, the US and Australia as well as international initiatives in the European context. Canada, however, has been lagging behind. Carol Perry, who oversees the data management programme at the University of Guelph, asserts that "other countries provide inspiring examples of what can be accomplished". Perry goes on to say that Canadians have a lot of expertise, but we are stymied by the absence of a national data management strategy and a lack of infrastructure. "We've been talking about these needs for several decades but very little has been accomplished." While individual institutions aspire to preserve and publish their research data, they don't have the resources to do it alone.

Fortunately, funders in the Canadian context are fully on board in recognizing the value of and need for data management and preservation. The three primary funders of research in Canada, while all supporting data management in principle, are at different stages in terms of how that is reflected in policy. CIHR requires that certain data types (bioinformatics, atomic, and molecular coordinate data) be submitted to a data repository and retained for a minimum of five years. The Social Science and Humanities
Research Council (SSHRC) informs grant recipients that all data must be preserved and made available for use by others within two years of the project’s completion. SSHRC also declares that "[c]osts associated with preparing research data for deposit are considered eligible expenses in SSHRC research grant programs" (Social Sciences and Humanities Research Council of Canada). NSERC has no independent policy at this time. There is a Tri-Council working group tasked with the coordination of access and preservation policies across the three agencies.

Work is underway to address the need for a national data management strategy and supporting infrastructure. The Research Data Strategy Working Group was formed in 2008 to "address the challenges and issues surrounding the access and preservation of data arising from Canadian research" (Government of Canada). They identified a number of key gaps in the existing environment, including policies, funding, data repositories, skills, standards, incentives, roles and responsibilities, and time. A Canadian Research Data Summit was held in Ottawa in September of 2011 followed by a Digital Infrastructure Summit in June of this year in Saskatoon. These summits have aimed to bring together key strategists to identify concrete plans to address the gaps.

**University of Guelph Library’s data repository**

While TDRs are primarily dark archives for purposes of preservation, a data repository is a platform for publication as well as preservation.

Early in 2011 the University of Guelph Library was awarded a $75,000 grant to develop an Ontario Agri-environmental Research Data Repository. Below is the preamble to the successful proposal:

> All too often the results of expensive and time-consuming research as represented by rich data sets are lost due to the absence of sound data management plans. Redundant research is undertaken because the previous research data is no longer available. Opportunities for analysis of data across time are lost along with the historical data sets. Even when data has been properly stored and preserved it benefits no one if it isn't easily discovered, retrieved and repurposed (Johnston).

The pilot project is designed to promote the need for data management plans as well as to establish a platform for easy submission of research data sets, robust and secure storage, discovery, retrieval, re-purposing and long-term preservation. The bulk of the grant money was used to hire a research data technician to assist researchers with preparing their data sets for submission to the repository.

Carol Perry explains that this pilot phase has enabled the Library to establish the core infrastructure including identification of appropriate standards and development of the requisite procedures and protocols. “We have many years of experience providing access to data, and now we are developing expertise in the preservation and publication of data generated by researchers on our campus.” As of this writing the
Ontario Agri-environmental Research Data Repository has published seven data sets with a number of others in various stages of development.

While part of the work of the Library’s service is educating researchers about the need for data management plans and about best practices for sound data management, there is still a need for the type of support provided by the research data technician. There are efficiencies and economies of scale when someone has developed the skills and experience to effectively ensure that data is clean and that the appropriate metadata has been developed. It is enough of a challenge to convince researchers of the need for this service without expecting them to actually do the work.

The repository is primarily based on Scholars Portal’s Dataverse platform providing all the needed features for access and control as well as utilities for online data analysis. The Library is also a participant in NRC’s DataCite Canada service providing additional capabilities for data citation and data linking. It is recognized that the data repository will work in tandem with Guelph’s DSpace institutional repository, called The Atrium. The raw data will reside in the data repository while the scholarly outputs arising from that data will be available in The Atrium.

Watch for part two in this series on digital preservation from Andrew Waller.

Works Cited


Lemay, Faye. Telephone interview. 17 July 2012.


Marks, Steven. Telephone interview. 3 Aug. 2012.

Perry, Carol. Personal interview. 9 Aug. 2012.
