Assessing the Scale of Challenges for Preserving Research Data

Umar Qasim
University of Alberta
2-10N Cameron Library
Edmonton, AB
+1 (780) 492-9861
umar.qasim@ualberta.ca

Chuck Humphrey
University of Alberta
2-10T Cameron Library
Edmonton, AB
+1 (780) 492-9216
chuck.humphrey@ualberta.ca

John Huck
University of Alberta
5-25E Cameron Library
Edmonton, AB
+1 (780) 248-1337
john.huck@ualberta.ca

Leanne Trimble
OCUL Scholars Portal
130 St George St, 7th Floor
Toronto, Ontario
+1 (416) 978-7217
leanne.trimble@utoronto.ca

Alex Garnett
Simon Fraser University
Burnaby, B.C.
+1 (778) 228-5110
garnett@sfu.ca

Dugan O'Neil
Compute Canada
Burnaby, B.C.
+1 (778) 782-5623
dugan.oneil@computecanada.ca

Sean Cavanaugh
University of Saskatchewan
Saskatoon, SK
+1 (306) 966-2674
sean.cavanaugh@usask.ca

Jason Knabl
Compute Canada
36 York Mills Road, Suite/Unité 505
Toronto, ON
+1 (613) 986-0350
jason.knabl@computecanada.ca

Jason Hlady
University of Saskatchewan
Saskatoon, SK
+1 (306) 966-2075
jason.hlady@usask.ca

Rachana Ananthakrishnan
Globus Computation Institute
University of Chicago
ranantha@uchicago.edu

Kyle Chard
Globus Computation Institute
University of Chicago
chard@uchicago.edu

Jim Pruyne
Globus Computation Institute
University of Chicago
pruyne@uchicago.edu

ABSTRACT

This poster reports on the outcomes of and lessons learned from a pilot project to test core components of a national research data management infrastructure service. A software stack consisting of Archivematica and Globus Publishing was used to pass datasets from an established domain repository through an archival processing pipeline and establish discovery and access layers from the output.

General Terms

Infrastructure opportunities and challenges; Preservation strategies and workflows; Innovative practice.

Keywords

Research data, Preservation, Access, Archivematica, Globus Publishing.

1. INTRODUCTION

The number of data repositories providing access to research data is growing at a rapid rate around the globe. Developments in data access, however, have outpaced advances in the digital preservation of research data, even though long-term access is dependent on properly archived content. An important reason for this has been the wide variety of research data, its volume, and the speed at which it is produced. Finding technologies to disseminate such data tends to be easier than establishing sound ways of producing archival copies of complex datasets. Key to addressing this challenge is building software that scales to the processing demands of diverse research data collections.

As an initial investigation into this challenge in preserving research data, Research Data Canada[1] (RDC) established a Federated Data Management Pilot Project[2] to build core components of a national research data management infrastructure service. The design of the pilot project involved taking datasets from an established domain repository, passing them through an archival processing pipeline, and then establishing discovery and access layers from the archival output.

[1] http://www.rdc-drc.ca/

[2] http://www.rdc-drc.ca/activities/federated-pilot/

2. PILOT CONFIGURATION

The Canadian Polar Data Network[3] (CPDN) provided its diverse collection of research data from the Canadian International Polar Year (IPY) for use in this pilot, as well as the time and expertise of staff from CPDN partner members. A software stack consisting of Archivematica[4] for archival processing and Globus Publishing[5] for the discovery and access platform was hosted by Compute Canada[6] (CC), which also contributed personnel. The pilot project's objective was to evaluate this specific configuration to understand better the requirements for a national preservation, discovery, and access platform.

Archivematica processed each dataset selected for this pilot as a Submission Information Package (SIP) to generate Archival Information Packages (AIPs) and Dissemination Information Packages (DIPs). All DIPs were moved to the Globus Publication platform for access and discovery.

The implementation challenge with Globus Publishing was to find a flexible batch process to ingest metadata and data files from an existing collection rather than from individual research projects. This required entering metadata in batch rather than inputting metadata manually and ingesting data in bulk instead of submitting data through individual projects. Transformation of existing metadata to conform to Globus Publishing's metadata model was another key step. Aspects of this project built upon the experiences of an earlier project at Simon Fraser University by extending the deposit functionality of Globus Publishing.

3. CONCLUSIONS

This pilot provides important insights into the requirements for implementing a production service based on the functions of this test. First, it demonstrated that automated processes could generate archival digital objects for research datasets and that these objects could be deposited with an access platform (Globus Publishing in this instance) and archived in preservation storage. Second, it demonstrated that, once ingested into a discovery and access platform, datasets were discoverable and retrievable under appropriate controlled access conditions. Third, it identified a need for upfront preparation of metadata by a metadata expert and for the intervention of a data curator to start and monitor the processing cycle. Fourth, it identified several improvements that will be necessary to assemble a small-scale production system based on this pilot's basic design.

All of the suggested improvements are incremental in nature and achievable through a next-step development process. In the pilot a separate workflow was used to transmit metadata to Globus Publishing; this step needs to be better integrated into the Archivematica pipeline. Developments in computational processing that enhance scalability are needed for pushing large digital objects through the pipeline. There is a clear need for Archivematica to better manage the processing of dataset-level metadata for discovery applications outside of Archivematica. Finally, the use of Archivematica's Format Policy Registry needs to be incorporated into the design to support normalization processing of the diverse file formats encountered in research datasets.

Overall, the pilot helped us to understand better the steps required to prepare research data for access and preservation and to anticipate what a successful national preservation, discovery, and access platform for research data might look like.

[3] http://polardatanetwork.ca/

[4] https://www.archivematica.org/

[5] https://www.globus.org/

[6] https://www.computecanada.ca/