

Theses Harvesting at Library and Archives Canada

Access Conference 2019

Edmonton, Alberta

September 30, 2019

Library and Archives Canada



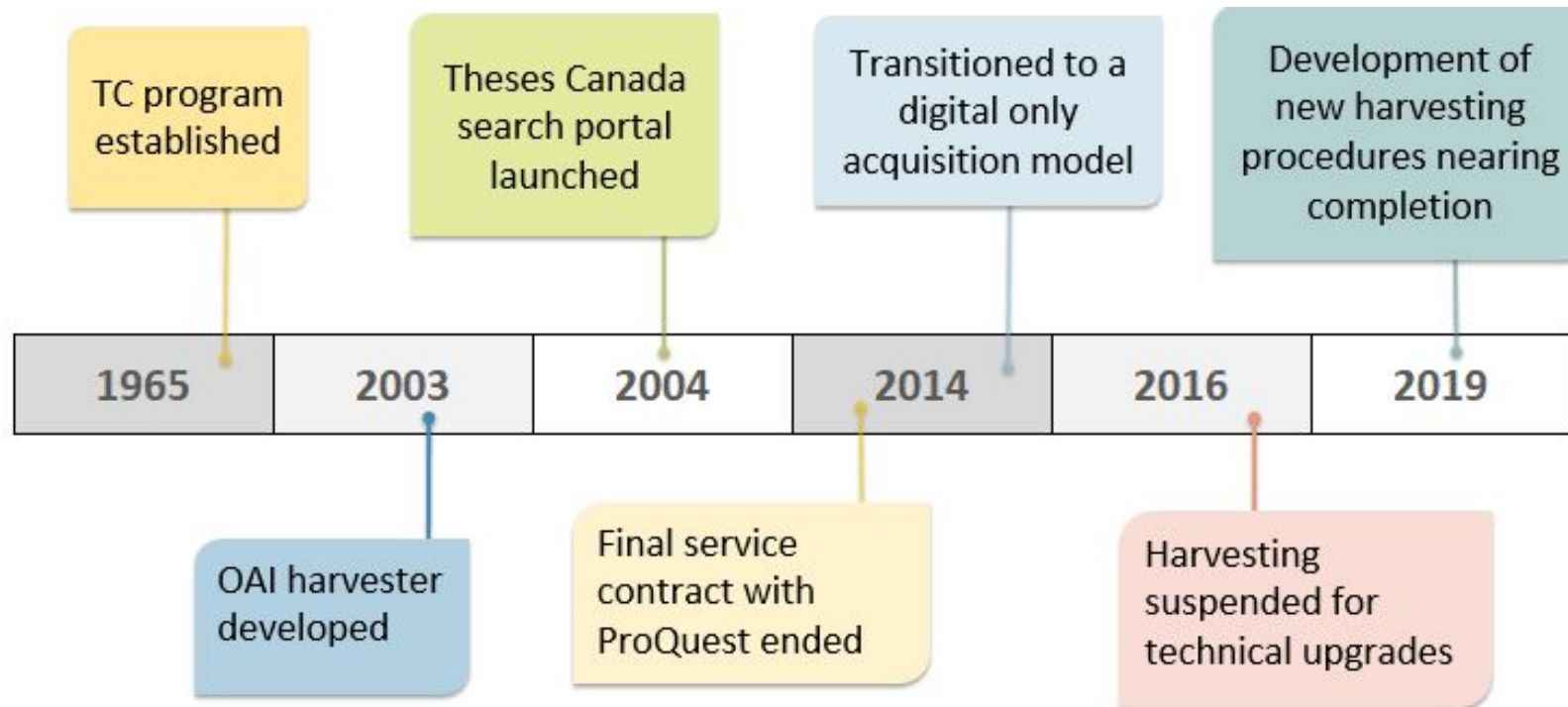
Library and Archives
Canada

Bibliothèque et Archives
Canada

Canada

Theses Canada (TC) program

Cooperative program between Library and Archives Canada (LAC) and over 50 Canadian universities



New systems

LAC has moved away from custom built systems and acquired:

- GoAnywhere Managed File Transfer:
 - harvester and workflow automation
- Preservica: active digital preservation
 - multiple copies, multiple locations
- OCLC WorldShare Management Services
 - shared library services platform



GO ANYWHERE®
Managed File Transfer



Preservica
Active digital preservation



OCLC®

Harvesting from institutional repositories

It all starts with an Excel sheet:

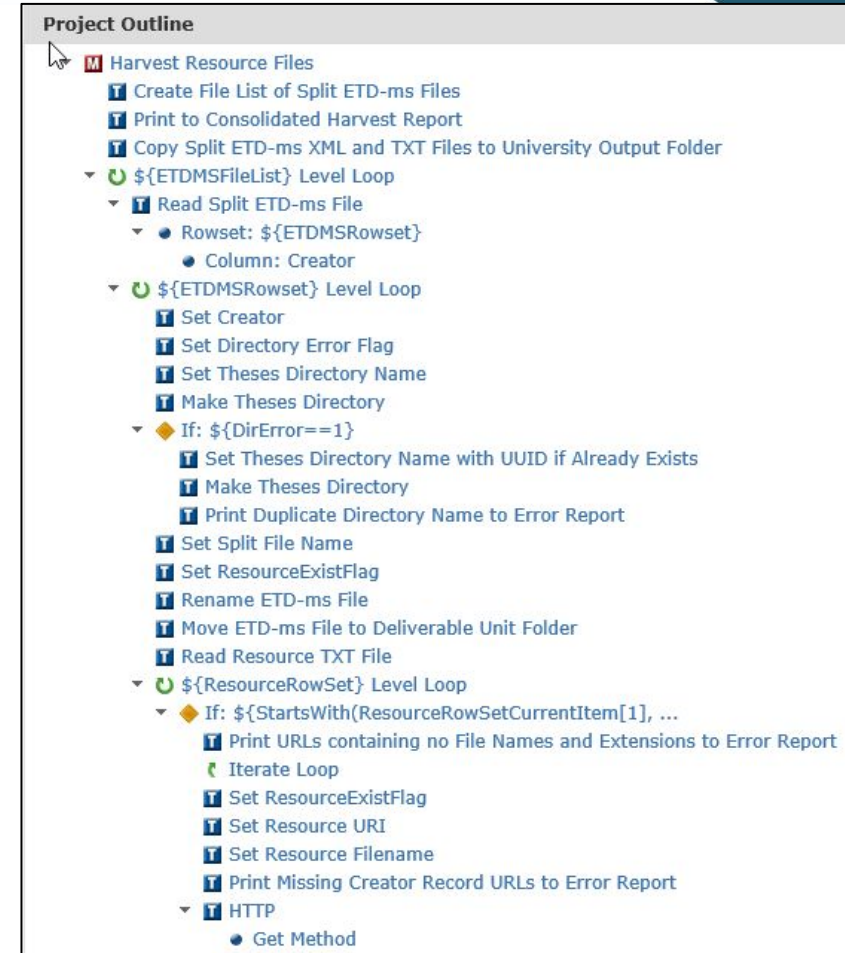
2	Carleton University	HTTPS	curve.carleton.ca	/oai-pmh/lac-theses/request	No_Set	oai_etdms
3	Concordia University	HTTPS	spectrum.library.concordia.ca	/cgi/oai2/request	oathesis	oai_etdms
4	Dalhousie University	HTTP	dalspace.library.dal.ca	/oai/request	com_10222_10559	oai_etdms
5	École nationale d'administration public	HTTP	espace.enap.ca	/cgi/oai2-etdms/request	74797065733D746865736973	oai_etdms
6	École Polytechnique de Montréal	HTTPS	publications.polymtl.ca	/cgi/oai2/request	bac-tc	oai_etdms
7	McGill University	HTTP	digitool.library.mcgill.ca	/OAI-PUB/request	eTheses	oai_etdms

We use the harvester in GoAnywhere:

Variable Name	Description	Original Value	New Value
FromDate	Not Specified	2018-08-01	2018-08-01
ToDate	Not Specified	2018-12-31	2018-12-31
ControlFileName	Not Specified	F:\Data\etheses\University List.xlsx	F:\Data\etheses\University List.xlsx
ControlFileStartRow	Not Specified	6	6
ControlFileEndRow	Not Specified	6	6

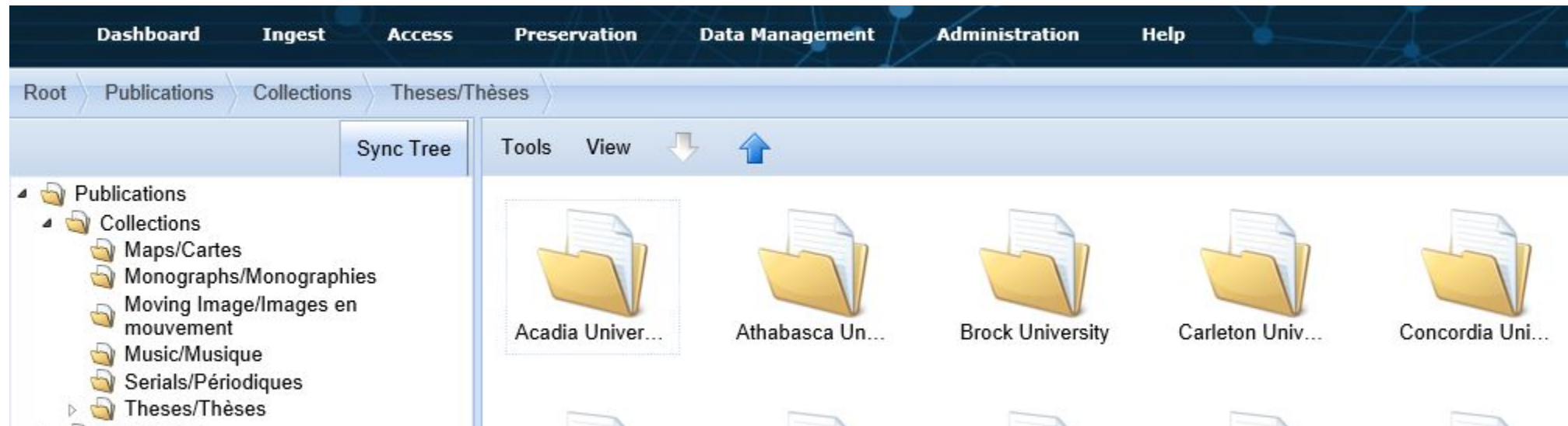
GoAnywhere workflow

- 14 modules to perform tasks including:
 - Harvest and split ETD-MS metadata
 - Download resource files (PDFs and other)
 - Transform ETD-MS into MODS
 - Arrange metadata and files into folders by author name
 - Identify errors and create reports (for each university plus a consolidated report)
 - produce a SIP (submission information package) for export to Preservica



Preservica

- GoAnywhere automatically triggers the Preservica ingest workflow and the folders are transferred to a review area
- Staff perform quality checks for each university then move the folders to a permanent collection area



Public access



The screenshot shows the Library and Archives Canada website. The header includes the logo and navigation tabs: 'Discover the Collection', 'Search the Collection', and 'Services for the Public'. A breadcrumb trail reads: Home → Services and Programs → Theses Canada → Search - Theses Canada. The main content area is titled 'Search - Theses Canada'. On the left, there is a sidebar with links: 'Theses Canada', 'Obtain a Thesis or Dissertation', 'Search' (highlighted), and 'Search Help'. The main search area contains two input fields: 'Any keyword' and 'Author', both with empty text boxes.

- MODS to MARC workflow to transform and send metadata to the LAC library catalogue, Aurora, is under development
- The metadata will include a link to the files in Preservica
- A separate theses search portal is updated nightly with data from the library catalogue
- Theses are also available in Aurora and in the union catalogue, Voila

Challenges

- Learning-by-doing as we tested and developed the workflow
- Dealing with variety of repositories and metadata, each with a different issue such as missing fields, no URL to the file, quirks in formatting the date parameters, expired security certificates, etc.
- What can universities fix, what can we fix? Trying to gauge when we should modify our workflow and when we should ask the universities to change something
- Creating quality MARC records from the MODS / ETD-MS metadata with no human intervention

Looking forward

- LAC remains strongly committed to the program
- Harvesting will begin soon with more universities than before
- We plan to harvest (or reharvest) the entire theses repository of each university (with approval) into Preservica
 - This way we can acquire supplementary files for older theses
- Improved preservation and long-term access ensured
- Flexible systems that we can efficiently modify and develop
- Integrated with LAC's main digital acquisitions stream

Thank you!

Arlene Whetter

arlene.whetter@canada.ca

Theses Canada

BAC.ThesesCanada-ThesesCanada.LAC@canada.ca

Library and Archives Canada

www.bac-lac.gc.ca

Telephone: 613-996-5115 or 1-866-578-7777 | TTY: 613-992-6969 or 1-866-299-1699

National
Capital Region

Vancouver
British Columbia

Winnipeg
Manitoba

Halifax
Nova Scotia



Library and Archives
Canada

Bibliothèque et Archives
Canada

Canada