Canadian Research Knowledge Network
Réseau canadien de documentation pour la recherche

Héritage Collection Update: RG-10 ICR Pilot Project

Jason Friedman, Senior Manager, Heritage Services

crkn-rcdr.ca

1

Canadiana Collections: Content

Canadiana

Digitized monographs, serials, government publications, and maps

Héritage

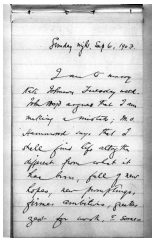
Digitized microfilm of archival primary-source materials from Library and Archives Canada

Canadiana

2

Héritage collection

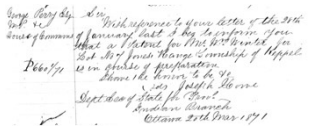
- Partnership with Library and Archives Canada (LAC)
- "C", "H" and "T" series of LAC microfilm
- 42+ million pages; 25,000+ reels, 900 collections
- Still in progress: 150,000 pages added annually



3

Discovery challenges

- Digitized microfilm
  - Limited finding aids/indexes
  - Limited metadata
  - Limited search capabilities
  - Lots of Handwriting means no OCR, no full text search



CRKN RCDR

4

ICR Pilot Project

5

Intelligent Character Recognition (ICR)

- Uses AI to perform character recognition
- Can be used on handwritten materials
- Environmental scan conducted in Fall 2022

6

Transkribus

- Produced by READ-COOP
- Used by researchers/historians working with parts of the Héritage collection
- Initial sample sent for testing
  - 1-2% error rate for typewritten materials
  - 5-7% error rate for handwritten materials
- Next step: pilot project

7

ICR Pilot Project

- Partnership with LAC
- Full scope: 2 million images
- Began in March 2023
- Initial batch of 600K+
- Additional 500K identified
- Approximately 1 million more images to be identified

8

Why RG-10?

- High priority collection
- Heavily used
- Lots of handwritten content
- Support claims researchers
- Variety of document types/quality

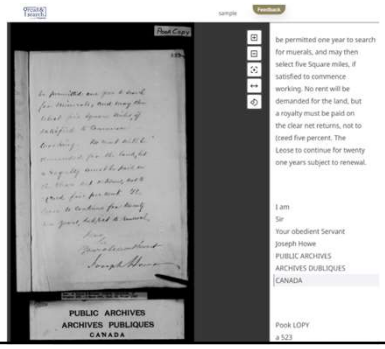
9

First Batch: Departmental Letterbooks

- All are handwritten with relatively consistent writing and format
- Covers a breadth of content of interest for variety of research topics
- They could benefit from ICR as they are poorly described
- Time period, 1871-1929 (low risk of sensitive content)
- Existing transcription data for some content from University of British Columbia/University of Saskatchewan for model training

10

Initial Results: “Out of the Box”



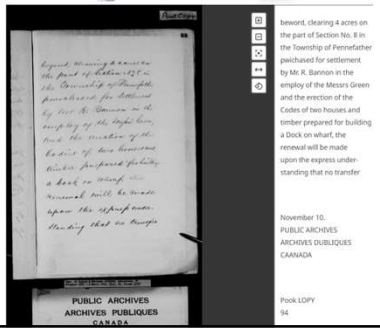
be permitted one year to search for mueals, and may then select five Square miles, if satisfied to commence working. No rent will be demanded for the land, but a royalty must be paid on the clear net returns, not to exeed five percent. The Leose to continue for twenty one years subject to renewal.

I am
Sir
Your obedient Servant
Joseph Howe
PUBLIC ARCHIVES
ARCHIVES DUBLIQUES
CANADA

Pook LOPY
a 523

11

Initial Results: “Out of the Box”



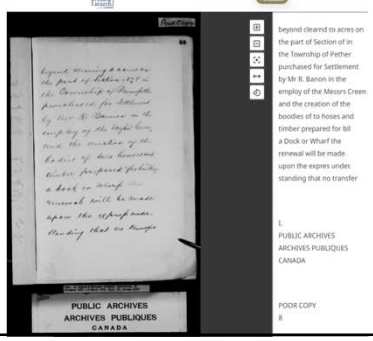
beword, clearing 4 acres on the part of Section No. 8 in the Township of Pennefather pwichased for settlement by Mr. R. Bannon in the employ of the Messrs Green and the erection of the Codes of two houses and timber prepared for building a Dock on wharf, the renewal will be made upon the express under-standing that no transfer

November 10.
PUBLIC ARCHIVES
ARCHIVES DUBLIQUES
CAANADA

Pook LOPY
94

12

Initial Results: “Out of the Box” (different model)



beyond clearnd to acres on the part of Section of in the Township of Pether purchased for Settlement by Mr R. Banon in the employ of the Messrs Green and the creation of the boodies of to hoses and timber prepared for bil a Dock or Wharf the renewal will be made upon the expires under standing that no transfer

I.
PUBLIC ARCHIVES
ARCHIVES PUBLIQUES
CANADA

POOR COPY
R

13

ICR Pilot Project: Next Steps

Training	Complete initial model training for Letterbooks
Receive	Receive initial results for verification/quality control
Review	In partnership with LAC, perform block review
Load	Load ALTO XML/PDFs onto Héritage platform
Transmit	Transmit additional image files to READ-COOP

14

How you can help

We still have approximately 1 million more images for the pilot, so suggestions are welcome!

Identified series include:

- Annuity paylists
- Census records
- Black series
- Thousand series
- Red series

And if you have transcriptions, please share.

Email: jfriedman@crkn.ca

15

Beyond ICR

16

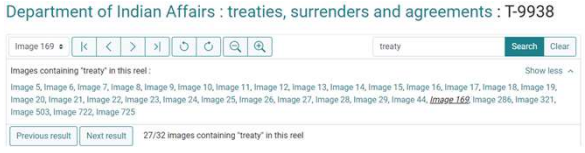
PDiiif: Custom PDF downloads from Héritage

- https://pdiiif.jbaiter.de/
- Instructions: https://www.crkn-rcdr.ca/en/navigating-collections
- Images only, no OCR

CRKN RCDR

17

Coming soon… improved search results display

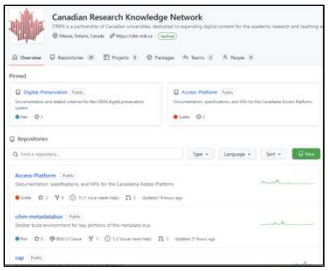


18

Future Enhancements

19

What is the Canadiana Platform today?



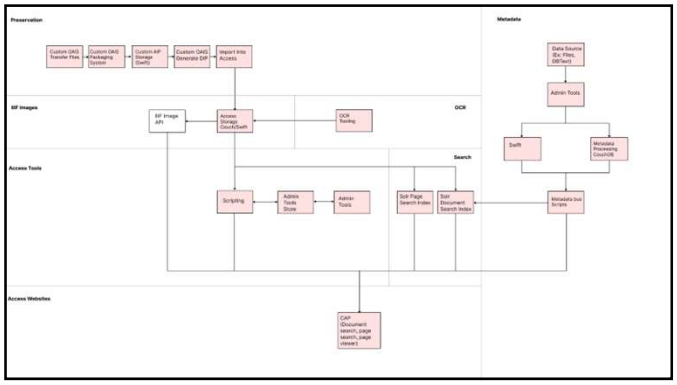
- Centralized web platform
- Monolithic slab of custom code
- Maintained by small team
- Technically open source code, available on GitHub

20

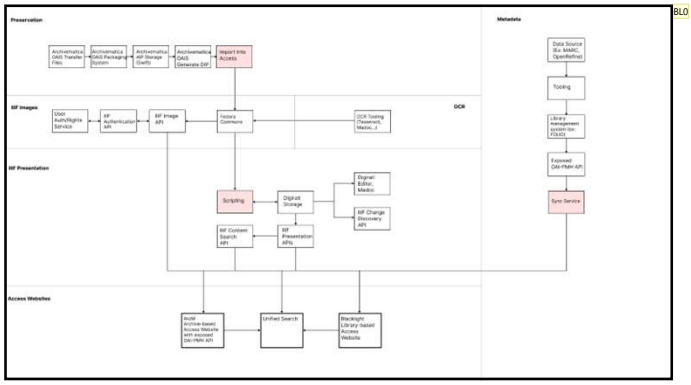
Replacing custom code with third-party solutions

- Archivematica for preservation
- Blacklight for improved search and discovery
- Digirati for collection management
- FOLIO for better description

21


Preservation
Metadata
IIIF Images
OCR
Access Tools
Search
Access Websites

22


Preservation
Metadata
IIIF Images
OCR
IIIF Presentation
Access Websites

23


Canadian Research Knowledge Network
Réseau canadien de documentation pour la recherche
Thank you! Questions?
jfriedman@crkn.ca
crkn-rcdr.ca

24

Slide 23

BL0 Russell said he would not be surprised if we could find a non-custom solution for the unified search at that point.
Brittny Lapierre, 2023-03-07T17:13:47.908