# COPPUL Digital Stewardship Network Community of Practice Mixer

May 7, 2025 @ 12pm Pacific / 1pm Mountain / 2pm Central

Hosted by the CDSN Community of Practice

## Quick links:

Sign up for the mailing list

## Suggested topics for future mixers:

- [ADD YOUR SUGGESTIONS HERE]

## AGENDA

- Welcome and introduction (Andréa Tarnawsky)
  - Code of Conduct
- Presentation (Carla Graebner and Nicholas Worby)
  - Canadian Government Information Digital Preservation Network
- Group discussion period

## UPCOMING MIXERS

- June & August 2025 – Date TBD

## Community notes

*If you aren't actively typing, please park your cursor here!*

- Presentation: A brief overview of the CGI-DPN
    - Started October 2012, with 11 institutions coming together
    - Goal to preserve collections of government information
        - Government of Canada publications catalogue
        - Pre-election Government of Canada and candidate websites
        - Provincial, thematic, and event specific collections
    - Options and considerations as they were developing the network
        - Internet Archive (Archive-It)
            - Could implement quickly, would not need to develop new plug-ins for LOCKSS network
            - Would require a consortial account to access Archive-It, would also require training on how to use
        - Access portal to point to LOCKSS (Lots of Copies Keep Stuff Safe)
            - LOCKSS is for storage, not really meant to provide access
        - Also looked at options for local hosting
    - In the end, went with Internet Archive and worked with COPPUL for the licensing, with CGI-DPN participants sharing costs
    - Currently the CGI-DPN has over 57 million documents archived
    - Web archiving
        - Not just preserving the content exactly as it is. If you've used the Wayback Machine, you know there are limitations on what we can capture. Work isn't fully automated either, you have to make decisions about how deep to go, which pages to capture, etc.
    - WARC files
        - ISO standard, basically a really complex zip file, written in such a way that it can be played back. Also includes metadata about how it was collected,

etc. lots of provenance metadata in there. Information to support de-duplication so you don’t crawl the same page again. Lots of information contained in WARC files!

- Different ways to interact with a WARC file
  - Text, web archive player like the Wayback Machine, etc.
- The web crawler can only capture what it can find, so we love publications! If the info lives in a database, we can’t get to that. Another issue is with websites changing over time, which can impact original order and context.
- Preservation challenges
  - Large files (4.6 TB in their collection), difficulty with local hosting
  - URLs not having a one-to-one relationship with a WARC file, it could be referencing other content
  - Playback reliant on a handful of tools that take a lot of work to get running, need access to a server, IT support, etc. which was part of why using the Internet Archive infrastructure was so appealing
  - Might need to do things like browser emulation for obsolete formats like Adobe Flash
- Additional challenges
  - Technology gets old – LOCKSS was the most appropriate at the time, but it might not be now / in future
  - Archive-It went down for 3 days during a crawl
  - Robots.txt protocol no longer working as well as it can, due to AI harvesters many websites are saying they don’t want to be crawled by AI so they are just blocking everything
  - Capacity

      - For many participants in the CGI-DPN, this is not their core responsibility, though members are advocating for the importance of this work
    - Return on investment
      - Generating an annual report on the activities of the CGI-DPN
  - Benefits of participating in the CGI-DPN
    - Collection development, preserving important resources
    - Distributed, collaborative work, able to learn skills from colleagues
  - Current and upcoming CGI-DPN activities
    - Crawls starting over the summer
    - Developing a risk management strategy
    - Recruitment and outreach to communicate the value of what we’re doing
  - Importance of building community
- Demo of backend Archive-It infrastructure, internal workflows and documentation
  - Data budget
  - Capturing Government of Canada info prior to the election, as websites are often updated with new information as a result
  - Also capturing local media coverage of the election, as well as candidate websites
  - 392 GB captured since March
  - Seeds are essentially URLs used to crawl
    - CGI-DPN members went through the Government of Canada website to identify seeds, use a Google Sheet to capture info on who has done what
    - Also have a QA process, going back to look for errors or missed information. Media-heavy websites (i.e. lots of videos) can make it hard to capture

    - Adding metadata to the content is a big deal
    - Original URL and Wayback URL
    - Try to capture as much info as possible, clicking around to the various links on websites
    - Using open source technology
        - Heritrix
        - Brozzler: Better for higher fidelity captures, scrolling, and opening menus, it's slower and acts more like a browser to avoid being blocked (which can be a challenge to deal with as it's often difficult to get in touch with folks who run the websites to restore access if you do get blocked)
    - Lots of back and forth involved to try and capture as much as possible
- Resources
    - How to's
        - Awesome web archiving
        - IIPC training materials
        - Web archiving at UBC Library
        - Archive-It video curriculum
    - Tools
        - Archive-It
        - Conifer
        - Browsertrix

Questions:

- Q: How does the CGI-DPN come together and make decisions? (In the context of elections, changes in government and resulting activities etc.)
    - A: Group meets monthly, sometimes more frequently to work on specific projects. The CGI-DPN has a pre-election toolkit, and action plan identifying certain tasks depending on when the election would be called. Because we had these protocols in place we were able to get things done. Outside of election years it can be a bit more mundane, might meet every two months, with a strategic planning session in spring to discuss what to do in the upcoming year.
    - A: Things were pretty intense in the Harper years, as back then LAC wasn't capturing everything and there was pressure for us to capture info, but we didn't necessarily know what LAC was doing, what their processing workflows were like, etc. which made coordinating activities a challenge
- Q: How do you define your thematic collections?
    - A: Some of them are institutionally focused, but are grouped together with ours. I haven't been so involved in that.
    - A: It can be hard to define scope, at UofT we've started requiring collections proposals with information. Haven't been doing that with CGI-DPN, but it might be nice to start preserving that information in provenance documents, as folks have come and gone from the CGI-DPN itself and sometimes that knowledge can become lost.
    - A: Just saw on the GOV DOC L list, the University of Minnesota has created a list of websites and a formalized process that we might be able to look at / adapt for our situation. We are always looking for people to join the CGI-DPN, you no longer need to host a LOCKSS box, can provide in-kind support instead. Please reach out to Carla or Nich if you're interested in joining!

- Q: Is there still work to be done for this most recent crawl?
    - A: Still a lot of QA to be done, comparing the crawl site to the live site.
    - A: You can view post-crawl reports and see if there's anything you've missed.
    - A: Another issue we've encountered is lots of candidates using social media, particularly Facebook to interact with their constituents. But there are ethical concerns about capturing that, when constituents start to interact with candidates on social media as they haven't consented to the captures.
    - A: Concerned about capturing social media for those reasons, the right to be forgotten, etc. Worry that captures of social media can be used by bad actors to target people. It's tricky if there are comments on the page, etc. and that was something that was discussed in the CGI-DPN.
- See discussion on the CWAC list re: WARC backups as a potential area of future interest