Shared services for digital preservation

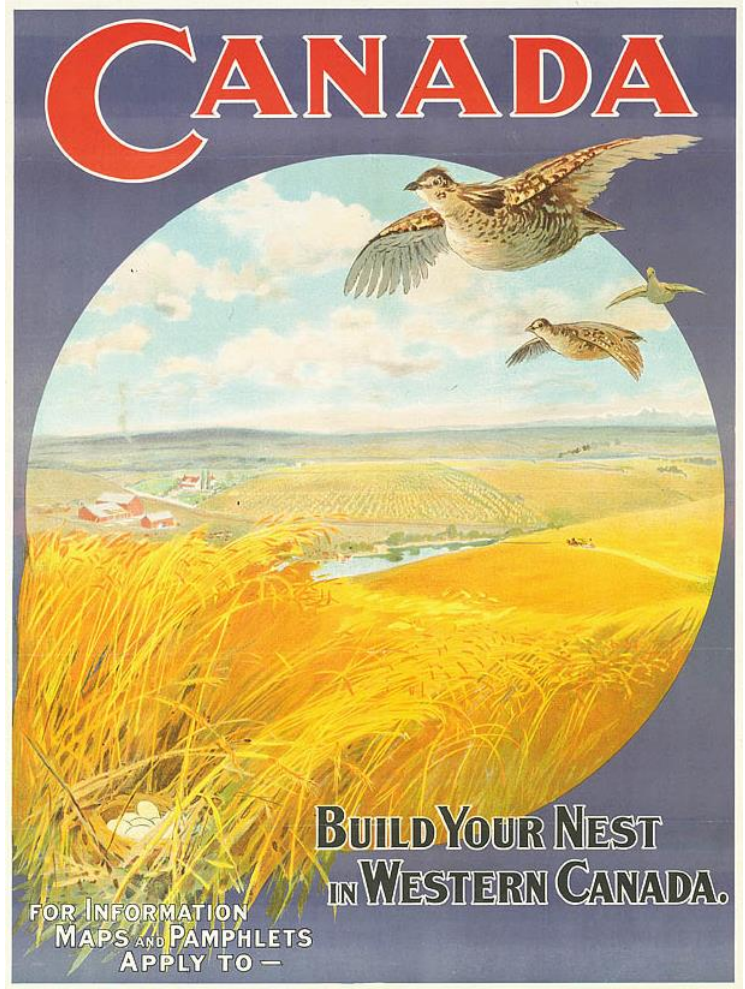
A library perspective

Corey Davis
Digital Preservation Network Manager, Council of Prairie and Pacific University Libraries (COPPUL)
Systems Librarian, University of Victoria

BCNET 2017, Vancouver, BC
Wednesday, April 26, 4:00pm-5:00pm

Council of Prairie and Pacific University Libraries

- Represents 22 university libraries in Western Canada
- "…leverage collective expertise, resources, and influence, increasing capacity and infrastructure, to enhance learning, teaching, student experiences, and research at our institutions."
- Digital preservation is one of three key strategic priorities moving forward



Digital dark age?

- “Much of our global cultural heritage, and our own individual and social imprint, is at serious risk of disappearing.”
    - Richard S. Whitt, Corporate Director for Strategic Initiatives at Google

Benign neglect does not work!

- “In the past, a do nothing approach, frequently characterised as ‘benign neglect’, has often been enough to ensure the survival and usability of physical data. However, computer systems and software applications change so rapidly there is no guarantee that existing data sources will be accessible and useable on future computing platforms or software versions.”

  - Andrew Wilson, Significant Properties Report: inSPECT Work Package 2.2, 2007.

Running to stand still…

- “Traditionally, preserving things meant keeping them unchanged; however our digital environment has fundamentally changed our concept of preservation requirements. **If we hold on to digital information without modifications, accessing the information will become increasingly difficult, if not impossible**”

  - Su-Sing Chen, “The Paradox of Digital Preservation”, Computer, March 2001, 2-6

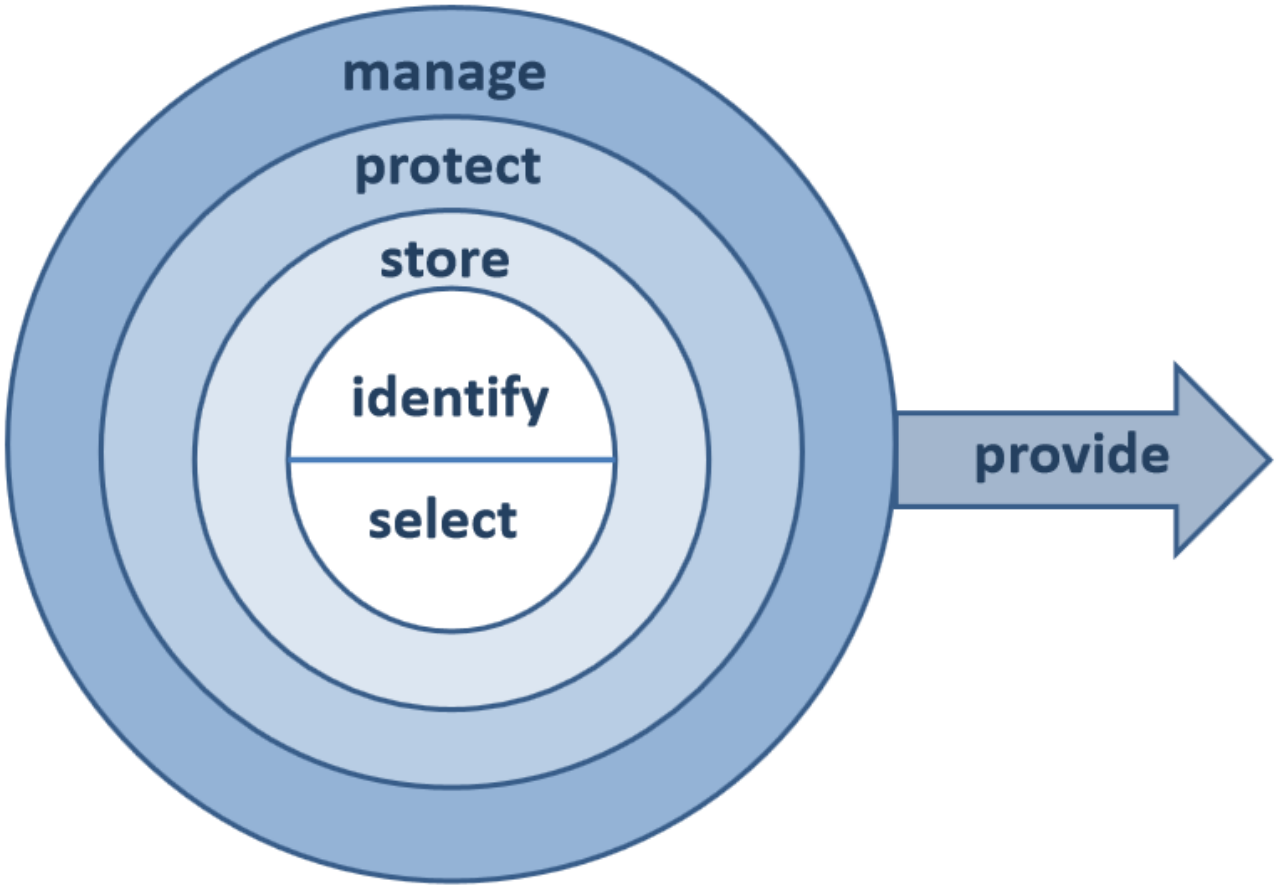
What is digital preservation?

- "…a **formal endeavor** to ensure that digital information of continuing value remains accessible and usable."
  - Digital Preservation Coalition

- "Digital preservation is the **active management** of digital content over time to ensure ongoing access. "
  - Library of Congress

Digital preservation consists of…

- Identification
- Selection
- Storing
- Protecting
- Managing
- Providing access



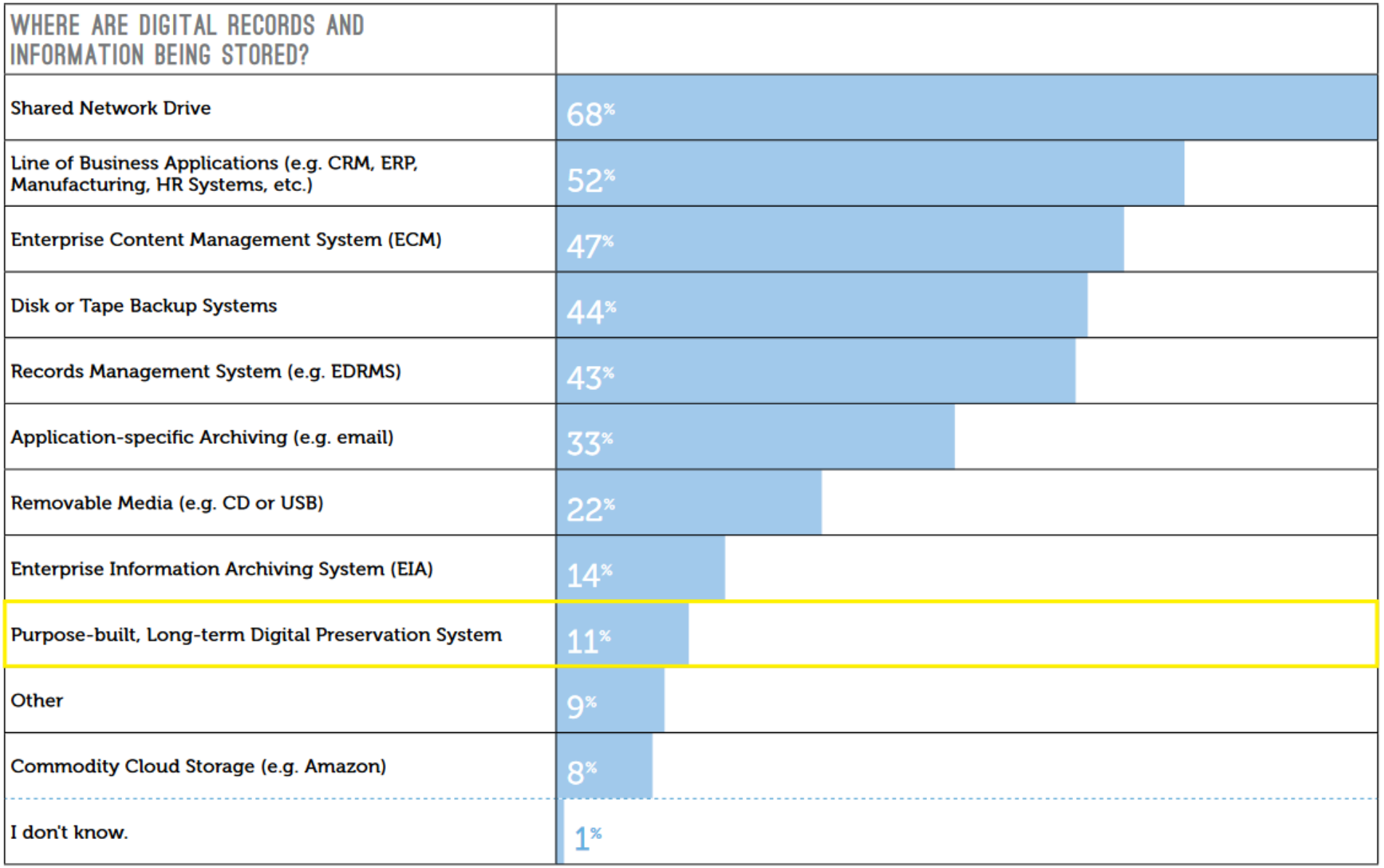


Chart from: The Governance of Long-Term Digital Information, IGI 2016 Benchmark. http://preservica.com/wp-content/uploads/sites/3/2016/05/The-Governance-of-Long-Term-Digital-Information-IGI-Benchmark-2016.pdf

- “Archivists, historians, and librarians—among many others—have been sounding the alarm about an impending ‘digital dark age’ and taking action to protect their digital information for decades.”

  - Barclay T. Blair, Executive Director and Founder, Information Governance Initiative

COPPUL Digital Preservation Network

- Expertise and resources
- Educational and professional development opportunities
- Advocacy, both local, regional, and national
- Collaboration with external partners
- Align COPPUL's efforts with emerging national and international standards, services, and infrastructure
- Training, technical support, and advice in relation to platforms and services

Today’s tour will include…

- Web archiving
- Preservation processing services
- Preservation storage

Web archiving

A case study in the use of Archive-It

Why archive the web?

- “The web, as it appears at any one moment, is a phantasmagoria. It’s not a place in any reliable sense of the word. It is not a repository. It is not a library. It is a constantly changing patchwork of perpetual nowness. You can't count on the web, okay? It’s unstable. You have to know this. ”
  - https://www.theatlantic.com/technology/archive/2015/10/raiders-of-the-lost-web/409210/?utm_source=twb

Ephemeral and drifting

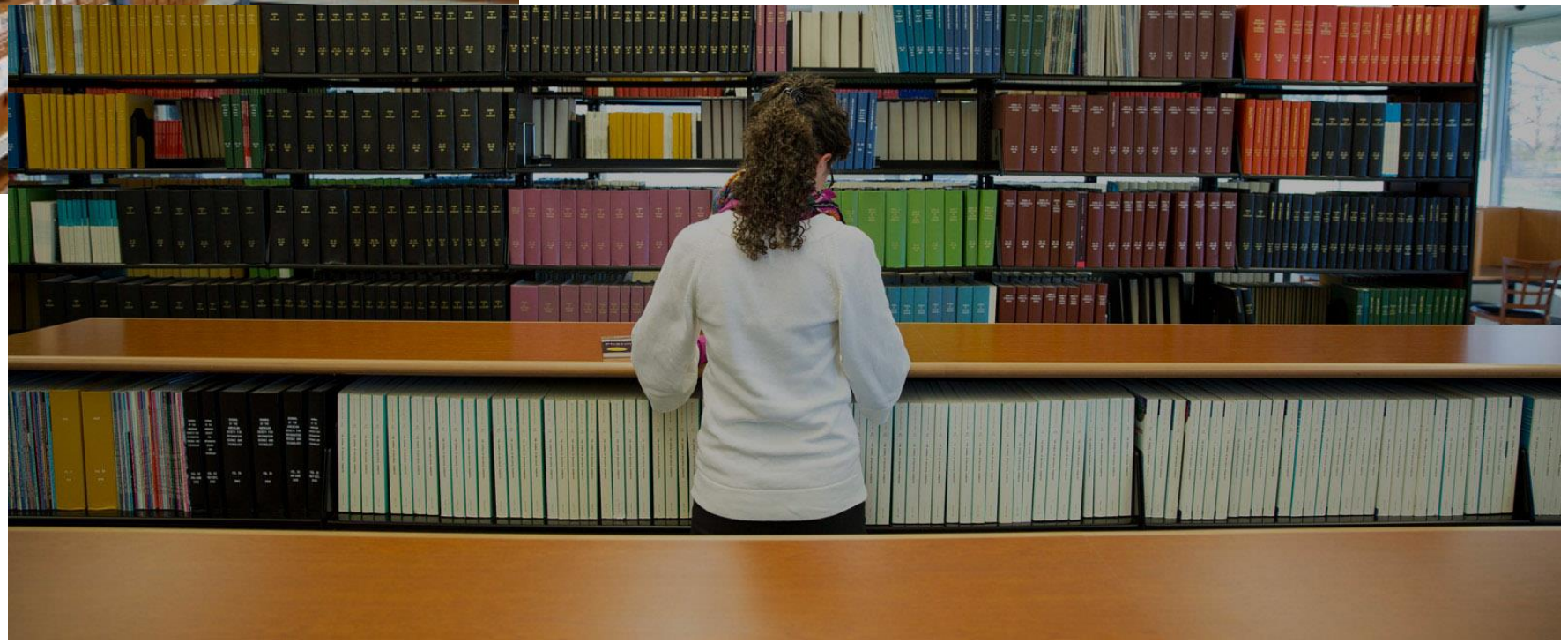
- “Fifty percent of URLS in the UK Web Archive have lost or missing content after one year. **After ten years nearly all content is moved, changed, missing or gone**.”
  - https://blogs.loc.gov/thesignal/2015/06/dodge-that-memory-hole-saving-digital-news/

- “We find that **for over 75% of references the content has drifted away from what it was when referenced**. These results raise significant concerns regarding the long term integrity of the web-based scholarly record and call for the deployment of techniques to combat these problems.”
  - http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0167475

Archive-It

- A hosted service of the Internet Archive
- Consortial license across 40+ institutions in Western Canada and Ontario
- Uses open source crawling and playback software to capture and provide access to snapshots of the web
- Produces standards-based WARC files

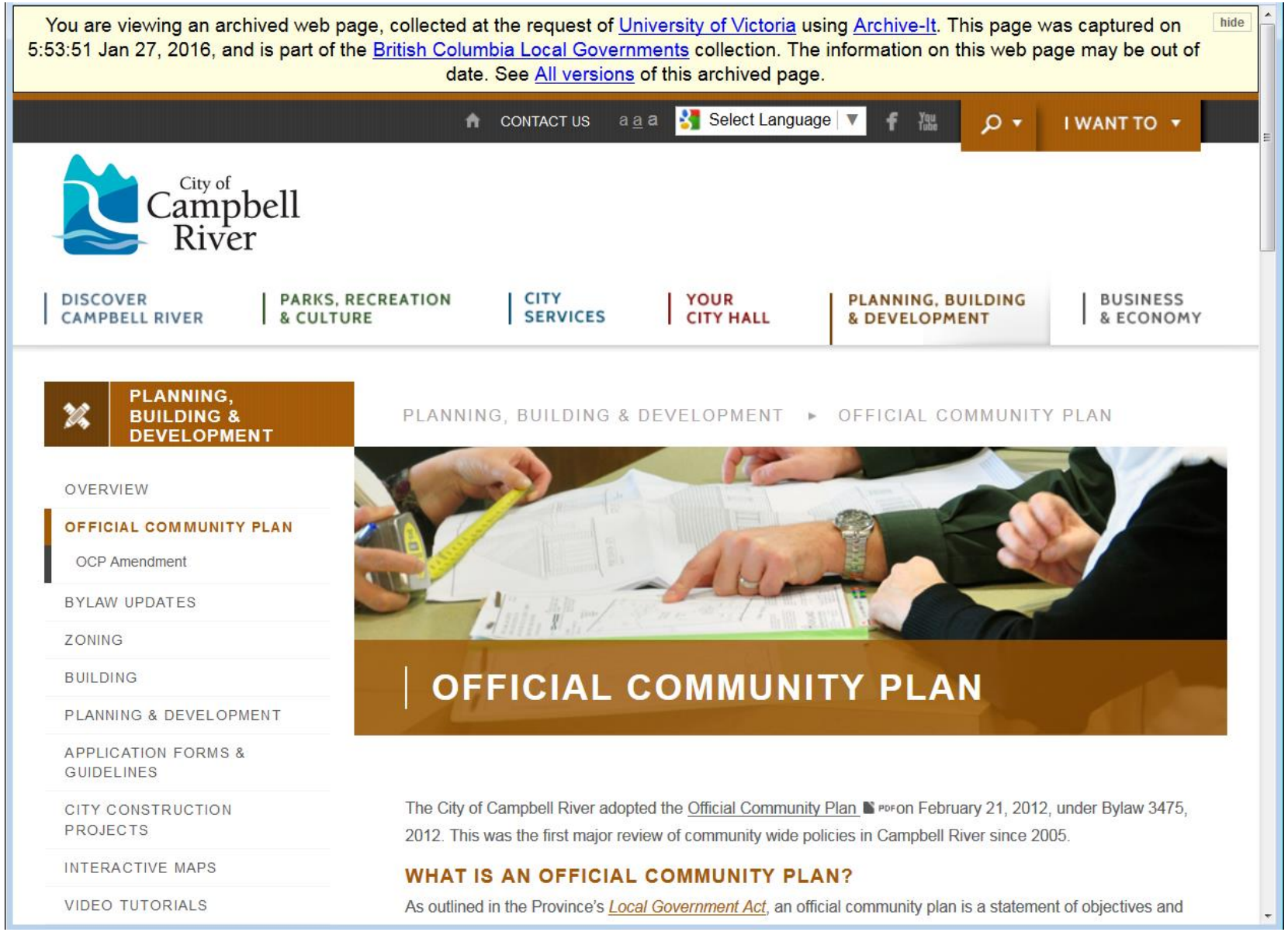
Corporate memory



Rough draft of history…

Access to government information

- "…facilitate democracy by helping to ensure that citizens have the information required to participate meaningfully in the democratic process and that politicians and bureaucrats remain accountable to the citizenry."
    - Mr. Justice La Forest, Supreme Court of Canada, 1997.



Before the web, the DSP

- “The Depository Services Program (DSP) was established in 1927 … to provide a comprehensive distribution source from which published Government of Canada (GC) information would be sent to academic, college, legislative and public libraries…”
  - http://www.publications.gc.ca/site/eng/programs/aboutDsp.html

NATIONAL POST

• FINANCIAL POST • NEWS • COMMENT • PERSONAL FINANCE • INVESTING • TECH • SPORTS • ARTS • LIFE • HEALTH •

NEWS CANADA POLITICS

CANADA

TRENDING Lotto Max | Donald Trump | Syria | Real estate | Vimy 100

Suspicions over library consolidation as critics complain Harper government is trashing important books

JEN GERSON | January 7, 2014 5:46 PM ET
More from Jen Gerson | @jengerson

Republish
Reprint

Wednesday December 14, 2016

Canadians help U.S. scientists protect climate change data

GUERRILLA ARCHIVING

Saving Environmental Data from Trump

Sear

Mor

Government of Canada | Gouvernement du Canada

Canada.ca | Services | Departments | Français

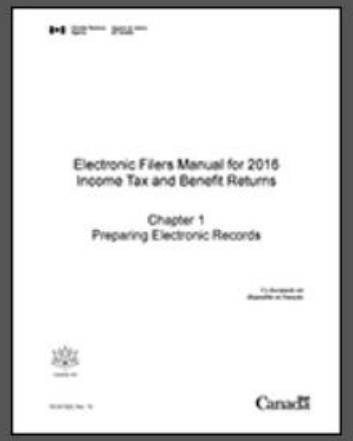
Government of Canada Publications

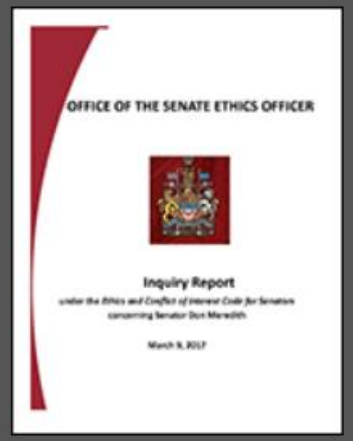
Canada

Search

Publications | Services | Resources

From Weekly Acquisitions List 17-11 (March 17, 2017)





NCBI Resources How To Sign in to NCBI

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed Search

Advanced Help

⚠ **Due to the lapse in government funding, PubMed is being maintained with minimal staffing. Information will be updated to the extent possible, and the agency will attempt to respond to urgent operational inquiries.**

Updates regarding government operating status and resumption of normal operations can be found at http://www.usa.gov.

PubMed

PubMed comprises more than 23 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

PubReader

A whole new way to read scientific literature at PubMed Central

Using PubMed

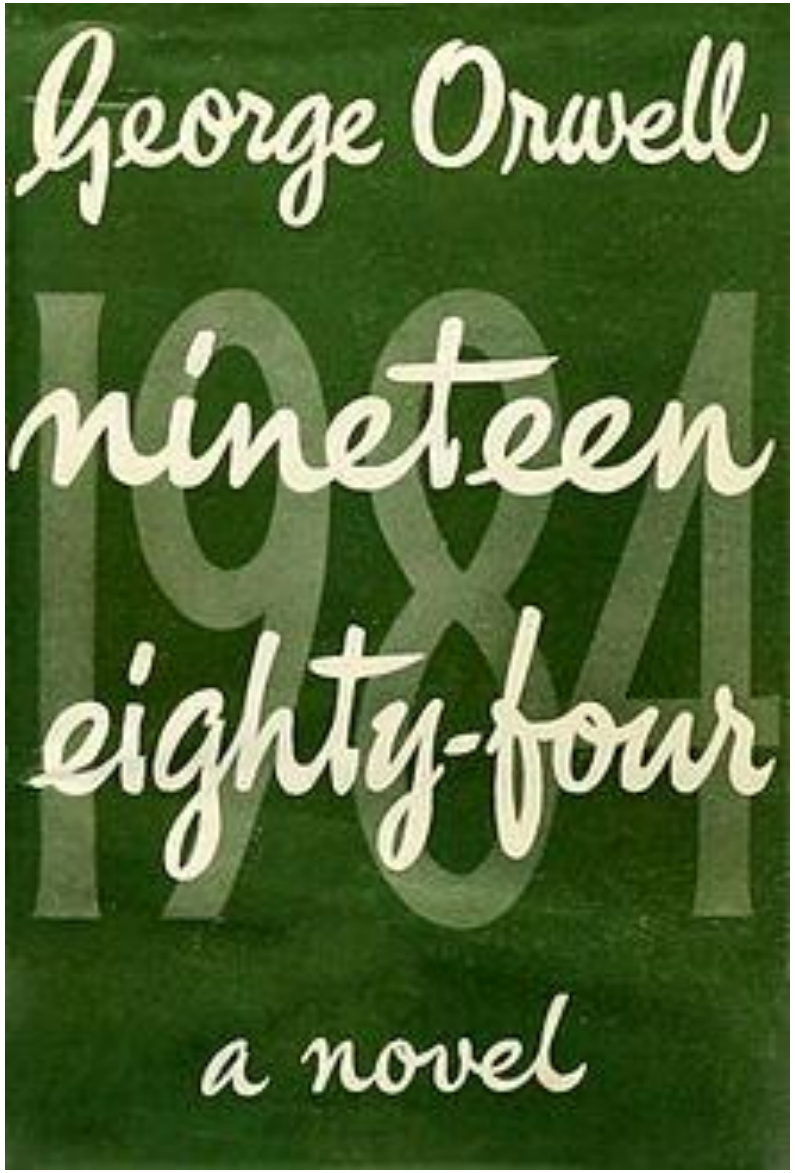
- PubMed Quick Start Guide
- Full Text Articles
- PubMed FAQs
- PubMed Tutorials
- New and Noteworthy

PubMed Tools

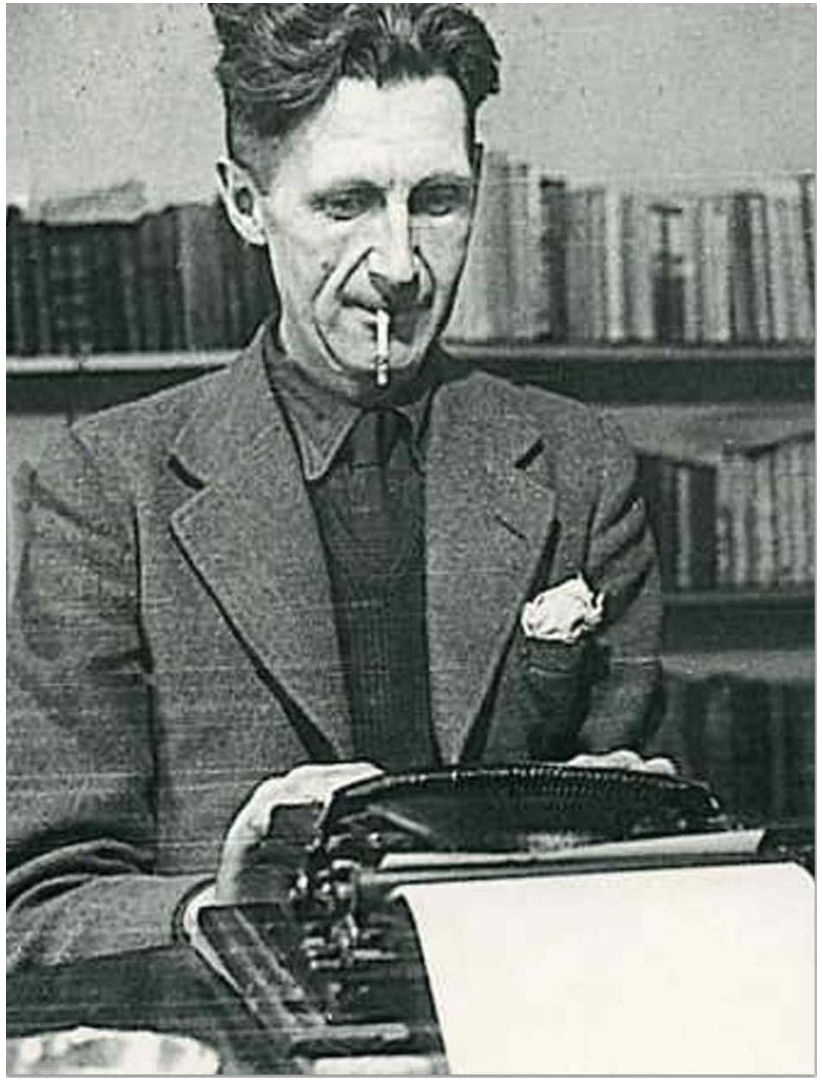
- PubMed Mobile
- Single Citation Matcher
- Batch Citation Matcher
- Clinical Queries
- Topic-Specific Queries

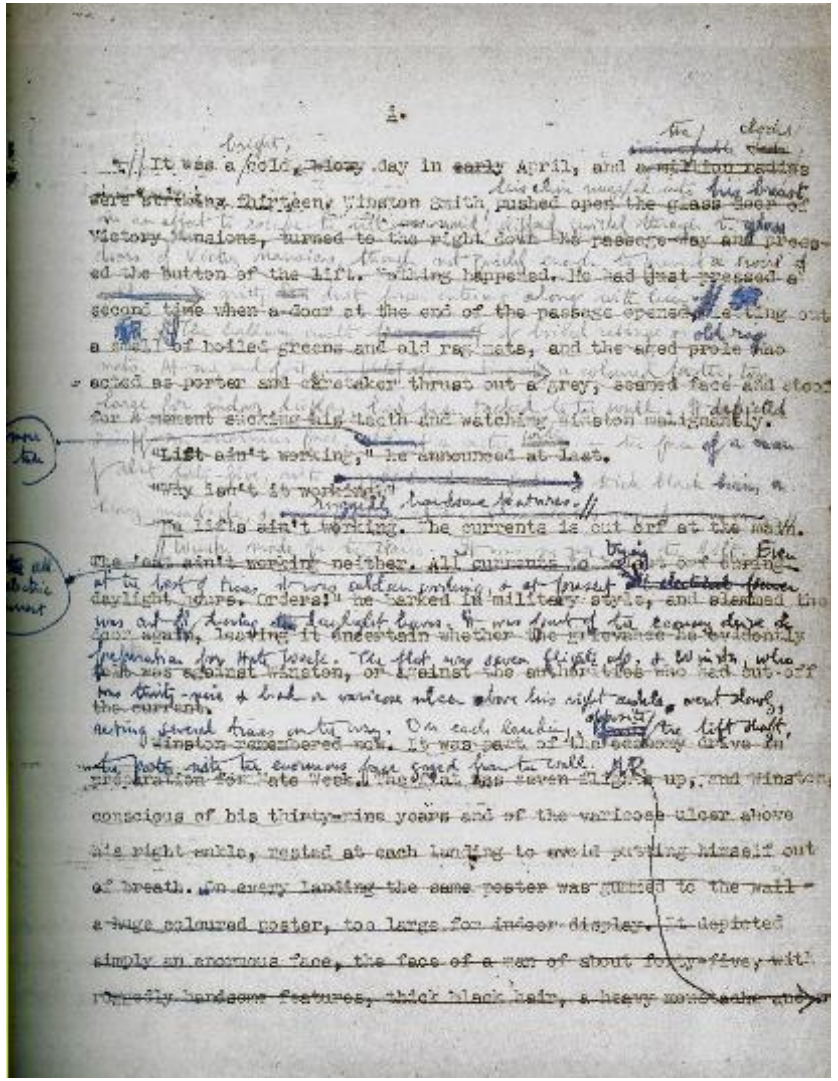
More Resources

- MeSH Database
- Journals in NCBI Databases
- Clinical Trials
- E-Utilities
- LinkOut



"The past was erased, the erasure was forgotten, the lie became the truth."





1.

It was a bright, cold, day in early April, and the clocks were striking thirteen. Winston Smith, his chin nuzzled into his breast in an effort to escape the wind, diffused with the stage of Victory Mansions, turned to the right down the passage-way and pressed the button of the lift. Nothing happened. He had just pressed a second time when a door at the end of the passage opened, letting out a smell of boiled greens and old rag mats, and the aged prole who acted as porter and caretaker thrust out a grey, seamed face and stood for a moment sucking his teeth and watching Winston malignantly.

"Lift ain't working," he announced at last.

"Why isn't it working?"

The lifts ain't working. The currents is cut off at the main. The lot ain't working neither. All currents to be cut off during daylight hours. Orders!" he barked in military style, and slammed the door again, leaving it uncertain whether the grievance he evidently felt was against Winston, or against the authorities who had cut off the current.

Winston remembered now. It was part of the economy drive in preparation for Hate Week. The flat was seven flights up, and Winston, conscious of his thirty-nine years and of the varicose ulcer above his right ankle, rested at each landing to avoid putting himself out of breath. On every landing the same poster was gummed to the wall - a huge coloured poster, too large for indoor display. It depicted simply an enormous face, the face of a man of about forty-five, with roggedly handsome features, thick black hair, a heavy moustache and

iBook

You
Tube

Mark J. Perry

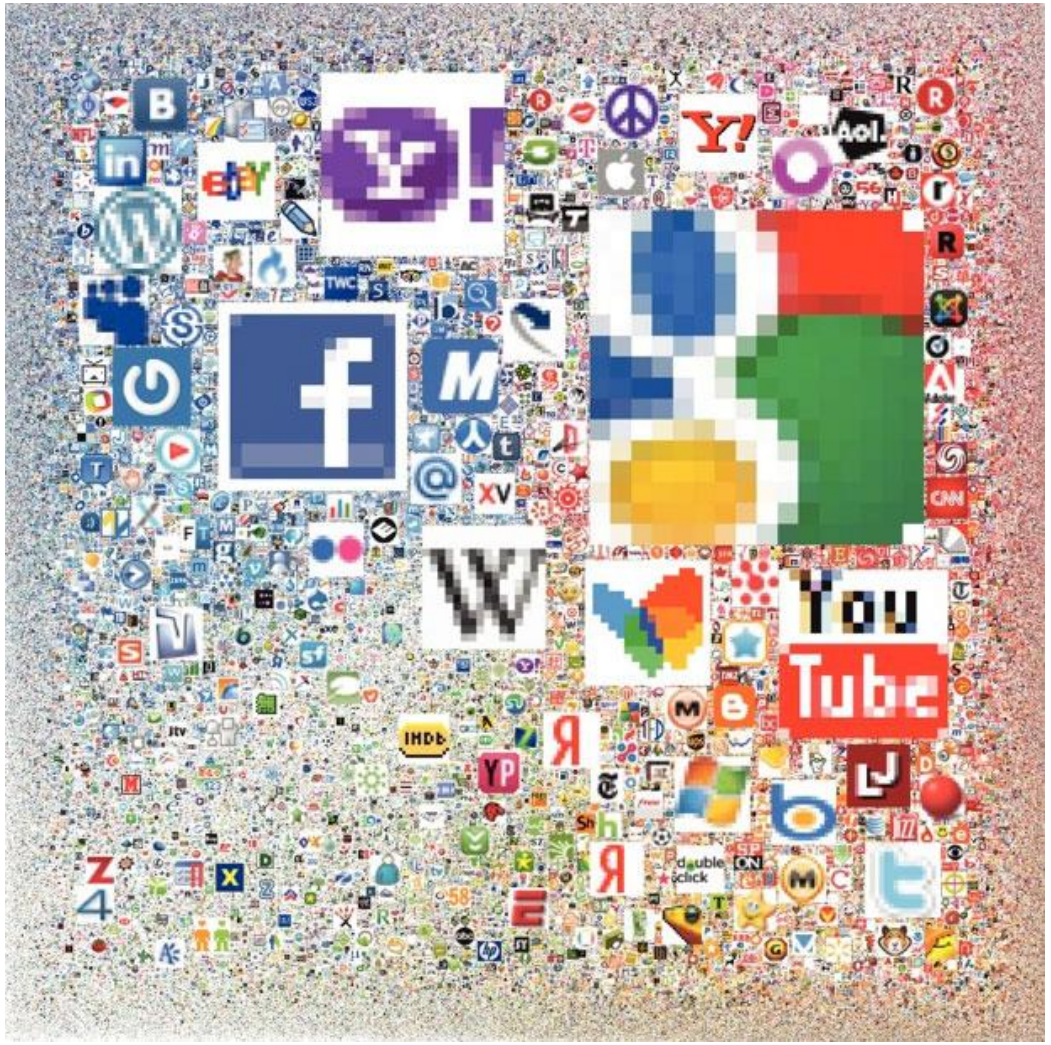
@Mark_J_Perry

October 12, 2015 7:33 pm | *AEIdeas*

Fortune 500 firms in 1955 v. 2015; Only 12% remain, thanks to the creative destruction that fuels economic prosperity

Carpe Diem

Only These 61 Companies Were in the Fortune 500 in Both 1955 and 2015		
3M	CVS	Lockheed Martin
Abbott Laboratories	Deere	Marathon Oil
Alcoa	Dow Chemical	McGraw Hill Financial
Alleghany	DuPont	Monsanto
Archer Daniels Midland	Exxon Mobil	Navistar
Ashland	Freeport-McMoRan	NCR
ATT	General Electric	Northrop Grumman
Avon Products	General Dynamics	Owens Corning
Boeing	General Mills	Owens-Illinois
BorgWarner	General Motors	PepsiCo



INTERNET ARCHIVE
INTERNET ARCHIVE

Fears of Trump prompt Internet Archive to make mirror site Canada

'Canadian law is actually quite friendly to digital libraries'

By Brady Strachan, CBC News Posted: Dec 02, 2016 7:00 AM PT | Last Updated: Dec 02, 2016 7:47 AM PT

Stay Connected with CBC News

Mobile Facebook Podcasts Twitter Alerts News

Weather

Severe weather warnings or watches in effect

Chetwynd Dawson Creek Fort St. John

Vancouver Kelowna Abbotsford Prince George Vict

Web archiving issues for COPPUL institutions

- Repatriating and storing WARC files in Canada
- Jurisdictional issues
- Copyright and other legal issues
- **Policies, procedures, workflows**
  - Appraisal and selection
  - Quality assurance and metadata
  - Etc., etc.

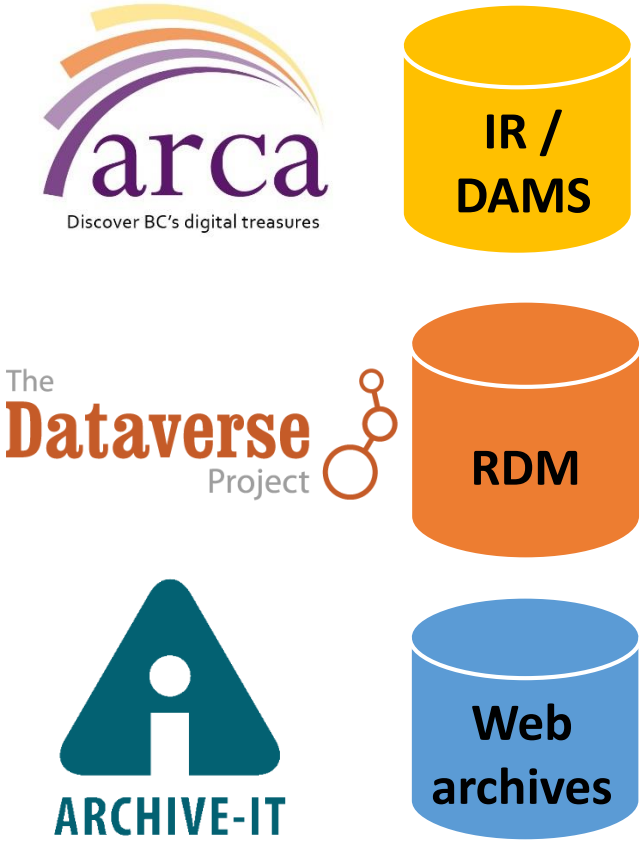
Libraries coming together…

- Expanding consortial license nationally
- Establishing a “Canadian Web Archiving Coalition” through CARL
- COPPUL looking to partner with Ontario Council of University Libraries (OCUL) for distributed archival storage of WARCs
- MOU with COPPUL institutions and U of Waterloo/U of Alberta to enable data-intensive analysis of Canadian web archives via Compute Canada infrastructure

Other platforms…

- Repository built on Islandora, a Canadian-developed open-source platform currently in use at over 100 public and private institutions worldwide
- Partnership w/ British Columbia Electronic Library Network (BC ELN)

- Open source research data repository software
- IQSS at Harvard leads development
- Canadian Working Group established under CARL's Portage Network

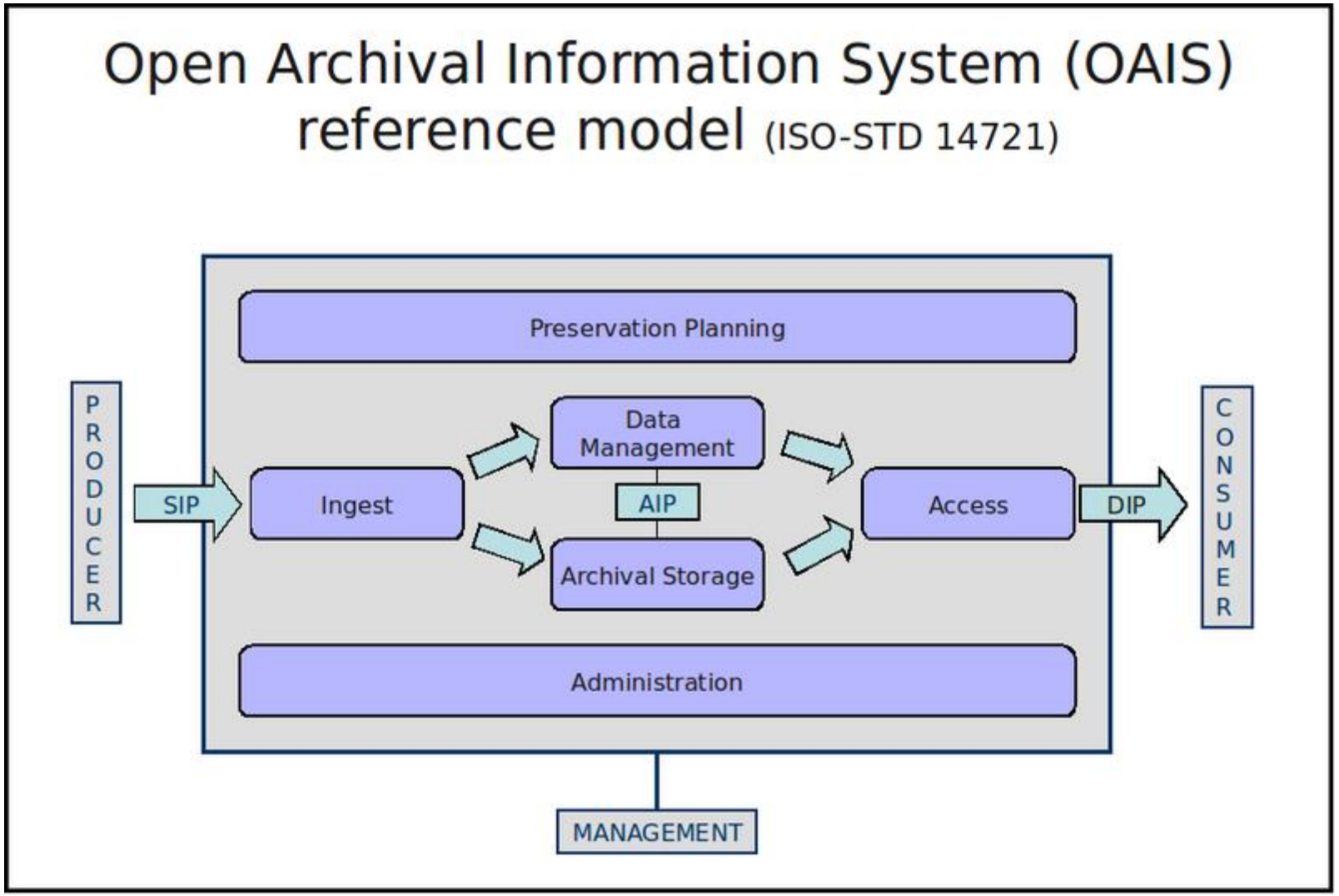

arca
Discover BC's digital treasures
IR / DAMS
The Dataverse Project
RDM
ARCHIVE-IT
Web archives

Processing for preservation: Archivematica

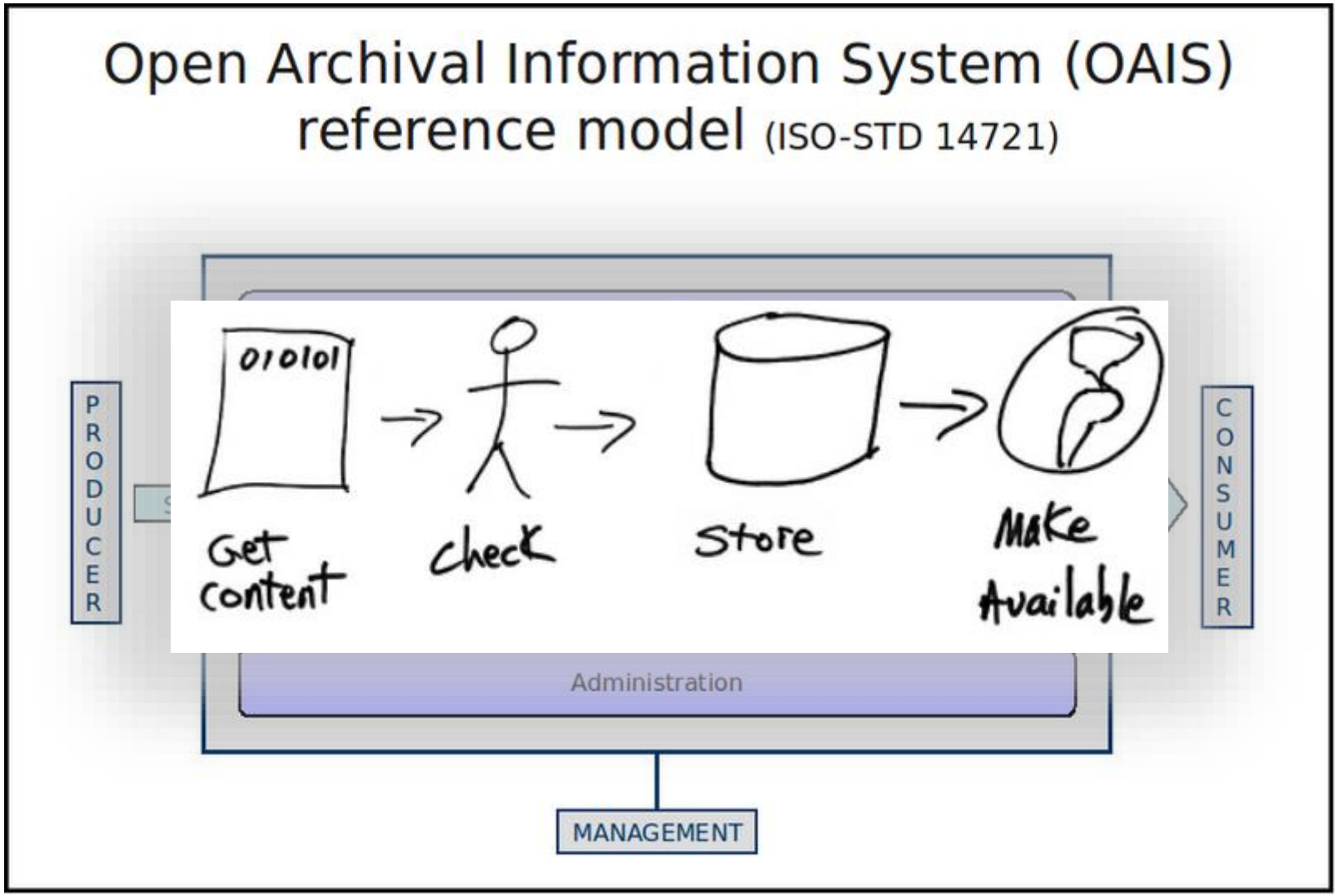
Archivematica-as-a-service as a digital preservation platform

archivematica

- “Archivematica provides an integrated suite of free and open-source tools that allows users to process digital objects from ingest to archival storage and access in compliance with the ISO-OAIS functional model and other digital preservation standards and best practices.”
  - https://wiki.archivematica.org/Overview

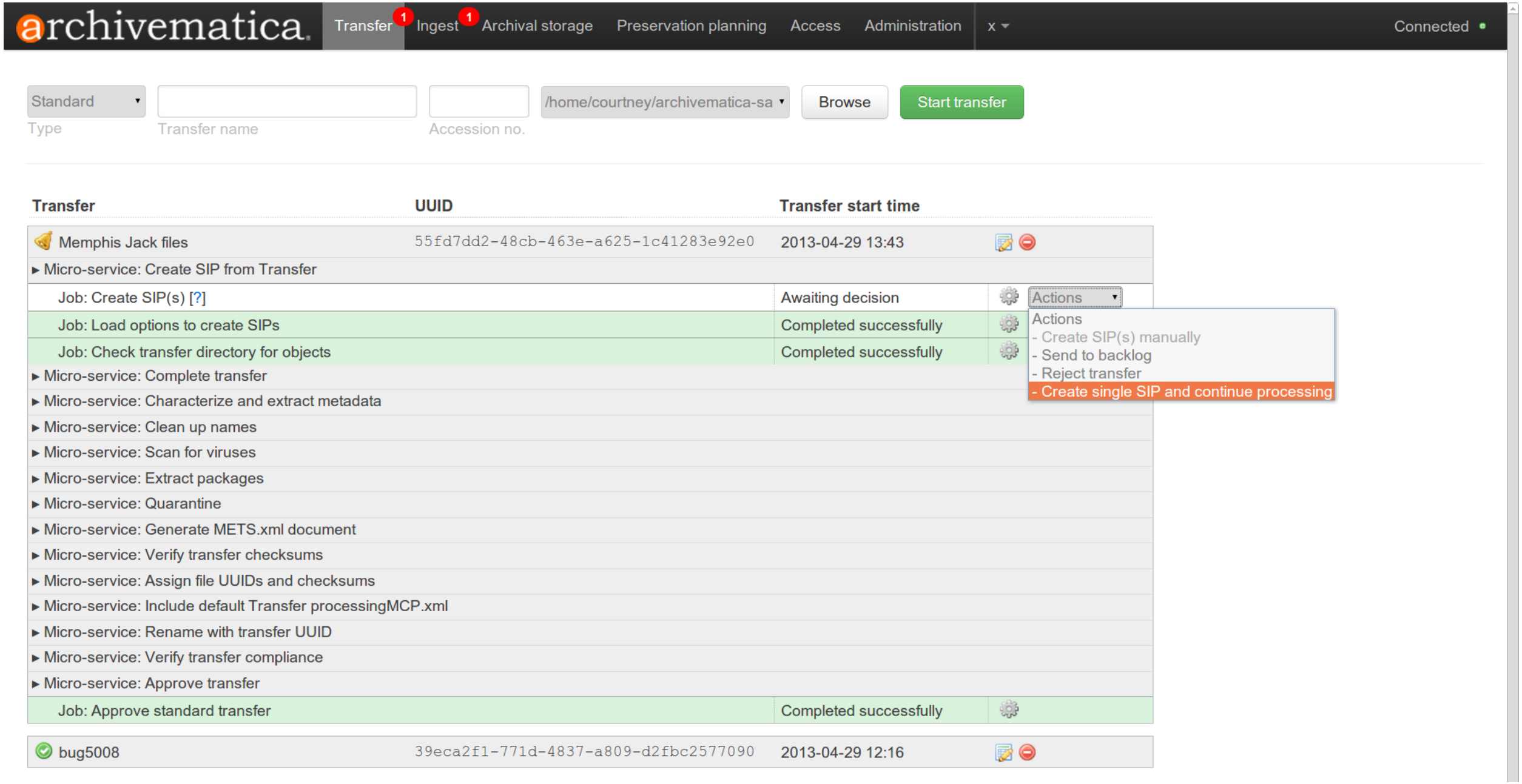

Open Archival Information System (OAIS) reference model (ISO-STD 14721)
Preservation Planning
PRODUCER
SIP
Ingest
Data Management
AIP
Archival Storage
Access
DIP
CONSUMER
Administration
MANAGEMENT

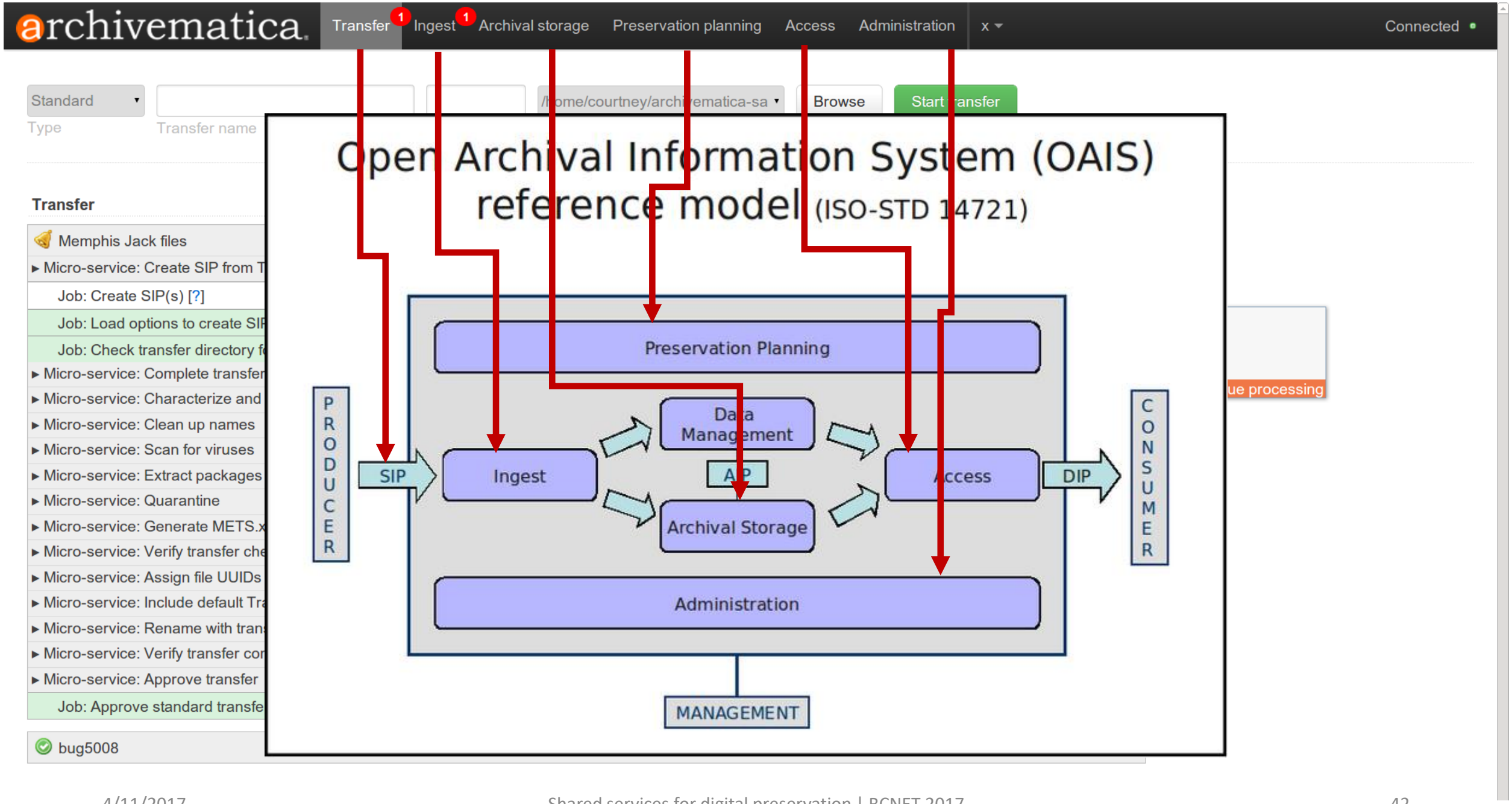
Open Archival Information System (OAIS) reference model (ISO-STD 14721)



Archivematica micro-services include

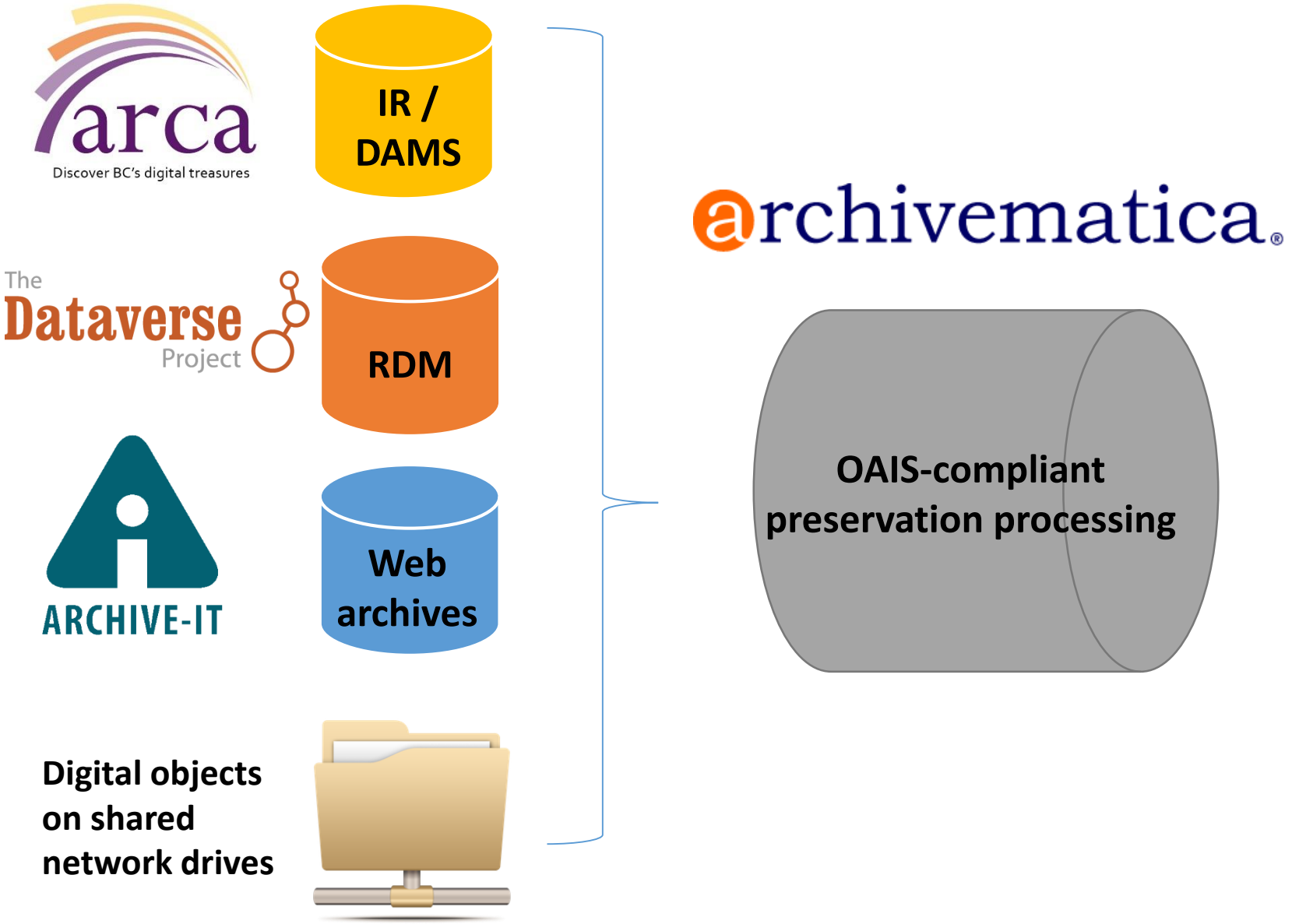
- Virus scans
- Checksums
- Format identification and validation
- Format normalization
- Creation of METS.xml files with PREMIS and other metadata
- Creation of Archival Information Packages (AIPs)


archivematica
Transfer 1
Ingest 1
Archival storage
Preservation planning
Access
Administration
x
Connected
Standard
Type
Transfer name
Accession no.
/home/courtney/archivematica-sa
Browse
Start transfer
Transfer
UUID
Transfer start time
Memphis Jack files
55fd7dd2-48cb-463e-a625-1c41283e92e0
2013-04-29 13:43
Micro-service: Create SIP from Transfer
Job: Create SIP(s) [?]
Awaiting decision
Actions
Job: Load options to create SIPs
Completed successfully
Job: Check transfer directory for objects
Completed successfully
Actions
- Create SIP(s) manually
- Send to backlog
- Reject transfer
- Create single SIP and continue processing
Micro-service: Complete transfer
Micro-service: Characterize and extract metadata
Micro-service: Clean up names
Micro-service: Scan for viruses
Micro-service: Extract packages
Micro-service: Quarantine
Micro-service: Generate METS.xml document
Micro-service: Verify transfer checksums
Micro-service: Assign file UUIDs and checksums
Micro-service: Include default Transfer processingMCP.xml
Micro-service: Rename with transfer UUID
Micro-service: Verify transfer compliance
Micro-service: Approve transfer
Job: Approve standard transfer
Completed successfully
bug5008
39eca2f1-771d-4837-a809-d2fbc2577090
2013-04-29 12:16


archivematica
Transfer
Ingest
Archival storage
Preservation planning
Access
Administration
Connected
Standard
Type
Transfer name
Browse
Start transfer
Transfer
Memphis Jack files
Micro-service: Create SIP from T
Job: Create SIP(s) [?]
Job: Load options to create SIP
Job: Check transfer directory f
Micro-service: Complete transfer
Micro-service: Characterize and
Micro-service: Clean up names
Micro-service: Scan for viruses
Micro-service: Extract packages
Micro-service: Quarantine
Micro-service: Generate METS.x
Micro-service: Verify transfer che
Micro-service: Assign file UUIDs
Micro-service: Include default Tra
Micro-service: Rename with trans
Micro-service: Verify transfer cor
Micro-service: Approve transfer
Job: Approve standard transfe
bug5008
Open Archival Information System (OAIS)
reference model (ISO-STD 14721)
Preservation Planning
PRODUCER
SIP
Ingest
Data Management
AIP
Archival Storage
Access
DIP
CONSUMER
Administration
MANAGEMENT
ue processing

Archivematica-as-a-service

- COPPUL promotes the service, signs up new institutions and provided seed funding for the one-time set-up costs
- Artefactual Systems provides installation, ongoing technical support and user training
- UBC provides a fee-based server hosting and digital object storage and back-up service


arca
Discover BC's digital treasures
IR / DAMS
The Dataverse Project
RDM
ARCHIVE-IT
Web archives
Digital objects on shared network drives
archivematica
OAIS-compliant preservation processing

Preservation storage: LOCKSS

Lots of Copies Keeps Stuff Safe

Preservation storage vs. standard IT storage

- Standard IT storage systems are generally designed for digital objects that are in **active use**, and generally include backup and other disaster recovery mechanisms
- Preservation storage tackles the **long-term viability** of digital objects and as a result addresses somewhat different but also overlapping risk scenarios

Standard IT storage threat scenarios

- Fire, flood, power failures
- Operator error
- Malicious attack
- Media degradation
- Obsolescence of storage systems and technologies

Preservation storage service threat scenarios

- Fire, flood, power failures
- Operator error
- Malicious attack
- Media degradation
- Obsolescence of storage systems and technologies
- **Organizational failure**
- **Economic failure**
- **Jurisdictional failure**

Standard IT storage and preservation storage services are **complementary**

Standard IT storage

- Designed for active use
- Robust disaster recovery
- Constantly refreshed media
- Crucial infrastructure to support digital preservation services

Preservation storage service

- Often characterized by different threat scenarios and governance structures that encourage inter-institutional collaboration
- Long-term thinking (i.e. 25-100 years)

Preservation storage service

- Stores standard AIPs w/ preservation-specific metadata
- **Independently** administered copies (no single point of failure)
- **Disperse content across several institutions or jurisdictions** to address large-scale political and/or economic threats and/or environmental threats
- Hardware diversity
- Polling and repair mechanisms to detect and fix “bit-rot”

A preservation storage service

- Undertake **long-term planning** that explicitly looks beyond the life of a particular institution
- Have **governance mechanisms** to address both short-term and long-term threats, including financial sustainability and succession planning
- Look beyond a **single institution or technological platform**, and work with other preservation initiatives to align infrastructure

What is LOCKSS?

- An open-source, library-led digital preservation system built on the principle that “lots of copies keep stuff safe”
- Through a LOCKSS distributed network, libraries cooperate with one other to ensure their preserved content remains authentic and authoritative over the long term

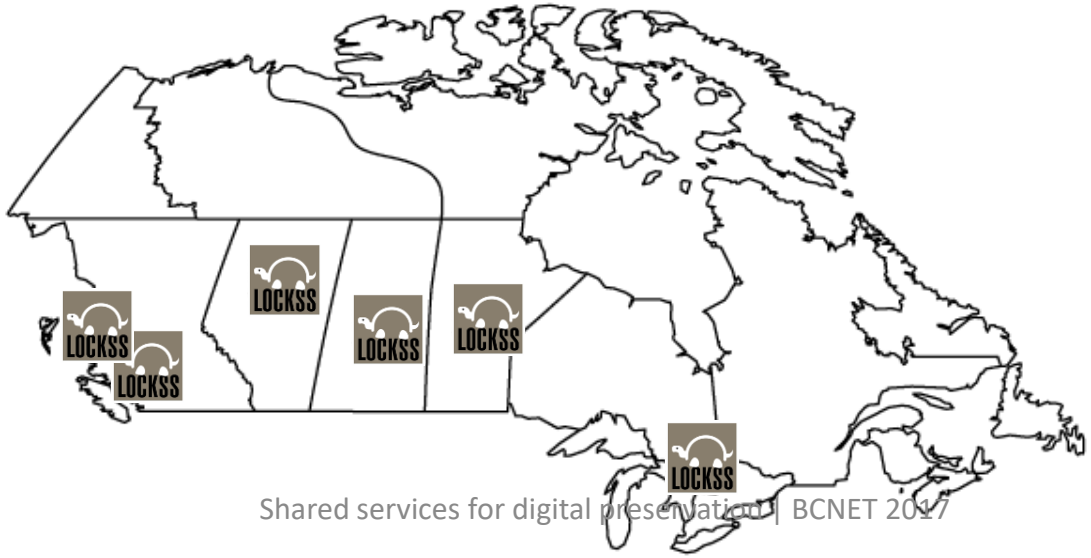
LOCKSS LOTS OF COPIES KEEP STUFF SAFE

What does a Private LOCKSS Network do?

- It **ingests** content and **preserves** it by continually comparing the content it has collected with the same content collected by other LOCKSS Boxes, and repairing any differences

WestVault

- COPPUL's emerging next generation Private LOCKSS Network
- Will run LOCKSS software at 5-7 university data centres with scalable storage
- **Ingests** content and **preserves** it by continually comparing the content it has collected with the same content collected by other LOCKSS Boxes, and repairing any differences



Putting it all together

Platforms + Processing + Storage

IR / DAMS

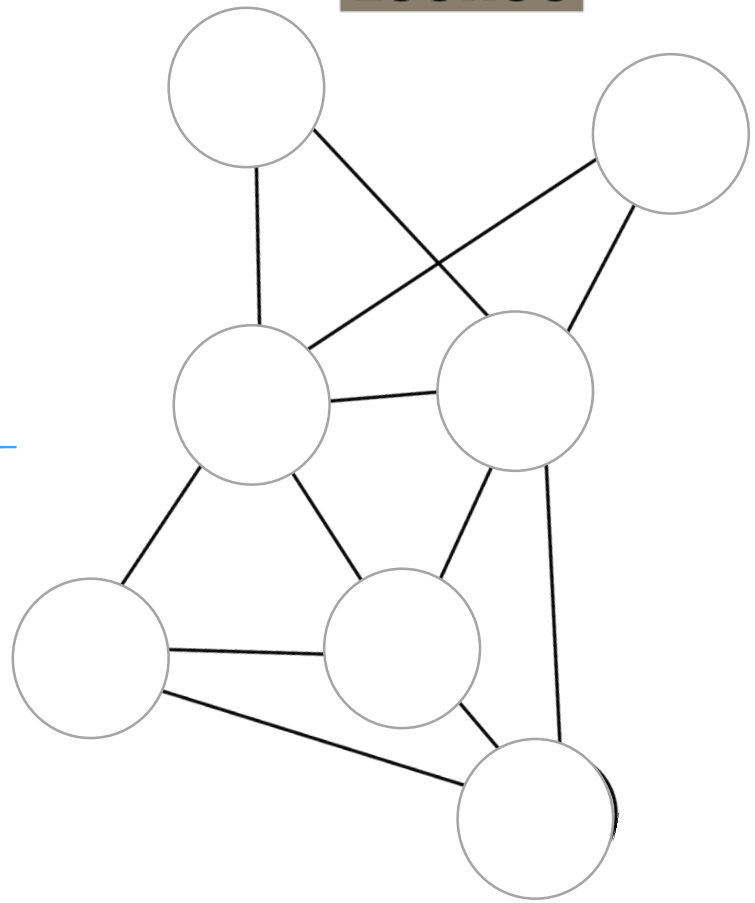
RDM

Web archives

Digital objects on shared network drives

archivematica

OAIS-compliant preservation processing



COPPUL Digital Preservation Network

- Its not just about the tools and services, its about helping members build the capacity to undertake digital preservation activities at their institutions, and helping them align their efforts with other nationally and internationally

To summarize…

- "…[i]f we are not to tolerate gaps in society's memory in the future, we must intervene in the present to secure the long-term availability of culturally significant digital materials."

  - Lavoie, B. F. (2006). The costs of digital preservation. Digital Preservation, 106-132.

Thanks!

- corey@coppul.ca