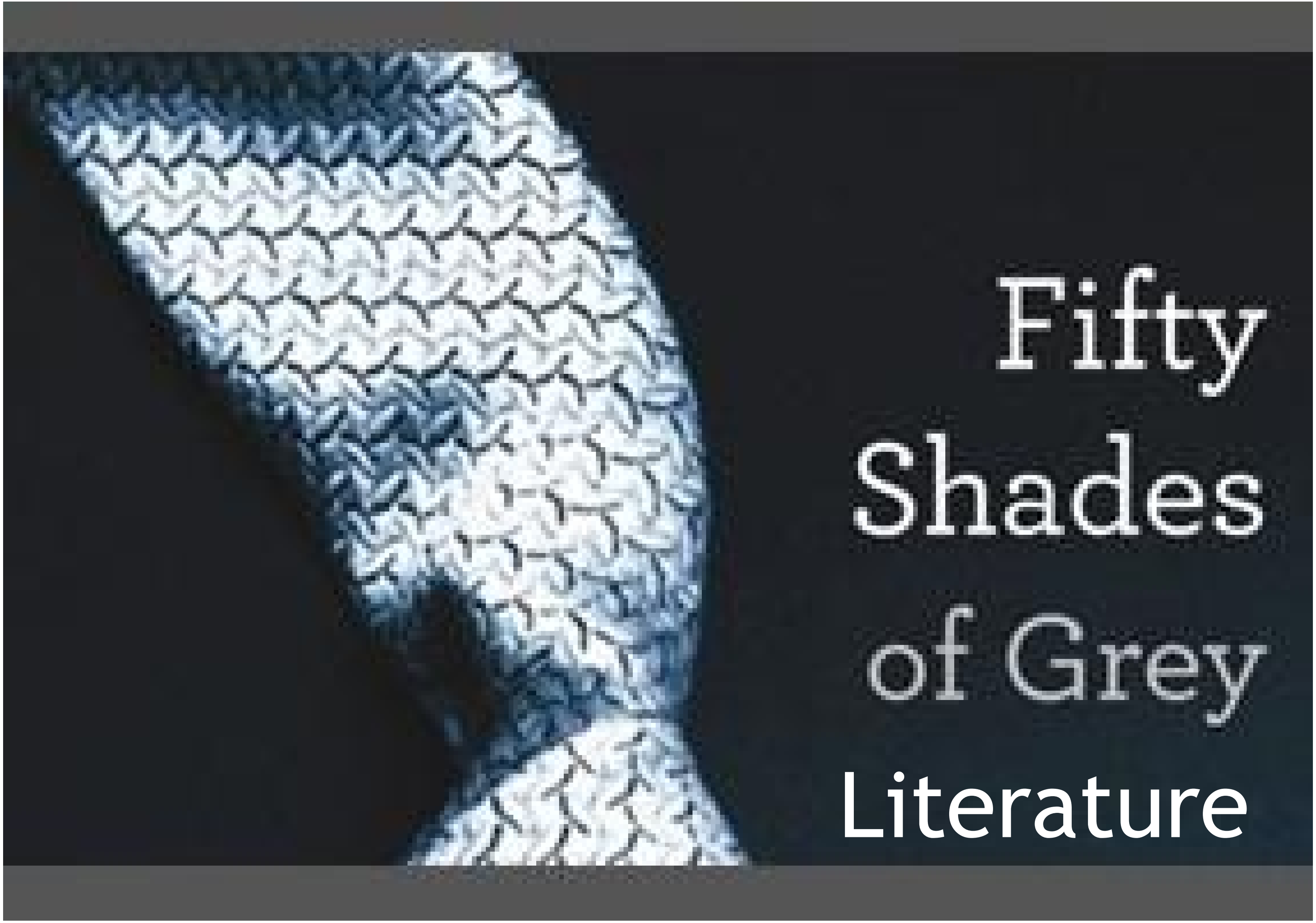
The Role of Grey Literature in Academic Library Collections: Discovering, Capturing, Preservation, & Access

Patti Sherbaniuk & Sean Luyk
University of Alberta Libraries

What We Will Cover

- Grey literature
    - What is it? Who produces it? Why do we want it?
- New born digital issues
- Web archiving at the UofA Libraries
    - Who and what is involved
    - Challenges
    - Business collections
    - Other collections
- What does it mean for you?

Grey Literature - What is It?



Definitions...

Luxembourg Definition (1997):
“Information produced by all levels of government, academics, business and industry in electronic and print formats **not controlled by commercial publishing** i.e. where publishing is not the primary activity of the producing body.” *Includes postscript added in 2004 at New York Conference
(Quoted in Schöpfel).

Definitions cont’d

Prague Definition (2010):

"Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are **protected by intellectual property rights**, of **sufficient quality** to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers i.e., where publishing is not the primary activity of the producing body."

Who Produces It?

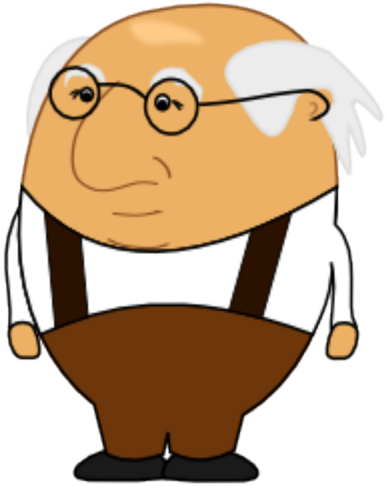
- International and governmental agencies
- Companies
- Non-profits
- Think tanks
- Universities, research institutes
- Professional associations
- Libraries, Archives, Museums
- Special Interest groups
- **can vary between disciplines

Why Do We Want It?

- Not indexed in major databases
- Not (always) actively collected
- Contributes to **comprehensive/balanced collections**
- Fulfill preservation mandates

Old Issues

- Access, preservation, ownership, cataloguing, **discoverability**

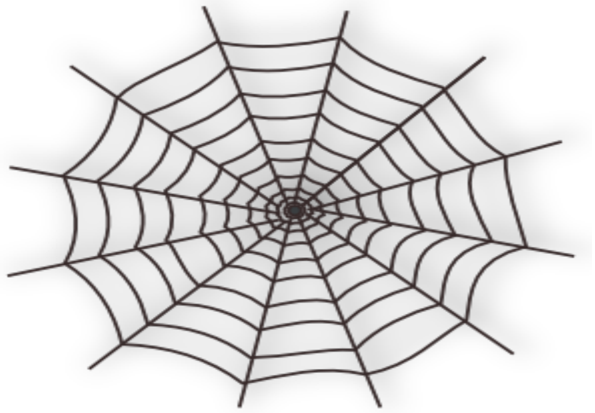


New Issues

- Changing collections: **born digital**, social media, open access, **discoverability**
- "greyness" will change (Gelfand and Lin 2013)

Why Web Archiving?

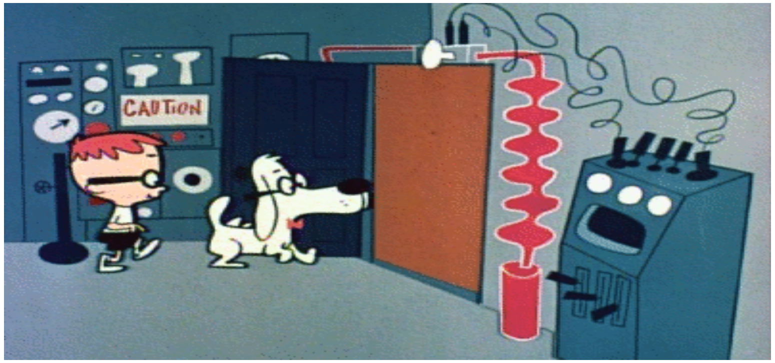
- Centered around the “grey literature problem”
- Grey literature more often born digital
- Viable way of collecting grey literature, using tools such as ArchiveIT



Internet Archive

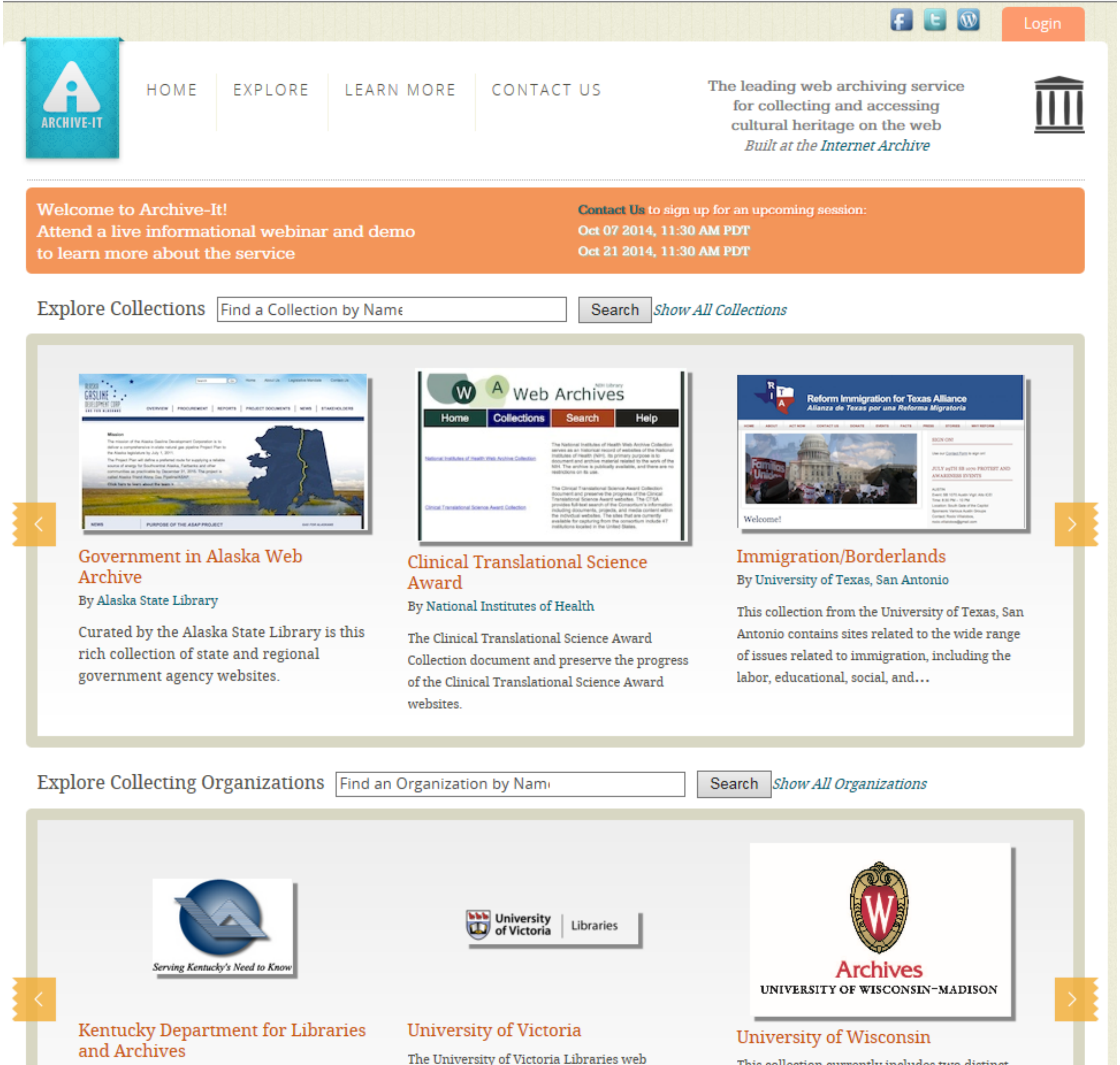
The Internet Archive is a 501(c)(3) non-profit that was founded to build an Internet library. Its purposes include offering permanent access for researchers, historians, scholars, people with disabilities, and the general public to historical collections that exist in digital format.
https://archive.org/about/

Collections are freely accessible to the public and displayed through the Internet Archive's Wayback Machine interface, highlighted on the UAL website, and for select literature - catalogue records the library catalogue.



- 2006- subscription web archiving service from the Internet Archive that helps organizations to harvest, build, and preserve collections of digital content.

https://www.archive-it.org/learn-more



The leading web archiving service
for collecting and accessing
cultural heritage on the web
Built at the Internet Archive

Welcome to Archive-It!
Attend a live informational webinar and demo
to learn more about the service

Contact Us to sign up for an upcoming session:
Oct 07 2014, 11:30 AM PDT
Oct 21 2014, 11:30 AM PDT

Explore Collections [Find a Collection by Name] Search *Show All Collections*

Government in Alaska Web Archive
By Alaska State Library
Curated by the Alaska State Library is this rich collection of state and regional government agency websites.

Clinical Translational Science Award
By National Institutes of Health
The Clinical Translational Science Award Collection document and preserve the progress of the Clinical Translational Science Award websites.

Immigration/Borderlands
By University of Texas, San Antonio
This collection from the University of Texas, San Antonio contains sites related to the wide range of issues related to immigration, including the labor, educational, social, and…

Explore Collecting Organizations [Find an Organization by Name] Search *Show All Organizations*

Serving Kentucky's Need to Know
University of Victoria
Libraries
Archives
UNIVERSITY OF WISCONSIN–MADISON

Kentucky Department for Libraries and Archives

University of Victoria
The University of Victoria Libraries web

University of Wisconsin
This collection currently includes two distinct

How it Works

Standard practice was to literally print off web site data and have it physically catalogued for a library collection.

- Based on “**seed URLs**” provided, an Archive-it crawler captures all unique URLs and archives them
- A seed is any url that you (as the curator) tell the crawler you want to capture. A seed could be:
  - an entire website
  - a specific part(directory) of a website
  - a specific url
- What exactly your seed url(s) are determine how much of each of those sites will be archived.

“seed metadata”

- Unique metadata can be added at the document level to help users locate things using keywords.
- IA provides 15 standard Dublin Core metadata fields.
- Additionally, **customized metadata fields** allow institutions to enter metadata field names and corresponding values beyond the standard Dublin Core fields.

“group level” and “group level metadata”
mention where you can edit

“collection level” and “collection level metadata”
note fields, image etc for the COLLECTION

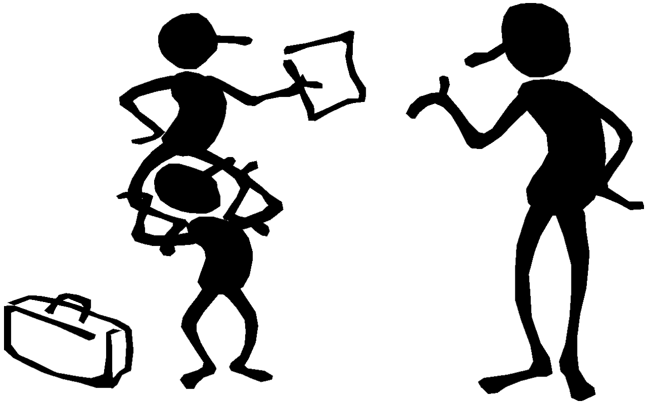
Who Uses It?

- 300 partner organizations in 48 U.S. states and 16 countries worldwide, including:
- College and University Libraries
- State Archives, Libraries, and Historical Societies
- Federal Institutions and NGOs
- Museums and Art Libraries
- Public Libraries, Cities and Counties

- Archive-It has collected **8,884,010,821** URLs for **2,638** public collections thus far...

Archive-it Canadian Partners

- Canada's National History Society
- Library and Archives Canada
- United Church of Canada Archives
- National Gallery of Canada Library and Archives
- Canada's National History Society
- COPPUL partners



US- Some Major Players

- Harvard
- Arizona State
- Brigham Young
- Columbia
- Cornell
- Duke
- Penn State
- Stanford
- University of Minnesota
- USC
- NASA Images
- Church of Latter-Day Saints
- High schools
- Public libraries
- and growing…...

In the Beginning…

- “rescue operation”

2010 Business Pilot Goals

- Learn about ArchiveIT software and web archiving in general
- Determine content for web archiving on Pembina Institute and Canadian Association of Petroleum Producers (CAPP) sites
- Set up crawls of sites
- Analyze captured materials to determine most appropriate means of providing access
- Web archiving permissions
- Procedures and guidelines

the PEMBINA institute

Sustainable Energy Solutions

CAPP

CANADIAN ASSOCIATION OF PETROLEUM PRODUCERS

Canada's Oil and Natural Gas Producers

Copyright and Ethical Crawling

- Cautious approach
- Consider impact on content providers
- Web archiving permissions template

Develop Workflow Guidelines

- Who does what, how are collections maintained, etc.
- Started longer process of development of policies and procedures
- Moving target, as releases remove need for specific procedures (e.g. de-duplication)

More Crawls for Business

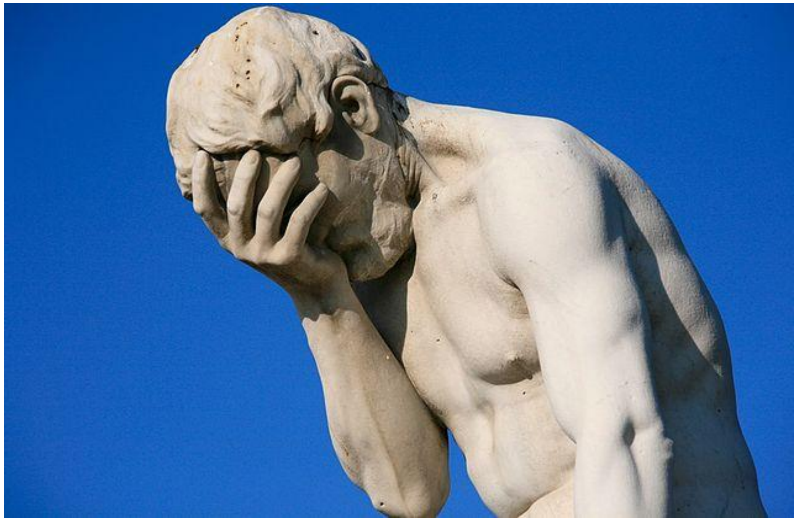
- Financial websites- Deloitte, BMO, Avison Young, Colliers
- Energy and the Environment- CAPP, CEMA, CEPA, CleanEnergy, Greenpeace Canada, CEAA-ACEE, NPA, Pembina Institute
- Oil Sands publications

Other Collections at the U of A

- Gov’t docs
- Energy/Environment
- Education curriculum
- Circumpolar
- Health Sciences
- Prairie Provinces
- Western Canadian Francophonie
- Humanities Computing
- U of A websites
- Idle No More

Issues

- Duplication (of web content- recrawling)- after “timing out”
- Duplication (between partners and the Internet Archive itself)
- Ethical crawling/permissions/copyright
- Cannot cherry pick
- Need precise URLS to capture specific content
- Staff time
- No formal “best practices”
- Social media?
- Sheer scope of the Web
- How to weed digitally?
- **Discoverability**



Our Users

Questions about our web archives

● Who are our existing users? Potential users?

● Where do they discover our archives? Where would they expect to?

● How do they discover our archives?

- do they search for, within and across our archives?

Linked to Metadata

- At what level? In what format? Editing existing metadata from AI

Business Pilot #2

- Pilot project started with 2 business-related websites as “guinea pigs” - CEPA (Canadian Energy Pipeline Association) and BMO
- Simple stylesheet to pull URLs and import and edit existing metadata from the AI feed
- Editing= add column headings to stylesheet- title, date, author, subject terms, description, relation, coverage, language
- Ongoing….

New and Improved

- Archive-it version 5.0- release late October will include a new interface, better crawl reports, access to statistics using Google analytics
- Splitting, sharing and merging collections and seeds
- Can resume crawls, export of metadata, removing "out of scope" data
- Social media archiving
- and much more!

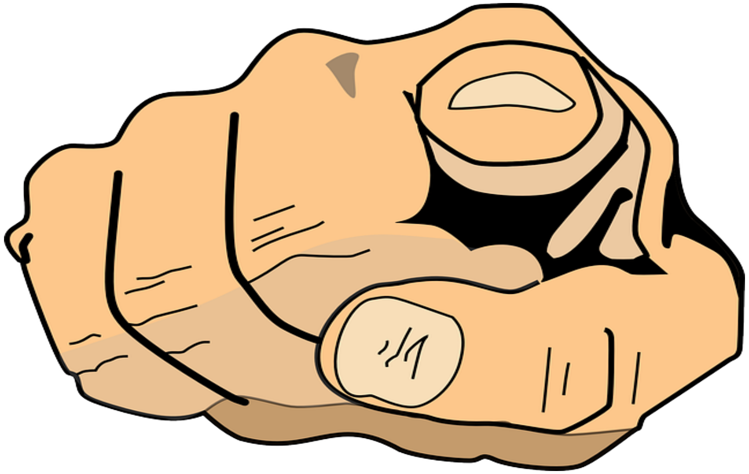
Web Archiving Life Cycle Model

"an attempt to incorporate the technological and programmatic arms of web archiving into a framework that will be relevant to any organization seeking to archive the web."

https://archive-it.org/static/files/archiveit_life_cycle_model.pdf

What it Means for You

- Free, searchable access from our catalogue
- Better access to grey literature in various disciplines that is:
  - difficult to uncover and
  - is born digital and at the risk of disappearing



The Road Forward

Thank You!

Patti Sherbaniuk | psherban@ualberta.ca

Sean Luyk | sean.luyk@ualberta.ca