Web Harvesting the Government of Canada sites

A Collaborative Effort

Sam-chin Li
Tom Smyth
Amanda Wakaruk

What is Web Harvesting

Web sites are collected via software that downloads code, images, documents, and others files essential to completely and faithfully reproduce the web site at the time of capture.

International Internet Preservation consortium

Why harvesting Government web sites

The web is often the only source for government information.

Web archives keep important records of the government and ensure democracy for now and for the future.

March 2013

- **Reduce Redundant, Outdated and Trivial Content (ROT Criteria)**
- **GoC Web Convergence Plan**
- **UofT starts to archive GoC web content**

Background

What is seed?

A URL appearing in a seed list as one of the starting addresses a web crawler uses to capture content. Also called a targeted URL.

Archive-IT glossary

Scope and Frequency

- Government of Canada Departments, Agencies, Crown Corporations, Special Operating Agencies and various affiliated organizations.
- One time only
- Daily and Semiannually
- Separate seeds for important resources

Seed List based on this Page

We started too late!

Many contents have already disappeared from the Web!

March 2013

Departments and Agencies

This page provides an alphabetical listing of links to Government of Canada Departments, Agencies, Crown Corporations, Special Operating Agencies and various affiliated organizations.

If the name of the organization you are looking for is not listed, try consulting the *Financial Administration Act*. It contains information on al current Government of Canada departments, agencies, Crown corporations and special operating agencies, and also includes a listing of organizations that no longer exist or that have been privatized.

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A

- Aboriginal Affairs and Northern Development Canada
  - First Nations Statistical Institute
  - Indian Oil and Gas Canada
  - Indian Residential Schools
- Agriculture and Agri-Food Canada
  - Canadian Biodiversity Information Facility
  - Canadian Pari-Mutuel Agency
- Assisted Human Reproduction Canada
- Atlantic Canada Opportunities Agency
- Atlantic Pilotage Authority
- Atlas of Canada - Natural Resources Canada
- Atomic Energy of Canada Limited
- Auditor General of Canada, Office of

Solution:
Wayback Machine

Strategy Changed

June, 2013

- **Checked the gaps in the Wayback Machine GoC Harvest**
- **Crawled the gaps**
- **Purchased the content**
- **Moved it to UofT Archive-IT account**

Purchased Wayback Machine Harvest from

December 2007 – March 2013

Archive-IT: University of Toronto

ARCHIVE-IT

Canadian Government Information Web Archive (University of Toronto)

Enter Web Address: http:// All Take Me Back

Searched for http://agr.gc.ca/ 191 Results RSS Me

Look up URL in general Internet Archive web collection Proxy Mod

* denotes when page was updated

Found 191 Captures between Dec 2, 2007 - Aug 13, 2013

2007	2008	2009	2010	2011	2012	2013
1 page	48 pages	12 pages	3 pages	15 pages	12 pages	100 pages
Dec 2, 2007 *	Feb 16, 2008	Jan 29, 2009	Feb 13, 2010	Jun 5, 2011	Feb 4, 2012	Jan 13, 2013 *
	Feb 19, 2008	Feb 6, 2009	Oct 7, 2010	Jul 6, 2011	Feb 5, 2012	Jan 14, 2013 *
	Feb 27, 2008	Feb 7, 2009	Dec 14, 2010	Jul 9, 2011	Feb 6, 2012	Mar 12, 2013
	Mar 8, 2008	Feb 25, 2009		Aug 5, 2011	Feb 15, 2012	May 9, 2013
	Apr 6, 2008	Feb 25, 2009		Aug 7, 2011	Feb 18, 2012	May 15, 2013
	Apr 8, 2008	Feb 26, 2009		Aug 11, 2011 *	Apr 10, 2012	May 16, 2013
	Apr 14, 2008	Mar 4, 2009		Aug 24, 2011 *	Apr 15, 2012	May 19, 2013
	May 27, 2008	Mar 14, 2009		Aug 25, 2011	Jun 26, 2012	May 22, 2013
	Jun 27, 2008	Mar 18, 2009		Sep 2, 2011	Jun 30, 2012 *	May 22, 2013 *
	Aug 9, 2008	Mar 25, 2009		Sep 24, 2011	Jul 28, 2012 *	May 22, 2013 *

UoT Archive IT: Two Collections Combined Seamlessly

ing an archived web page, collected at the request of University of Toronto using Archive-It. This page was capture
04, 2013, and is part of the Canadian Government Information collection. The information on this web page may be
date. See All versions of this archived page.

Example of an archived page: Canada News Centre

Workflow

- Select sites
- Crawl test to identify the scope
- Schedule the frequency
- Harvesting
- Quality Assurance
- Patch crawls
- Create metadata

Seed Selection & Quality Assurance

How can we make sure the important content is archived in a such **short timeframe** ?


HELP!

Library community involvement

URLs Recommended by subject specialists:

- 46

Libraries Volunteered their time:

- Industrial Relations & Human Resources
- OISE

Seeds

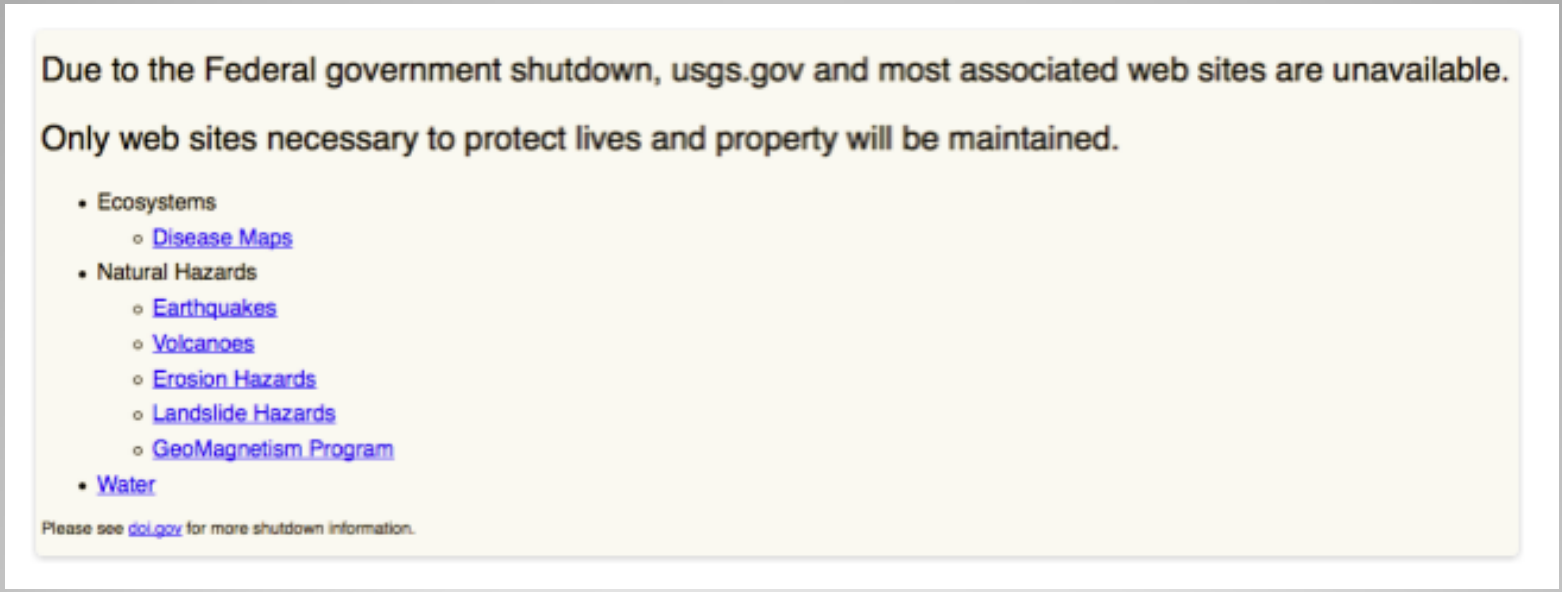
- Captured: 146
- Transferred from Global Wayback: 418

http://www.archive-it.org/collections/3608

University of Toronto Archive-IT

The Team

- **Sam-chin Li, Government Information Librarian**
- **Don McLeod, Archive-IT account coordinator at UofT**
- **Nicholas Worby, Graduate Student Library Assistant**



"Announcing the first ever Archive-It US Government Shutdown Notice Awards!"

Archive-IT Blog http://blog.archive-it.org/2013/10/11/announcing-the-first-ever-archive-it-us-government-shutdown-notice-awards/

Web Content Precarity

1996: Internet Archive founded and starts web harvesting activities

2001: Wayback Machine launches

2002: First release of Heritrix (web crawler)

2003: International Internet Preservation Consortium (IIPC)

2006: Archive-It, one of the first web archiving services

2009: WARC file, format generated by Heritrix, becomes an ISO standard

2009: University of Alberta Libraries (UAL) starts web harvesting with Archive-IT

Adapted from Scott Reed's slides: Wakaruk, Lau, Reed, "Community Tools and Best Practices for Harvesting and Preserving At-Risk Web Content." Association of Canadian Archivists, June 2013 http://hdl.handle.net/10402/era.32072

Web Harvesting

Web Harvesting at UAL 2013

Current Status (October 2013)

- 13 collections, > 67 million documents crawled (3.6 TB), ~600 active seeds, >1700 websites: ~50-60 are gc.ca
- government web content distributed across multiple collections: e.g., Energy & Environment, Health Sciences Grey Literature, Idle No More, Circumpolar, Government Information Collection (~133 seeds, mostly Alberta)

UAL Collection Development Committee

Born Digital Working Group (2009-2014)

- Leads the Libraries' direction related to born digital collections.
- Members: AUL, Collections Manager; Librarians: Digital Initiatives, Digital Repository Services, Metadata, Govt Info, Circumpolar, Business, Humanities and Social Sciences

Policies and Procedures

- Policy and Procedures for Born Digital Collections
- Metadata and Cataloguing Policy
- Web Archiving Guidelines (2): Selectors and Collection Managers
- Web Archiving Policy; Web Archiving Agreement

Governance and Workflow

ACCESS: contents will be made publicly available

OWNERSHIP: remains with the website owner

AUTHORIZATION: we cannot authorize re-use

TAKE-DOWN: we will evaluate requests from content owners to remove content

LIABILITY: we will not be held liable for actions of website owners

http://www.library.ualberta.ca/webarchive/

UAL Web Archiving Policy (excerpts)

LIBRARIES

DATABASES | JOURNALS | SUBJECTS | LIBRARIES | MY ACCOUNT | SERVICES

SEARCH ALL UOFA ARCHIVE-IT COLLECTIONS

Search Web Archive

BROWSE ARCHIVE-IT COLLECTIONS

All University of Alberta Websites

Alberta Education Curriculum Collection

Canadian Business Grey Literature Collection

Circumpolar Collection

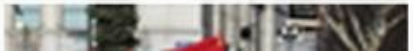
Energy/Environment Collection

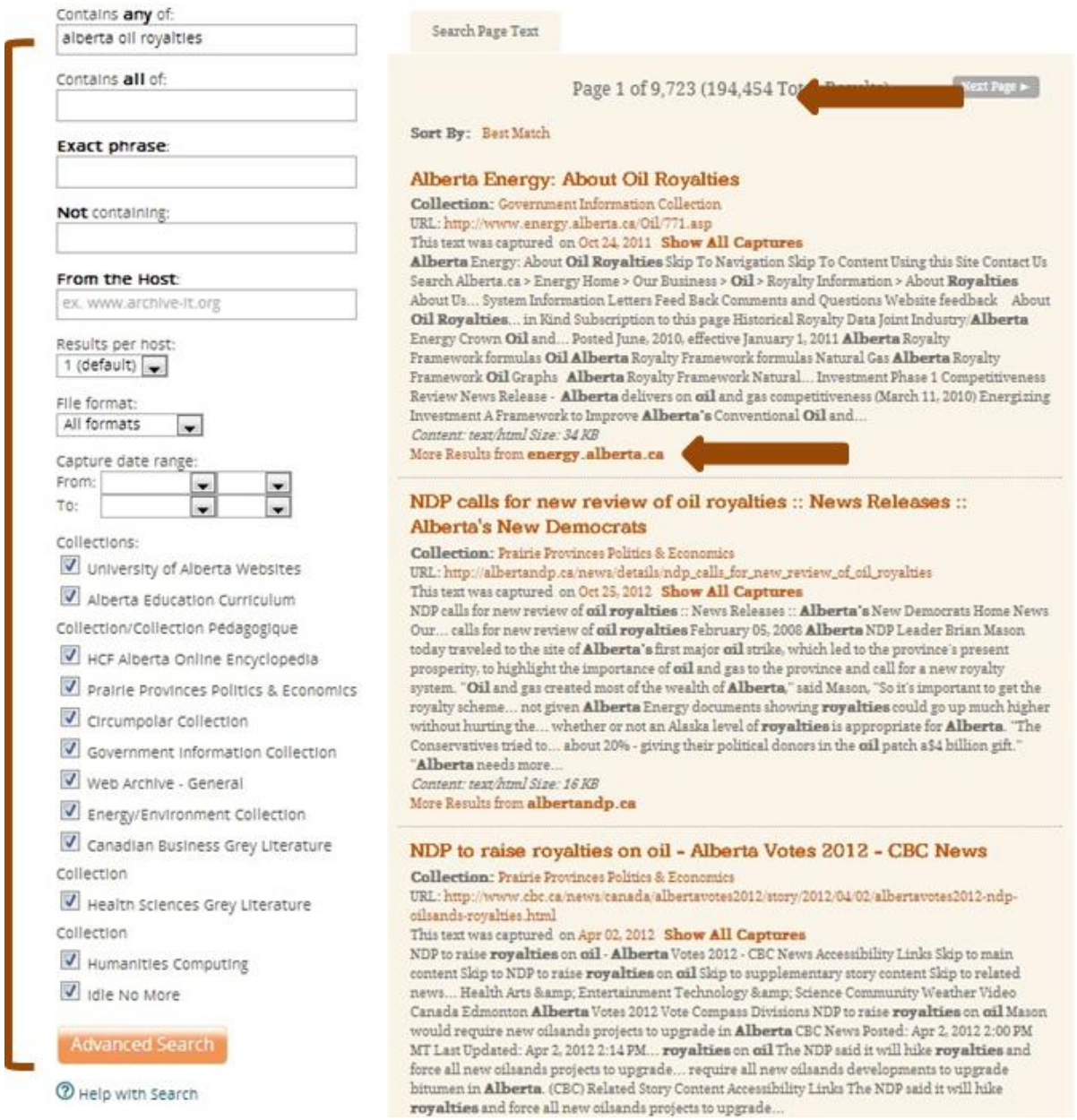
Government Information Collection

Health Sciences Grey Literature Collection

Heritage Community Foundation Online Encyclopedia

Idle No More




Contains any of:
alberta oil royalties
Contains all of:
Exact phrase:
Not containing:
From the Host:
ex. www.archive-it.org
Results per host:
1 (default)
File format:
All formats
Capture date range:
From:
To:
Collections:
University of Alberta Websites
Alberta Education Curriculum
Collection/Collection Pedagogique
HCF Alberta Online Encyclopedia
Prairie Provinces Politics & Economics
Circumpolar Collection
Government Information Collection
Web Archive - General
Energy/Environment Collection
Canadian Business Grey Literature
Collection
Health Sciences Grey Literature
Collection
Humanities Computing
Idle No More
Advanced Search
Help with Search
Search Page Text
Page 1 of 9,723 (194,454 To
Next Page ►
Sort By: Best Match
Alberta Energy: About Oil Royalties
Collection: Government Information Collection
URL: http://www.energy.alberta.ca/Oil/771.asp
This text was captured on Oct 24, 2011 Show All Captures
Alberta Energy: About Oil Royalties Skip To Navigation Skip To Content Using this Site Contact Us Search Alberta.ca > Energy Home > Our Business > Oil > Royalty Information > About Royalties About Us... System Information Letters Feed Back Comments and Questions Website feedback About Oil Royalties... in Kind Subscription to this page Historical Royalty Data Joint Industry/Alberta Energy Crown Oil and... Posted June, 2010, effective January 1, 2011 Alberta Royalty Framework formulas Oil Alberta Royalty Framework formulas Natural Gas Alberta Royalty Framework Oil Graphs Alberta Royalty Framework Natural... Investment Phase 1 Competitiveness Review News Release - Alberta delivers on oil and gas competitiveness (March 11, 2010) Energizing Investment A Framework to Improve Alberta's Conventional Oil and...
Content: text/html Size: 34 KB
More Results from energy.alberta.ca
NDP calls for new review of oil royalties :: News Releases :: Alberta's New Democrats
Collection: Prairie Provinces Politics & Economics
URL: http://albertandp.ca/news/details/ndp_calls_for_new_review_of_oil_royalties
This text was captured on Oct 25, 2012 Show All Captures
NDP calls for new review of oil royalties :: News Releases :: Alberta's New Democrats Home News Our... calls for new review of oil royalties February 05, 2008 Alberta NDP Leader Brian Mason today traveled to the site of Alberta's first major oil strike, which led to the province's present prosperity, to highlight the importance of oil and gas to the province and call for a new royalty system. "Oil and gas created most of the wealth of Alberta," said Mason, "So it's important to get the royalty scheme... not given Alberta Energy documents showing royalties could go up much higher without hurting the... whether or not an Alaska level of royalties is appropriate for Alberta. "The Conservatives tried to... about 20% - giving their political donors in the oil patch a$4 billion gift." "Alberta needs more...
Content: text/html Size: 16 KB
More Results from albertandp.ca
NDP to raise royalties on oil - Alberta Votes 2012 - CBC News
Collection: Prairie Provinces Politics & Economics
URL: http://www.cbc.ca/news/canada/albertavotes2012/story/2012/04/02/albertavotes2012-ndp-oilsands-royalties.html
This text was captured on Apr 02, 2012 Show All Captures
NDP to raise royalties on oil - Alberta Votes 2012 - CBC News Accessibility Links Skip to main content Skip to NDP to raise royalties on oil Skip to supplementary story content Skip to related news... Health Arts & Entertainment Technology & Science Community Weather Video Canada Edmonton Alberta Votes 2012 Vote Compass Divisions NDP to raise royalties on oil Mason would require new oilsands projects to upgrade in Alberta CBC News Posted: Apr 2, 2012 2:00 PM MT Last Updated: Apr 2, 2012 2:14 PM... royalties on oil The NDP said it will hike royalties and force all new oilsands projects to upgrade... require all new oilsands developments to upgrade bitumen in Alberta. (CBC) Related Story Content Accessibility Links The NDP said it will hike royalties and force all new oilsands projects to upgrade...

Metadata Needs Assessed on Collection Basis

Mixed model of metadata:

- Traditional document level: full, brief, use warrant access points
- Non-traditional: AI text search/widgets, AI metadata

Ongoing:

- Exploring options for metadata creation (e.g., copy, outsourcing)
- Investigating discovery access options (e.g., EDS, AI Portal, others?)
- Developing use case studies (sub-group of BDWG)

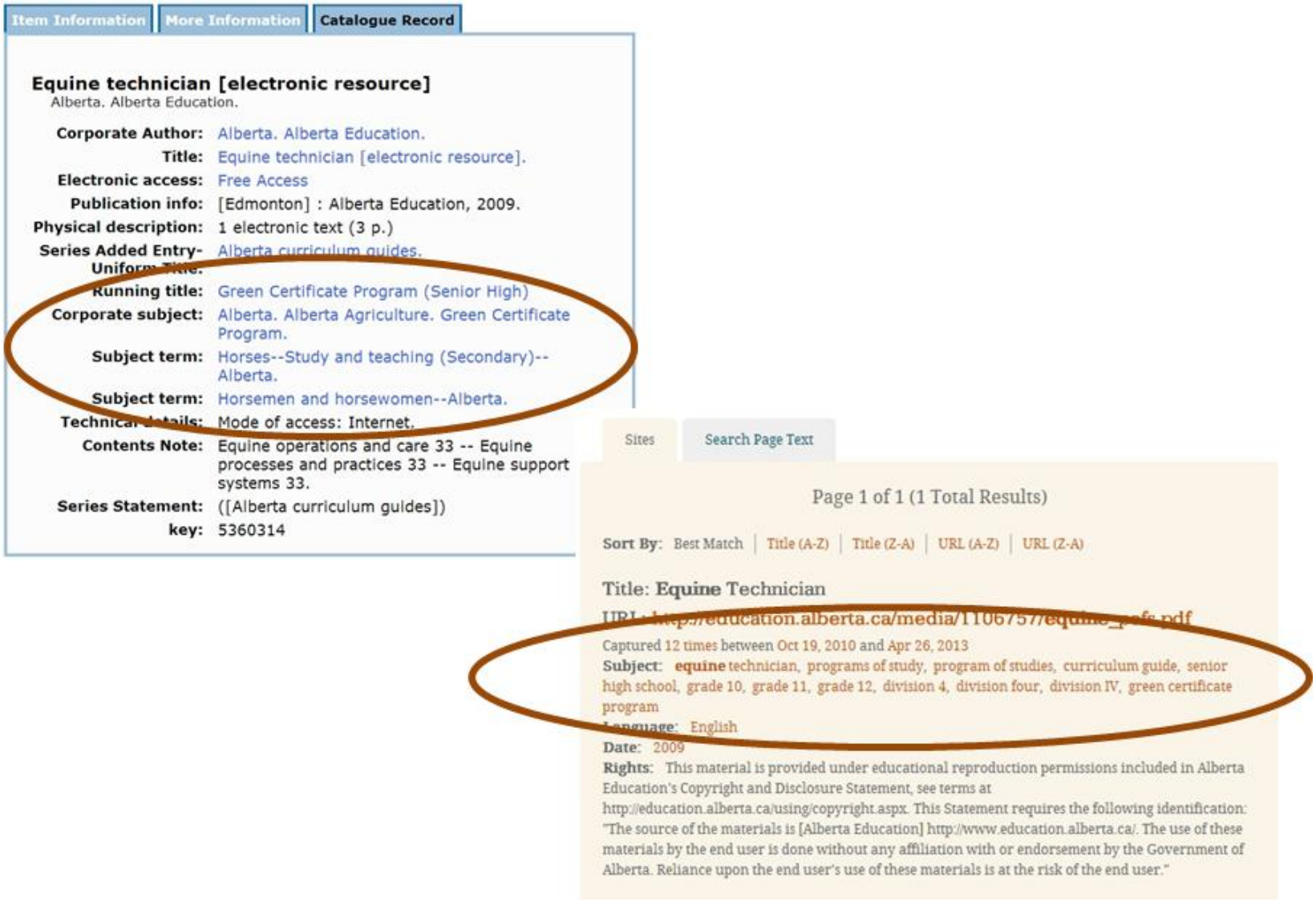
Access = Metadata

Item Information | More Information | Catalogue Record

Mis-à-jour Giant [electronic resource] : Projet d'assainissement de la mine Giant

Projet d'assainissement de la mine Giant (Canada)

Corporate Author:	Projet d'assainissement de la mine Giant (Canada)
Title:	Mis-à-jour Giant [electronic resource] : Projet d'assainissement de la mine Giant.
Electronic access:	Free Access
Publication info:	[Yellowknife] : Indian and Northern Affairs Canada, [2008]
Physical description:	1 online resource.
General Note:	Distributed by the Government of Canada Depository Services Program.
General Note:	"QS-Y289-005-FF-A1".
General Note:	Also published under the title: The Giant update, Giant Mine Remediation Project.
Added Entry-Corporat:	Canada. Indian and Northern Affairs Canada.
Added Entry-Corporat:	Northwest Territories.
Govt Canada class#:	R76-1/2007F-PDF
key:	6062544


Item Information
More Information
Catalogue Record
Equine technician [electronic resource]
Alberta. Alberta Education.
Corporate Author: Alberta. Alberta Education.
Title: Equine technician [electronic resource].
Electronic access: Free Access
Publication info: [Edmonton] : Alberta Education, 2009.
Physical description: 1 electronic text (3 p.)
Series Added Entry-Uniform Title: Alberta curriculum guides.
Running title: Green Certificate Program (Senior High)
Corporate subject: Alberta. Alberta Agriculture. Green Certificate Program.
Subject term: Horses--Study and teaching (Secondary)--Alberta.
Subject term: Horsemen and horsewomen--Alberta.
Technical details: Mode of access: Internet.
Contents Note: Equine operations and care 33 -- Equine processes and practices 33 -- Equine support systems 33.
Series Statement: ([Alberta curriculum guides])
key: 5360314
Sites
Search Page Text
Page 1 of 1 (1 Total Results)
Sort By: Best Match | Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)
Title: Equine Technician
URL: http://education.alberta.ca/media/1106757/equine_pafs.pdf
Captured 12 times between Oct 19, 2010 and Apr 26, 2013
Subject: equine technician, programs of study, program of studies, curriculum guide, senior high school, grade 10, grade 11, grade 12, division 4, division four, division IV, green certificate program
Language: English
Date: 2009
Rights: This material is provided under educational reproduction permissions included in Alberta Education's Copyright and Disclosure Statement, see terms at http://education.alberta.ca/using/copyright.aspx. This Statement requires the following identification: "The source of the materials is [Alberta Education] http://www.education.alberta.ca/. The use of these materials by the end user is done without any affiliation with or endorsement by the Government of Alberta. Reliance upon the end user's use of these materials is at the risk of the end user."

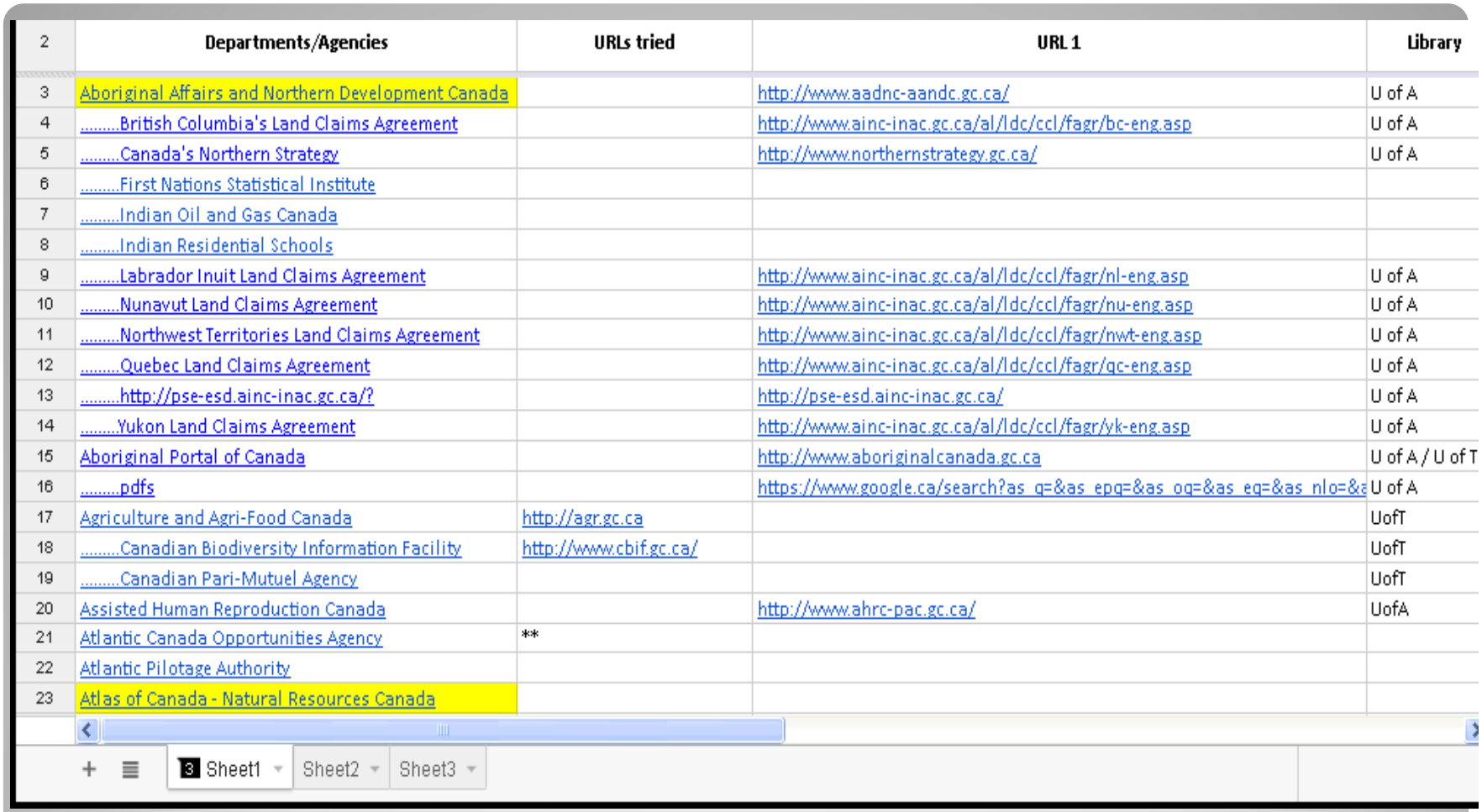
March, 2013

- About 400 seeds
- Content disappearing on a daily basis
- Deadline: July 31, 2013

Mission Impossible !

Collaboration

UofA and UofT
COPPUL members
LAC



2	Departments/Agencies	URLs tried	URL 1	Library
3	Aboriginal Affairs and Northern Development Canada		http://www.aadnc-aandc.gc.ca/	U of A
4	.....British Columbia's Land Claims Agreement		http://www.ainc-inac.gc.ca/al/ldc/ccl/fagr/bc-eng.asp	U of A
5	.....Canada's Northern Strategy		http://www.northernstrategy.gc.ca/	U of A
6	.....First Nations Statistical Institute			
7	.....Indian Oil and Gas Canada			
8	.....Indian Residential Schools			
9	.....Labrador Inuit Land Claims Agreement		http://www.ainc-inac.gc.ca/al/ldc/ccl/fagr/nl-eng.asp	U of A
10	.....Nunavut Land Claims Agreement		http://www.ainc-inac.gc.ca/al/ldc/ccl/fagr/nu-eng.asp	U of A
11	.....Northwest Territories Land Claims Agreement		http://www.ainc-inac.gc.ca/al/ldc/ccl/fagr/nwt-eng.asp	U of A
12	.....Quebec Land Claims Agreement		http://www.ainc-inac.gc.ca/al/ldc/ccl/fagr/qc-eng.asp	U of A
13	.....http://pse-esd.ainc-inac.gc.ca/?		http://pse-esd.ainc-inac.gc.ca/	U of A
14	.....Yukon Land Claims Agreement		http://www.ainc-inac.gc.ca/al/ldc/ccl/fagr/yk-eng.asp	U of A
15	Aboriginal Portal of Canada		http://www.aboriginalcanada.gc.ca	U of A / U of T
16	.....pdfs		https://www.google.ca/search?as_q=&as_epq=&as_oq=&as_eq=&as_nlo=&a	U of A
17	Agriculture and Agri-Food Canada	http://agr.gc.ca		UofT
18	.....Canadian Biodiversity Information Facility	http://www.cbif.gc.ca/		UofT
19	.....Canadian Pari-Mutuel Agency			UofT
20	Assisted Human Reproduction Canada		http://www.ahrc-pac.gc.ca/	UofA
21	Atlantic Canada Opportunities Agency	**		
22	Atlantic Pilotage Authority			
23	Atlas of Canada - Natural Resources Canada			

Sheet1 Sheet2 Sheet3

March, 2013
Seed List #1: UofA and UofT

	A	B	C	
1	**What does yellow mean?**	Yellow highlighting was used by U of T and U of A to indicate instutitions that deemed critical to capture and preserve		
2	Ministry of Finance : Results-based planning (documents)	Legislative Assembly		
3	**Departments/Agencies**	**Library**	**URL 1**	
168	North American Agreement on Environmental Cooperation (NAAEC) - Environment Canada	UVIC	http://www.cec.org/Page.asp?PageID=1226&SiteNodeID=	
169	Registry of the Specific Claims Tribunal of Canada	UVIC	http://www.sct-trp.ca/hist/hist_e.htm	
170	Seniors - Human Resources and Skills Development Canada	UVIC	http://seniors.gc.ca/eng/index.php	
171	Species at Risk Public Registry - Environment Canada	UVIC	http://www.sararegistry.gc.ca/default_e.cfm	
172	Staff of the Non-Public Funds, Canadian Forces - National Defence and the Canadian Forces	UVIC	https://www.cfpsa.com/en/Pages/default.aspx	
173	Statistical Survey Operations - Statistics Canada	UVIC		
174	Veterans Affairs Canada	UVIC		
175	Veterans Review and Appeal Board Canada	UVIC	http://www.vrab-tacra.gc.ca/Home-accueil-eng.cfm	
176	Pacific Fisheries Resource Conservation Council	UVIC	http://www.fish.bc.ca/	
177	Canada Research Chairs - Industry Canada	UVIC / UofT	Part of the Industry Canada crawl done by UofT	
178	**How to use this spreadsheet:**			
179	Please enter your institution's name in Column B next to the agencies that you will *intend* to harvest with Archive-IT. Many of us do not have final approval to purchase ArchiveIt so this is just our first draft.			
180	Please locate the main URL for the sites that you intend to harvest and add to URL 1 column. Use the notes field as needed. At this time the file sizes of most of these agencies isn't known.			
181	This checklist was adapted from the UofT and UofA Archive-It working document, please see the following url for the updated crawl list https://docs.google.com/spreadsheet/ccc?key=0AvlHesVQk7rbdHktdmMzUTRRYkdaTnVLNHpnYjAwWUE#gid=0			

May, 2013
Spreadsheet #2: UofA, UofT, some COPPUL members

Conversations with LAC

March, 2013

- Sent letters to LAC for meeting

May, 2013

- Teleconference (UofA, UofT and LAC)
- Lobbied for final harvesting of the entire GoC domain before the proposed GC web convergence plan is implemented
- Lobbied for future harvesting of the GoC domain
- Share and maintain seed list

July, 2013

- Questions for LAC about recommencement of web harvesting gc.ca domain

August, 2013

- Teleconference (UofA, UofT and LAC)
- LAC confirmed the harvesting of the GoC domain

November, 2013

- Presentation at the Gov Info Day (UofT, UofA, LAC)

Government Information Day Roundtable

We’d love to hear your suggestions

Next Steps