Capturing the Web: Web Archiving in Cultural Heritage Institutions

Jackie Dooley, OCLC Research
Jason Kovari, Cornell University
Christie Peterson, George Washington University
Kristen Yarmey, University of Scranton
What is Web Archiving?

Capture, preservation and access of web content

For more, see: http://www.netpreserve.org/web-archiving/overview
WEB ARCHIVING
A Small Sampling of Capture Tools

Jason Kovari  |  Cornell University Library  |  jak473@cornell.edu
Head of Metadata Services  |  Web Archiving Coordinator
RBMS  |  2016.06.23
Tooling - Considerations

(very) Basic:

• Web content type
• Scale of effort
• IT support
• Overall infrastructure
Web Archiving: Activity Overview (selected)

- Management (collection and resource)
- Capture
- Quality assessment
- Discovery
- Preservation

For what do we need tooling?
For what do we have tooling?
Common Tool Vocabulary

- Heritrix
- WARC
- Wayback
2013 NDSA Survey

92 Institutions Responded

• 52% Universities
• 18% State & Federal Government
• 14% Archives

NDSA Survey

2013 Respondents:
• 63% : external services
• 20% : in-house services
• 16% : in-house & external

2013 External services:
- 70% : Archive-It
- 17% : CDL's WAS

Hosted Solutions
Developer: Internet Archive

Bundles open-source tools
Collection administration
Seed administration

<table>
<thead>
<tr>
<th>Seed URL</th>
<th>Group</th>
<th>Status</th>
<th>Frequency</th>
<th>Type</th>
<th>Access</th>
<th>Last Crawl</th>
<th>Captures</th>
<th>Wayback</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.state.gov/g/ses/climate/">http://www.state.gov/g/ses/climate/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Public</td>
<td>Nov 27, 2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.nature.org/initiatives/climatechange/">http://www.nature.org/initiatives/climatechange/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Public</td>
<td>Nov 27, 2015</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Crawl administration
Webrecorder.io

Enter a url here

Record

Preview before recording

Project of: Rhizome
Capture & Access
Local Solutions
Capture & Access

http://machawk1.github.io/wail/
Wget

Command line utility

Since 1996

https://www.gnu.org/software/wget/
Available Tools

IIPC Tools & Software:
http://www.netpreserve.org/web-archiving/tools-and-software
Collection Development for Web Archives

Christie Peterson
The George Washington University
It’s the same!
The records management tradition
What will you find in the Web Site Archives?

The North Carolina State Government Web Site Archives & Access Program (WSAAP) is composed of web sites created by North Carolina's state agencies. The WSAAP currently contains archived web sites back to 1996, and it is updated on an ongoing basis. These sites may be viewed for free by all users. Sites in the WSAAP were collected using an automated web crawler operated by the Internet Archive. (To learn more about the Internet Archive’s Archive-It service, click here.)

Through the WSAAP users can access important state government information that was once available on the web but that may no longer exist there or in any other location. This information includes, but is not limited to:

- Board and Commission meeting minutes;
- publications that were only available on the Web;
- governmental policies;
- historically significant photographs and videos;
- state-wide statistics; and
- social networking sites.

The topics covered by sites in this collection cover all of the diverse areas that government agencies encompass. Moreover, when seen as a whole the WSAAP documents how North Carolina has developed its web presence over time and the different trends in government priorities.

All web sites in the WSAAP are captured at least on an annual basis. Most state agency web sites are captured quarterly or more frequently. For more information on the how we select and collect sites, please see Our Web Site Archives & Access Program: History and Procedures.

Finally, users are reminded that the materials included in the WSAAP are made available for use in research, teaching, and private study, pursuant to the U.S. Copyright law. Whenever materials from our collection are used in a publication or other product, we request that the copy carry a credit line stating “Courtesy of the Department of Natural and Cultural Resources.”
The rare books & manuscripts tradition
Peterson’s 3 categories of collection development strategies

- Provenance-based
- Subject-based
- Event-based
Provenance-based web collecting
Canadian Political Parties and Political Interest Groups

Collected by: University of Toronto

Archived since: Oct, 2005

Description: Canadian Political Parties and Political Interest Groups will archive the websites of all of the national Canadian political parties, and a number of special interest groups across the political spectrum.

Subject: Politics & Elections
Collector: University of Toronto

Narrow Your Results

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here

Page 1 of 1 (74 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Cosmopolitan Party of Canada
URL: http://agoracosmopolite.com/
Captured 59 times between Oct 4, 2005 and Nov 9, 2012
Subject: Cosmopolitan Party of Canada

Title: Bloc québécois
URL: http://blocquebecois.org/
Captured 32 times between Feb 1, 2009 and May 8, 2016

https://archive-it.org/collections/227
Canadian Political Parties and Political Interest Groups

Collected by: University of Toronto

Archived since: Oct, 2005

Description: Canadian Political Parties and Political Interest Groups will archive the websites of all of the national Canadian political parties, and a number of special interest groups across the political spectrum.

Subject: Politics & Elections

Collector: University of Toronto
Welcome to our Website

The CHP is Canada's only pro-life, pro-family federal political party, and the only federal party that endorses the principles of the Preamble to the Charter of Rights and Freedoms in the Canadian Constitution: 'Canada was founded upon principles that recognize the supremacy of God and the rule of law.'

Please plan to attend the October National Convention in Winnipeg!
(Registration Form)
(Note: Completed Registration Forms may be faxed to David Reimer at 204-346-3861)
(Tentative Schedule)
(Venue: "Victoria Inn")
Grassroots Activist School

Note: Three new links have been added:
How'd They Vote? (Check your MP's voting record)
Canadian Forum (Please participate in this Christian Political forum)
Skeena-Duilksley Valley, B.C.
Welcome to the Elections Canada Advisory Committee. Our mission is to demonstrate excellence and leadership in electoral matters. Through the valued efforts of staff members and election officers, the agency serves the needs of electors and legislators alike in an innovative, cost-effective manner.

Marijuana Party Whip Capitalizes Again

...but bureaucratic hurdles are STILL making life tough for fringe parties!

Following Elections Canada Conferences of Friday September 23 2005, Marijuana Party Whip Carol Taylor reports that the two topics originally on the agenda remain in treatment with the support of not only delegates from other parties, but most encouragingly, from the conference chair herself. See the full story.

Cet article en français

Featured Articles:

Become a Marijuana Party Candidate | Participate in the Democratic Process.

Pot Prohibition History
Web Archives for Longitudinal Knowledge (WALK) Portal

This website is home to the Web Archives for Longitudinal Knowledge (WALK) Project, an envisioned Canadian national Web Archiving portal. Spearheaded by the University of Waterloo, York University, and the University of Alberta, we plan to bring together interested Canadian partners to provide access to their collections.

Currently, this is a prototype site providing access to one such archive, the University of Toronto’s Canadian Political Parties and Political Interest Groups collection. This website allows you to search content from 50 political parties and political interest groups, from October 2005 to March 2015.

Curious how the Liberal Party of Canada responded to the 2008 financial crisis (a search for “recession” in 2008, liberal.ca)? How the Canadian Centre for Policy Alternatives reacted to Michael Ignatieff? Now you can check it all out.

Options include:
- Basic keyword searching [Example: "Rob Ford", only Liberal.ca]
- Graphing trends over time [Example: Liberal Opposition Leaders, 2005-2016]
- Advanced search, including words in proximity to each other [Example: environmental and tax within 25 words of each other]

Below, here are all of the links for the entire time period, visualized below.
Using Extracted Names to Explore Web Archives

6 AUGUST 2015

Yesterday morning, I used Mathematica’s new geographic processes to use our Warbase NER output in order to generate maps based on web archived locations – I then wondered what we could do with the frequency of individuals.

One challenge I have often encountered with NER output is the “what now?” question. We generate fantastic lists of frequently appearing individuals, organizations, and people, but apart from exploring tables or finding use for network analysis, I haven’t been terribly compelled by this (one exception is the Trading Consequences ‘location cloud’ visualization, which we’re currently trying to rip off/borrow from).

It should be no surprise that in 2009, the prominent NDP leader Jack Layton was the most frequent person mentioned on their site.

Mathematica has powerful integration with Wolfram Alpha’s surrounding databases, which can yield amounts of information as...
Subject-based web collecting
Legal Blawgs Web Archive

Collection Period:
March 1, 2007 to Present (ongoing archive).

This is a selective collection of authoritative sites associated with American Bar Association approved law schools, research institutes, think tanks, and other expertise-based organizations. These sites contain unique, born digital content. These blogs contain journal-style entries, articles and essays, discussions, and comments on emerging legal issues, national and international.
About this Collection

Collection Period:
March 1, 2007 to Present (ongoing archive).

This is a selective collection of authoritative sites (associated with American Bar Association approved law schools, research institutes, think tanks, and other expertise-based organizations) that contain unique, born digital content. These blogs contain journal-style entries, articles and essays, discussions, and comments on emerging legal issues, national and international. Additional sites are selected incrementally.
Human Rights
Collected by: Columbia University Libraries
Archived since: May, 2008

Description: An initiative of CULs Center for Human Rights Documentation and Research, the Human Rights Web Archive is an effort to preserve and ensure access to freely available human rights resources created mainly by non-governmental organisations, national human rights institutions, and individuals.

Subject: Society & Culture, Human rights, Non-governmental organizations, Human rights workers, National human rights institutions, Web archives

Creator: Columbia University Libraries, Center for Human Rights Documentation and Research
Collaborators: Columbia University Libraries, Center for Human Rights Documentation and Research

Narrow Your Results

Enter search terms here

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived web pages.

Page 1 of 8 (717 Total Results)

Title: Women’s Initiatives for Gender Justice
URL: http://4genderjustice.org/

Description: "The Women’s Initiatives for Gender Justice (Women's Initiatives) is an international women’s human rights organisation that advocates for gender justice through the International Criminal Court (ICC) and through domestic mechanisms, including peace negotiations and justice processes. We work with women most affected by the conflict situations under investigation by the ICC, and collaborate with more than 8,000 grassroots partners, activists and members across multiple armed conflicts."
Captured 6 times between Jul 1, 2015 and Apr 4, 2016
Videos: 33 Videos Captured

Subject: Human rights, Women’s rights, Sex discrimination against women
Human Rights

Collected by: Columbia University Libraries

Archived since: May, 2008

Description: An initiative of CUL’s Center for Human Rights Documentation and Research, the Human Rights Web Archive is an effort to preserve and ensure access to freely available human rights resources created mainly by non-governmental organizations, national human rights institutions, and individuals.

Subject: Society & Culture, Human rights, Non-governmental organizations, Human rights workers, National human rights institutions, Web archives

Creator: Columbia University Libraries, Center for Human Rights Documentation and Research

Collector: Columbia University Libraries, Center for Human Rights Documentation and Research
China's Olympic Torch relay set to continue through Tibet

Following the conclusion of a meeting of the International Olympic Committee (IOC) in Athens today, IOC Chairman Jacques Rogge confirmed that the Olympic Torch will still ...

ICT delivers UN statement on illegal killings by security forces in Tibet

On 3 June, ICT delivered an oral statement at the UN Human Rights Council in Geneva on the use of extrajudicial and summary executions in Tibet ....

U.S. Senators have added to the Congressional concern about the situation in Tibet by asking Secretary Rice to press for the release of Tibetan protestors, including Jamyang Kyi

U.S. Senators Gordon Smith (R-OR), John Kerry (D-MA), and Russell Feingold (D-WI) have written to Secretary Condoleezza Rice expressing...
الوعي المصري

مدونة الوعي المصري - على مزاج صادقته التي مش تيسلع عند حد

روضتي الفضول دم عن طريق الدخل
مصممها برسالة ثانية أتفرجها معا
فؤاد للشهيد شديد مجهول من ذات الألوترا والمحرر صغير في القرون

Wednesday, 26 January 2011

the martyr of Snez

Video: شهيد السويس من داخل المشتركة

An Independent Egyptian Blog
Blogger in Chief
Wael Abbas

destinia.com
You are viewing an archived web page, collected at the request of Columbia University Libraries using Archive-It. This page was captured on 2016-04-04 20:22:57 and is part of the Human Rights collection. The information on this web page may be out of date. See All versions of this archived page. Metadata

Cop Watch NYC exists to help New Yorkers document and expose unprofessional police behavior and hold the police accountable. The site can be used as a resource for individuals, organizations and teams of concerned citizens who.

See Something? Say Something.
If you've witnessed discriminatory or unjust policing, report it.

Join us.
Help your community deter police misconduct.

Holding the Police Accountable

Crooke avenue
After a nice quiet and peaceful BBQ ended literally 100s of police showed up in battle mode. I assume because of the large amount of people that attended the BBQ. They went crazy showing physical altercations, kicking, punching, I was pushed, then grabbed, body slammed, punched in back of my head, stepped on, and maced. I was arrested and charged with obstruction of justice and resisting arrest. I go back to court October 23 2014. Hopefully and most likely they dismiss it.
Event-based web collecting
The End of Term Web Archive captures and saves U.S. Government websites at the end of presidential administrations. Beginning in 2009, the EOT has thus far preserved websites from administration changes in 2009 and 2012, and is currently preparing for the 2016 electoral season.

Archive Scope

The End of Term Web Archive contains federal government websites (.gov, .mil, etc) in the Legislative, Executive, or Judicial branches of the government. Websites that were at risk of changing (e.g., whitehouse.gov) or disappearing altogether during government transitions were captured. Local or state government websites, or any other site not part of the federal government domain were out of scope.
Collection
September 11, 2001 Web Archive

Featured Content

About this Collection

Collection Period:
September 11, 2001 to December 1, 2001.

The September 11, 2001 Web Archive preserves the web expressions of individuals, groups, the press and institutions in the United States and from around the world in the aftermath of the attacks in the United States on September 11, 2001. The selected websites are comprised broadly of United States and non-United States government sites, press, corporate/business, portal, charity/civic, advocacy/interest, religious, school/educational, individual/volunteer, professional organizations sites, and others.
About this Collection

Collection Period:
September 11, 2001 to December 1, 2001.

The September 11, 2001 Web Archive preserves the web expressions of individuals, groups, the press and institutions in the United States and from around the world in the aftermath of the attacks in the United States on September 11, 2001. The selected websites are comprised broadly of United States and non-United States government sites, press, corporate/business, portal, charity/civic, advocacy/interest, religious, school/educational, individual/volunteer, professional organizations sites, and others.
Internet Archive Global Events

Archive-It Partner Since: Feb, 2006
Organization Type: Other Institutions
Organization URL: http://www.archive-it.org
Description: These collections of global events have been created by the Archive-It team in conjunction with curators and subject matter experts from institutions around the world.

Narrow Your Results

Sites and collections from this organization are listed below. Narrow your results at left, or enter a search query below to find a collection, site, specific URL or to search the text of archived webpages.

Enter search terms here

Collections  Sites  Search Page Text

#blacklivesmatter Web Archive

Archived since: Aug, 2014
Description: In August of 2014, the Archive-It team put out a call for URLs related to the August 9th shooting of Michael Brown in Ferguson, MO. Soon after we were joined by Archive-It partner
Earthquake in Haiti

Collected by: Internet Archive Global Events

Archived since: Jan, 2010

Description: This collection is currently documenting the events of the January 2010 earthquake in Haiti and the aftermath, including the rescue efforts from around the world and the stories and circumstances of the Haitian people. Archive-It partners Library of Congress, Bibliothèque Nationale de France, Virginia Tech and CTRnet, and University of Texas Libraries have all contributed websites for this collection.

Subject: Spontaneous Events, Society & Culture, earthquake, poverty, global aid, infrastructure, survivors

Date: 2010-01-21 00:00:00
#blacklivesmatter Web Archive

Collected by: Internet Archive Global Events

 Archived since: Aug, 2014

Description: In August of 2014, the Archive-It team put out a call for URLs related to the August 9th shooting of Michael Brown in Ferguson, MO. Soon after we were joined by Archive-It partner Washington University in St. Louis, MO who contributed lists of URLs and solicited URLs from communities in and around Ferguson. Maryland Institute for Technology in the Humanities contributed a significant number of URLs extracted from tweets about Michael Brown and Ferguson. In December we expanded the collection to include content related to the July 17th chokehold death of Eric Garner in Staten Island, NY, and the reactions around the country to the acquittal of the police officers in both cases. The collection then progressed further to reflect the movement building in reaction to the perceived unjust police violence against people of color. Today the #blacklivesmatter Web Archive contains news articles, blogs, social media, and other websites related to Ferguson, Michael Brown, Eric Garner, protests and demonstrations around the country, and seeks to document the movement as it grows and evolves. Seed URL nominations are continuing to be accepted. If you would like to contribute to the collection please copy and paste the following URL into your browser:

https://docs.google.com/forms/d/1uTyINjaFgLmzirxqGzGbs8avb5_1xJJ2ostRhvxXo/viewform

Note: Many of these sites contain videos, some of which may not play back in the page. To view archived videos click on the Videos link in the yellow banner at the top of the archived web pag
A Ferguson Twitter Archive

August 30, 2014

If you are interested in an update about where/how to get the data after reading this see here.

Much has been written about the significance of Twitter as the recent events in Ferguson echoed round the Web, the country, and the world. I happened to be at the Society of American Archivists meeting 5 days after Michael Brown was killed. During our panel discussion someone asked about the role that archivists should play in documenting the event.

There was wide agreement that Ferguson was a painful reminder of the type of event that archivists working to “interrogate the role of power, ethics, and regulation in information systems” should be documenting. But what to do? Unfortunately we didn’t have time to really discuss exactly how this agreement translated into action.

Fortunately the very next day the Archive-It service run by the Internet Archive announced that they were collecting seed URLs for a Web archive related to Ferguson. It was only then, after also having finally read Zeynep Tufekci's terrific Medium post, that I slapped myself on the forehead … of course, we should try to archive the tweets. Ideally there would be a “we” but the reality was it was just “me”. Still, it seemed worth seeing how much I could get done.

twarc

I had some previous experience archiving tweets related to Aaron Swartz using Twitter’s search API. (Full disclosure: I also worked on the Twitter archiving project...
DocNow is a tool and a community developed around supporting the ethical collection, use, and preservation of social media content.

About

DocNow responds to the public’s use of social media for chronicling historically significant events as well as demand from scholars, students, and archivists, among others, seeking a user-friendly means of collecting and preserving this type of digital content. DocNow has a strong commitment to prioritizing ethical practices when working with social media content, especially in terms of collection and long-term preservation. This commitment extends to Twitter’s notion of honoring user intent and the rights of content creators. The project is a collaborative effort among the University of Maryland, University of California at Riverside, and Washington University in St. Louis. We are extremely grateful for funding from the Andrew W. Mellon Foundation.
Building and archiving event web collections: A focused crawler approach

Mohamed M. G. Farag
Dept. of Computer Science, Virginia Tech
Blacksburg, VA 24061 USA
mmagdy@vt.edu

Edward A. Fox
Dept. of Computer Science, Virginia Tech
Blacksburg, VA 24061 USA
fox@vt.edu

ABSTRACT
In this paper, we present a new approach for building and archiving web collections about events. Our approach combines the traditional focused crawling technique with event modeling and representation.

1. INTRODUCTION
In the NSF funded Integrated Digital Event Archiving and Library (IDEAL) project [1] we are building an integrated system to collect, archive, organize, analyze (identify topics, summarize), access (search, browse), and visualize webpages and tweets about real world events (disasters as well as community and government activities). Event archiving is different from Domain/Site-based or Topic-based archiving. The first involves archiving a specific domain/website with all or some of the underlying subdomains/structure. The second covers a given number of webpages related to a user-defined topic.

We have identified and employed three approaches for archiving webpages about events:

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>No. of Seeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama University Shooting</td>
<td>116</td>
</tr>
<tr>
<td>April 16 Archive</td>
<td>88</td>
</tr>
<tr>
<td>Chile Earthquake</td>
<td>19</td>
</tr>
<tr>
<td>Nevada air race crash</td>
<td>64</td>
</tr>
<tr>
<td>China Floods</td>
<td>60</td>
</tr>
<tr>
<td>Encephalitis (India)</td>
<td>59</td>
</tr>
<tr>
<td>Hurricane Irene</td>
<td>70</td>
</tr>
</tbody>
</table>

Table 2. Sample of tweet collections

<table>
<thead>
<tr>
<th>Collection</th>
<th>Keywords/Hashtags</th>
<th># Tweets</th>
</tr>
</thead>
</table>

A few words about format
VR Works—For Your Business!

You've got customers to serve and deadlines to meet. You need skilled, dependable people. We're VR—Vocational Rehabilitation—and we refer skilled, trained applicants to businesses all across North Carolina.

Last year NC employers hired over 10,000 individuals who'd been served by VR.

To these employees, we provided training or education. To many we provided on-the-job training and job-coaching, at no cost to their employers.

Now these employees are bringing to their jobs a wide range of skills and experience.

VR can help you cut hiring costs and reduce training costs.

- You get customized consultation and follow-up. Services may cover any employee and such problems as drug and alcohol abuse.
- We can help you retain workers who acquire a disability.
- When you hire a VR-referred employee, you may qualify for tax credits.
- We can help you determine if and how the ADA applies to you.

The bottom line: let VR bring you the skills your business needs.

Some Jobs Filled By VR Consumers:
- Assistant managers
- Auto mechanics
- Auto body repairers
- Barbers
- Bookkeepers
- Brick masons
- Computer operators
- Computer programmers
- Clerical workers
- Construction workers
- Customer service reps
- Day care workers
- Dispatchers
- Electricians
- Electrician's helpers
- Forklift operators
- Groomers
- Landscapers
- Medical assistants
- Medical transcriptionists
- Medical technicians
- Mechanics
- Paralegals
- Pharmacy technicians
- Police officers
- Product inspectors
- Programmers
- Sales representatives
- Secretaries
- Transportation workers
- Truck drivers
- Welders
- Wire line workers
In conclusion
Developing Best Practices to Meet User Needs

The OCLC Research Library Partnership Web Archiving Metadata Working Group

Jackie Dooley
Program Officer
OCLC Research
THE PROBLEM

What happened?
The page you visited could not be found.

What can I do?
If you're a visitor to this site, please try back a bit later.

If you are the owner of this site, please visit Typepad Status for network updates or open a ticket from within your account. If you are unable to open a ticket, please contact us via support@typepad.com.
• Archived websites often are *not easily discoverable* via search engines or library and archives catalogs and finding aid systems, *which inhibits use.*

• *Absence of community best practices for descriptive metadata was the most widely-shared web archiving challenging* identified in a 2015 survey of the OCLC Research Library Partnership.
OUR APPROACH

Web Archiving Metadata Working Group

CHARGE: The OCLC Research Library Partnership Web Archiving Metadata Working Group will evaluate existing and emerging approaches to descriptive metadata for archived websites and will recommend best practices to meet user needs and to ensure discoverability and consistency.

The Problem

Archived websites often are not easily discoverable via search engines or library and archives catalogs and finding aid systems, which inhibits use.
Working Group Charge

The OCLC Research Library Partnership Web Archiving Metadata Working Group will evaluate existing and emerging approaches to descriptive metadata for archived websites and will recommend best practices to meet user needs and to ensure discoverability and consistency.

http://www.oclc.org/research/themes/research-collections/wam.html
Projected Outcomes

• *Best practices for descriptive metadata* that bridge existing approaches (bibliographic and archival).

• A *report on user needs* will inform community-wide understanding of documented needs and behaviors as evidence to underlie the metadata best practices.
WORKING GROUP MEMBERS
Trevor Alvord, Brigham Young University
Alexis Antracoli, Princeton University
Penny Baker, Clark Art Institute
Kate Bowers, Harvard University
Lori Dedeyan, UC Berkeley
Evan Echols, University of Delaware
Karen Stoll Farrell, Indiana University
Rick Fitzgerald, Library of Congress
Ben Goldman, Penn State University
Rebecca Guenther, consultant
Claudia Horning, UCLA
Chad Hutchens, University of Wyoming
Deborah Kempe, Frick Museum
Tammi Kim, University of Nevada Las Vegas

Jason Kovari, Cornell University
Rosalie Lack, California Digital Library
Eilidh MacGlone, National Library of Scotland
Matthew McKinley, UC Irvine
Allison O’Dell, University of Florida
Joy Panigabutra-Roberts, University of Cairo
Dallas Pillen, University of Michigan
Lily Pregill, Frick Museum
Mary Samouelian, Harvard Business School
Aislinn Sotello, UC San Diego
Alex Thurman, Columbia University
Jessica Venlet, MIT
Olga Virakhovskaya, University of Michigan

Jackie Dooley & Dennis Massie, OCLC Research
World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists, Policy, November's W3 news, Frequently Asked Questions.

- What's out there?
  - Pointers to the world's online information, subjects, W3 servers, etc.
- Help
  - on the browser you are using
- Software Products
  - A list of W3 project components and their current state. (e.g. Line Mode, X11 Viper, NeXTStep, Servers, Tools, Mailrobot, Library)
- Technical
  - Details of protocols, formats, program internals etc
- Bibliography
• Call for volunteers, November 2015
• WAM launched, January 2016
• 2016
  • Desk research
  • Study existing practices
  • Obtain ongoing community input
• 2017
  • Publish!
DESK RESEARCH
• Literature review
  • User needs
  • Metadata needs from user & library/archives perspectives
  • Tools with metadata components

• Abstract readings

• Synthesize key concepts
STUDY USER NEEDS
Why bother?

• A necessary prelude to development of metadata best practices
• Take an approach based in reality of user needs and behaviors
• Users don’t necessarily utilize the usual discovery tools to locate archived websites
• Many users don’t realize that libraries & archives do web archiving for user access
Methodology

• Read
• Abstract
• Synthesize
• Write

[Rinse and repeat for metadata]
STUDY METADATA NEEDS

About the Team

http://www.dolekemp96.org
Methodology

• *Analyze* descriptive metadata standards & local guidelines
• *Evaluate* existing records “in the wild”
• *Compare* data elements
• *Differentiate* bibliographic & archival approaches
• *Identify* issues specific to web archiving
• *Incorporate* findings from user needs and metadata literature reviews
• *Prepare* best practices document
Descriptive standards under review

- *Describing Archives: A Content Standard* (SAA)
Local guidelines under review

• Archive-It
• Harvard
• Library of Congress
• New York Art Resources Consortium (NYARC)
• University of Illinois
• University of Michigan
• University of Texas
Analysis of local guidelines

• Data elements
  • Definitions?
  • Most common?
  • Core?
    • Same concept, different elements?
    • Same element, different concepts?

• Associated structure & content standards

• Presence/absence of elements specified in descriptive standards
Analysis of existing records

- Data sources
  - WorldCat
  - ArchiveGrid
  - Archive-It

- Types of record
  - Bibliographic: MARC, Dublin Core, etc.
  - Archival: MARC, finding aids, local

- Identify patterns and inconsistencies in local practice
Sample web-specific issue #1

• Institution responsible for the crawl is variably tagged as ...
  • Creator/contributor
  • Repository
  • Publisher
  • Selector
  • Statement of responsibility
Sample web-specific issue #2

- Is the site owner the ...
  - Publisher?
  - Creator?

P.S. Are websites published?
Sample web-specific issue #3

• Useful dates?
  • Capture
  • Beginning/end of site’s existence
  • Content
  • Copyright
What are your web-specific descriptive issues?
WHAT’S NEXT?

What will the Internet look like in 100 years?

WHAT WILL THE INTERNET LOOK LIKE IN 100 YEARS?

We can’t predict the future with 100% accuracy, whether it is 10 years from now or 100 years from now, but we can look at where today’s technology is headed for a glimpse of what the Internet may be like in the future. At the pace things are moving, it’s possible some of us may even be here to see it turn into reality.

IN 100 YEARS...

It’s highly likely something new and more involved will replace the Internet.
Community awareness & input

• Discussion with constituencies
• Presentations
• Published article in *Journal of Western Archives*
• Listservs
• ...
Produce the two outputs

• Report on user needs
• Best practices for descriptive metadata
• Aiming for early 2017
DISCUSSION
Because what is known must be shared.
Web Archiving with Limited Resources:
A Nickel’s Worth of Free Advice

Accompanied by visual inspiration from the Zaner-Bloser Penmanship Collection

Kristen Yarmey (@kristenyty)
Associate Professor and Digital Services Librarian
University of Scranton
#rbms2016
Capture All the URLs: First Steps in Web Archiving

- [2014 article](#) in *Pennsylvania Libraries: Research & Practice*
- [2015, 2014, 2013](#) presentations

- Collaborators:
  - Alexis Antracoli
  - Steven Duckworth
  - Judy Silva
  - Kristen Yarmey

Big Web, Small Staff: Web Archiving with Limited Resources

- [2015 article](#) in *Archival Outlook*
- 2015 Society of American Archivists presentation

- Collaborators:
  - Kelli Bogan
  - Rebecca Petersen
  - Rachel Taketa
  - Kristen Yarmey

E. L. Brown, [Sample illuminated album page](#), 1937, Horace G. Healey Collection
What do I know about web archiving?
What do I know about web archiving?

Getting started

Tomorrow

is the fool's workday

Getting started
Getting started

Don’t go it alone.
Don’t go it alone.
Archive-It

• Subscription service
• Branch of nonprofit Internet Archive
• Crawls, harvests, and hosts web content, using open source tools and standard formats
• Yearly fees, based on data budget
• Support and training
• User community (465+ organizations)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9 pages</td>
<td>10 pages</td>
<td>7 pages</td>
<td>13 pages</td>
<td>13 pages</td>
<td>17 pages</td>
<td>43 pages</td>
<td>15 pages</td>
<td>2 pages</td>
<td>1 page</td>
<td>4 pages</td>
<td>20 pages</td>
<td>45 pages</td>
<td>55 pages</td>
<td>4 pages</td>
</tr>
</tbody>
</table>

* denotes when page was updated.
Making the case for web archiving

“Show Me”
The Record News

University Receives $6 Million Gift

University Hosts Sixth Annual "Blessing of the Books"

Position Created To Improve Student-Neighbor Relations

Royal News

Biosignals

Billboard Events

Royal Events

Featured Videos
Gather ye rosebuds while ye may.
Old time is still a flying;
And this same flower that smiles today,
To-morrow will be dying.
Records Management Advisory Board
Electronic Records Subcommittee

[an error occurred while processing this directive]

SORRY!
We cannot find that page
Please use the following site map to locate the page you are looking for.

- Information For...
  - Future Students
  - Current Students
  - Faculty & Staff
  - Alumni
  - Parents
  - Community
  - Consumers
- About Us
  - About Us Homepage
  - Office of the President
  - Mission
  - Outcomes & National Recognition
  - Facts About Scranton
  - Campus Map
- Academics
  - Academics Homepage
  - Provost/VP Academic Affairs
  - Undergraduate Programs
  - Graduate Programs
  - Graduate & Continuing Ed
  - Colleges & Departments
Making the case for web archiving
MIDDLE STATES ACCREDITATION ACTIVITIES

The University of Scranton is accredited by the Middle States Commission on Higher Education (MSCHE), one of six regional accreditors within the United States. This institutional accreditation was last reaffirmed in 2008, following the University's successful decennial self study.

Colleges and university accreditation typically follows the following cycle:

**Self Study**, an in-depth institutional analysis that takes place every ten years. The self study requires the institution to review its programs, services, and operations with respect to the fourteen standards for Middle States accreditation, and results in a series of recommendations for improvement. The self study is reviewed by a team of peer reviewers, who can make recommendations of their own to the the institution, and by leadership at MSCHE.
Making the case for web archiving
Selection

2015 INTERNATIONAL CONFERENCE ON ASIAN STUDIES
BRINGING THE WORLD TO NORTHEASTERN PENNSYLVANIA
March 28 – 29, 2015
The University of Scranton, Pennsylvania, USA

You are viewing an archived web page, collected at the request of University of Scranton using Archive-It. This page was captured on 17:10:26 Jan 22, 2015, and is part of the University of Scranton Web collection. The information on this web page may be out of date. See all versions of this archived page at University of Scranton. Date: 160444 Jul 18, 2014, and is part of the University of Scranton Social Media collection. The information on this web page may be out of date. See all versions of this archived page at University of Scranton. Date: 160385 Jul 18, 2014.
Description
Quality Assurance (QA)

Staffing

In conclusion


Penmanship study of Longfellow’s Learn to Labor and to Wait, undated, Horace G. Healey Collection


Penmanship study, undated, Horace G. Healey Collection
Resources

Articles I’ve contributed to:
- Capture All the URLs: First Steps in Web Archiving (Pennsylvania Libraries: Research and Practice, Fall 2014)
- Big Web, Small Staff: Web Archiving with Limited Resources (Archival Outlook, Sept/Oct 2015)

Archive-It
- Archive-It Help
- Archive-It Blog
- 2016 Partner Meeting (August 2 in Atlanta, Georgia)
- Archive-It MidAtlantic User Group (Google Group)

Web Archiving
- International Internet Preservation Consortium (IIPC)
- Society of American Archivists (SAA) Web Archiving Roundtable
- The Signal (LoC digital preservation blog)
- Truman, Web Archiving Environmental Scan (2016)
- NYARC Web Archiving Quality Assurance documentation (2015)
- UNT Quality Assurance Practices in Web Archiving (August 2014)
- Goethals et al, Facing the Challenge of Web Archives Preservation Collaboratively (D-Lib Magazine, May/June 2015)
- Ayala, Web Archiving Bibliography (June 2013)
- Pennock, DPC Technology Watch Report on Web Archiving (March 2013)
Thank you!

Jackie Dooley, OCLC Research
Jason Kovari, Cornell University
Christie Peterson, George Washington University
Kristen Yarmey, University of Scranton