

Collaborative Web Archiving with Ivy Plus / Borrow Direct

Anna Perricci

June 4, 2015

Columbia University

Thanks for joining us today!

Web Resources Archiving Collaboration

Many thanks to the Mellon Foundation

Building collaborations among

- Web archiving communities
- Other research libraries
- Users and potential users of web archives
- Website creators

Andrew W. Mellon Foundation support for CUL web archiving

Grant projects

- Collection Building for Web Resources (2008-2009)
1 FTE: project librarian
- Web Resources Collection Program Development (2009-2012)
3 FTE: 2 web curators, 1 programmer
- **Web Resources Archiving Collaboration (2013-2015)**
2 FTE:
1 project librarian
1 bibliographic assistant

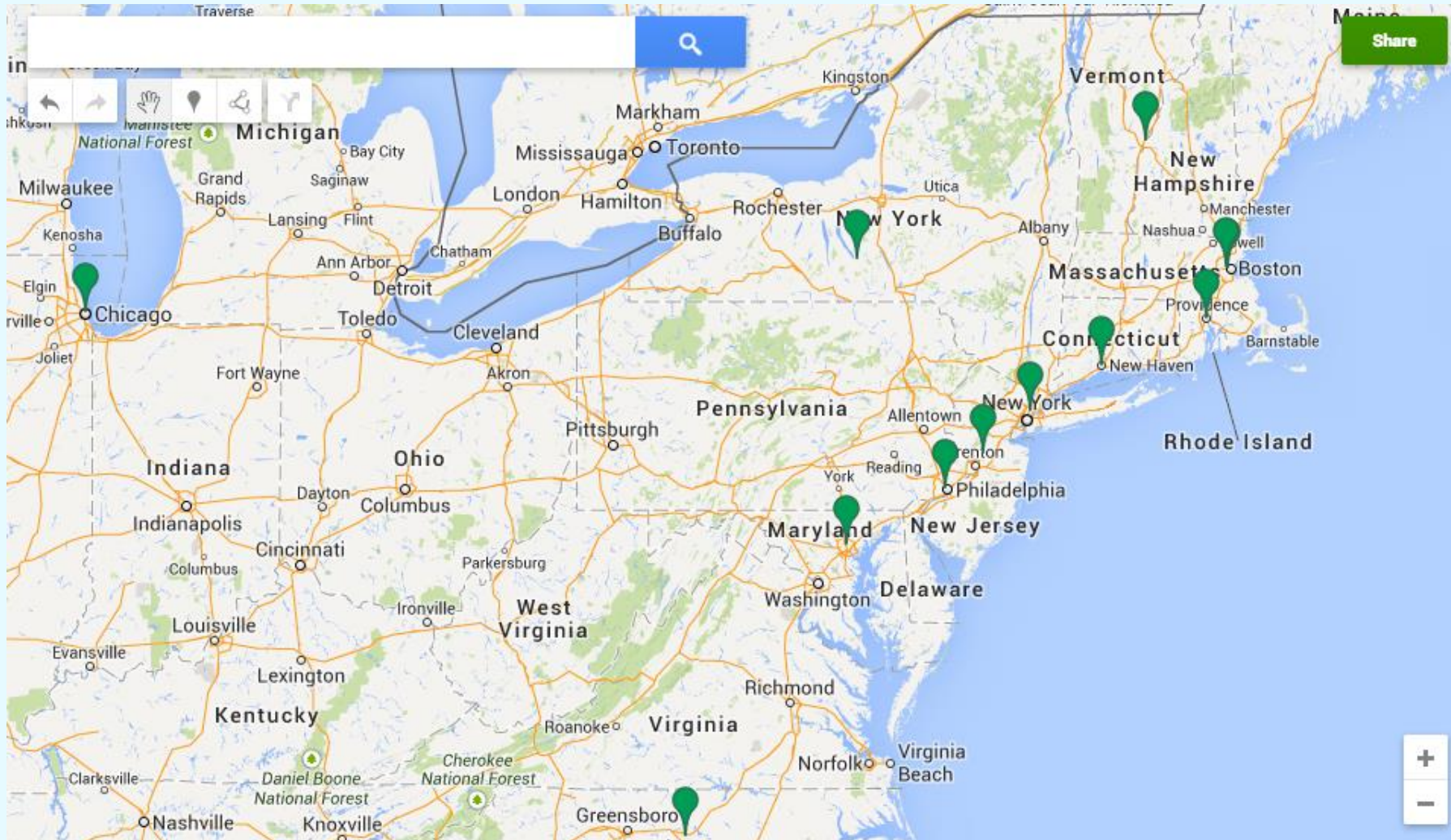
Project elements

- Incentives grants to advance web archiving tools
 - And this conference
- Citation analysis (**corrections/amendments/updates in red**)
 - Humans testing ~~about 1900~~ **2061** URLs from citations in scholarship on Human Rights published in ~~6~~ **9** major journals **in this field** in 2010
 - ~~46%~~ **Around 50%** don't **work or have major content drift/**lead to cited content—now what?
 - Leveraging APIs to determine if cited content in HRWA and/or the Internet Archive; **part of this process involved assistants** looking for **missing** content on live web
- Interviews with scholars to enrich use case development
- Outreach to site creators
- Best practices for site creators
- Collaborative collection building through Ivy Plus / Borrow Direct

Members of Ivy Plus / Borrow Direct

- Brown University
- Columbia University
- Cornell University
- Dartmouth College
- Duke University (new—welcome!)
- Harvard University
- Johns Hopkins University
- MIT
- Princeton University
- Yale University
- University of Chicago
- University Pennsylvania

Ivy Plus



Where we start

Borrow Direct / Ivy Plus



Archive-It

The screenshot shows a web browser window displaying the Archive-It Partner Home page. The browser's address bar shows the URL: https://partner.archive-it.org/archiveit/partner/home.html?accountId=730&cid=62478. The page has a blue header with the Archive-It logo and navigation links: Home, Collections, Crawls, Reports, Access, Research Services, Help Documentation, and Submit a Question. The main content area is titled 'Columbia University Libraries Consortial Collections - Web Archive' and 'Partner Home'. It features two tables: 'Current Subscription (started Jul 1 2014)' and 'All Subscription Periods'. The 'Current Subscription' table shows metrics like Documents Crawled (1,940,193), Subscription Document Budget (13,750,000), Document Budget Used (14.1%), Data Archived (103 GB), Data Budget (1,024 GB), Data Budget Used (10.1%), and Total Active Seeds (210). The 'All Subscription Periods' table shows Documents Crawled (2,877,202) and Data Archived (167 GB). Below these tables is a section for 'Active Collections' with a table listing collections like 'Anna's Art & Music: Example Account', 'Collaborative Architecture, Urbanism, and Sustainability Web Archive (CAUSEWAY)', and 'Contemporary Composers Web Archive (CCWA)', along with their last completed and next scheduled crawl dates. A 'Welcome to Archive-It' sidebar on the right provides an overview of account activity and links to help resources. The footer includes the Internet Archive logo and version information: 'Internet Archive - Archive-It Web UI 4.9-SNAPSHOT-prod-20150417-0005'.

Basis for comparison



76.1" of snow in Providence and Boston got 110.6"

<http://www.weather.com/news/news/new-england-boston-record-snow-tracker>

Seeds & seeds



StaffWeb Home Page x Archive-It: Collaborative Ar... x Massachusetts Smart Grow... x Internet Archive Wayback... x Massachusetts Smart Grow... x

https://partner.archive-it.org/archiveit/partner/collection/seeds.html?selectedTab=SEMIANNUAL&firstResult=0&order=uri+asc&accountId=730&collectionId=4638&cic... capturable

Most Visited Columbia University Columbia University Li... Web Resources Archivi... Archive This

Back to Collection Management **Import Seed Metadata** **Add Seeds** **Bulk Edit** **Run Test Crawl** **Draw Selected Seeds** **Delete**

All (151) Active (144) Inactive (7) One-Time (5) Daily (6) Weekly (8) Monthly (15) Bi-monthly (4) Quarterly (17) Semiannual (81) Annual (8)

81 Record(s)

<input type="checkbox"/>	Settings	URL ↑	Metadata	Status	Frequency	Type	Updated
<input type="checkbox"/>	Settings	http://asechicago.org/	Edit	Active	Semiannual	Default	Dec 23, 2014 11:59 AM
<input type="checkbox"/>	Settings	http://barbaracampagna.com/	Edit	Active	Semiannual	Default	Dec 23, 2014 11:59 AM
<input type="checkbox"/>	Settings	http://buffalocentralterminal.org/	Edit	Active	Semiannual	Default	Dec 23, 2014 11:59 AM
<input type="checkbox"/>	Settings	http://cargocollective.com/pratt/	Add	Active	Semiannual	Default	Mar 6, 2015 3:26 PM
<input type="checkbox"/>	Settings	http://ccmht.org/	Add	Active	Semiannual	Default	Mar 6, 2015 3:26 PM
<input type="checkbox"/>	Settings	http://chicagolakesidedevelopment.com/news/	Add	Active	Semiannual	Default	Mar 6, 2015 3:26 PM
<input type="checkbox"/>	Settings	http://chicagopatterns.com/	Edit	Active	Semiannual	Default	Dec 23, 2014 11:59 AM
<input type="checkbox"/>	Settings	http://designmuseumboston.org/	Add	Active	Semiannual	Default	Mar 6, 2015 3:26 PM
<input type="checkbox"/>	Settings	http://environment.yale.edu/uri/	Edit	Active	Semiannual	Default	Dec 23, 2014 11:59 AM
<input type="checkbox"/>	Settings	http://grayareaphilly.org/	Edit	Active	Semiannual	Default	Dec 23, 2014 11:59 AM
<input type="checkbox"/>	Settings	http://largetlots.org/	Edit	Active	Semiannual	Default	Dec 23, 2014 11:59 AM
<input type="checkbox"/>	Settings	http://ma-smartgrowth.org/	Edit	Active	Semiannual	Default	Jan 28, 2015 12:06 PM
<input type="checkbox"/>	Settings	http://mass-ave.org/	Add	Active	Semiannual	Default	Mar 6, 2015 3:26 PM
<input type="checkbox"/>	Settings	http://newarkriver.wordpress.com/	Edit	Active	Semiannual	Default	Dec 23, 2014 11:59 AM
<input type="checkbox"/>	Settings	http://nhpt.org/	Edit	Active	Semiannual	Default	Jan 28, 2015 12:09 PM
<input type="checkbox"/>	Settings	http://pacny.net/	Edit	Active	Semiannual	Default	Dec 23, 2014 11:59 AM
<input type="checkbox"/>	Settings	http://ppsri.org/	Edit	Active	Semiannual	Default	Dec 23, 2014 11:59 AM
<input type="checkbox"/>	Settings	http://recreationmass.org/	Edit	Active	Semiannual	Default	Dec 23, 2014 11:59 AM

To activate or deactivate a single seed, click the Settings link for that seed. You will then see an Activation Status area where you can select to activate or deactivate the seed.

Activate or Deactivate Multiple Seeds

Select the check box to the left of each seed you would like to activate/deactivate. Then select the "Bulk Edit" option and choose the action you would like (Activate or Deactivate).

Test Crawls

To run a test crawl, select the seeds you would like to crawl by checking the box next to the seed. You can review the seeds you have selected by clicking on the "Selected" tab. When you are ready to start your crawl press the "Run Test Crawl" button.

[Learn more about test crawls.](#)

Frequently Asked Questions

- [What is the difference between Active and Inactive?](#)
- [How can I change my seeds' crawl frequency?](#)

10:04 AM 5/29/2015

Focusing on bright days and growth



**Collaborative
web archiving
pilots projects
supported/furthered
by existing &
developing cohorts
of peers focused on
cooperative collection
development initiatives**



Contemporary Composers Web Archive (CCWA)

By the numbers:

- 11 curators participating
- 56 sites currently available in Archive-It all with MARC records in WorldCat
 - Russell Merritt (music cataloger) collaboratively developed MARC records for composers websites; further cataloging of sites might happen in 2CUL
 - 268,828 URLs and 27 GB archived

Outreach

- SAA presentation on MARC records for CCWA
- Over 30 sites tested for quality by five music librarians; bibliographic assistant on the grant tested all sites in collection

CCWA

- Goal: preserve copies of present and future manifestations of the websites of notable contemporary composers in a secure digital archive to guarantee the continuing availability of these extremely important but potentially ephemeral documents for researchers and scholars seeking to study the careers of contemporary composers
- Selection process so far as been via an analysis of collection data and direct nominations (e.g. faculty)
 - Composers who have a website and scores are collected by 6 or more BD libraries
 - Initially contemporary loosely defined as living or deceased after 1950

Collaborative Architecture, Urbanism and Sustainability Web Archive (CAUSEWAY)

By the numbers:

- Curators from 9 Ivies Plus institutions (up to 20 seeds per institution)
- 144 seed URLs active (over 100 harvested and being released as sites are tested, cataloged and assigned metadata in Archive-It)
- 51 GB of content archived (over 1 million URLs so far)
- Over 80 sites available in Archive-It (over 60 of these sites have MARC records and Dublin Core metadata to facilitate access via Archive-It)

CAUSEWAY themes

- Urban Fabric (e.g. historic preservation, urban renewal, urban preservation)
- Public Space (e.g. parklands, community gardens)
- Community Activism (e.g. historic preservation initiatives associations)
- Each librarian is making nominations focused on the geographic region in which her or his institution is located

Climate change pilot & lessons learned so far

- 156 seeds nominated by at least 27 selectors from 6 institutions

Selectors from a great range of fields:

- Wide variety of area studies
 - Social science
 - Science and environmental science
 - Medical, Law, Special Collections, Preservation
 - Collection Development Associate University Librarians
- A lot of enthusiasm for topic, potential recognized

Climate change pilot & lessons learned so far

- 156 seeds nominated by at least 27 selectors from 6 institutions

Selectors from a great range of fields:

- Wide variety of area studies
 - Social science
 - Science and environmental science
 - Medical, Law, Special Collections, Preservation
 - Collection Development Associate University Librarians
- A lot of enthusiasm for topic, potential recognized

The quick intro to web archiving covered early

- Web archiving entails a multifaceted approach to preserve web-based materials (e.g. websites) and ensure ongoing access to collected content
- The main elements of web archiving are
 - Selection
 - Collection
 - Metadata assignment/cataloging
 - Quality assurance
 - Access
 - Long-term stewardship
- Columbia University Libraries Web Resources Collection Program
 - Engages in all of the above steps plus a policy to ask permission before collecting

Planning / Curation / Selection

From collection development policy to leveraging subject expertise

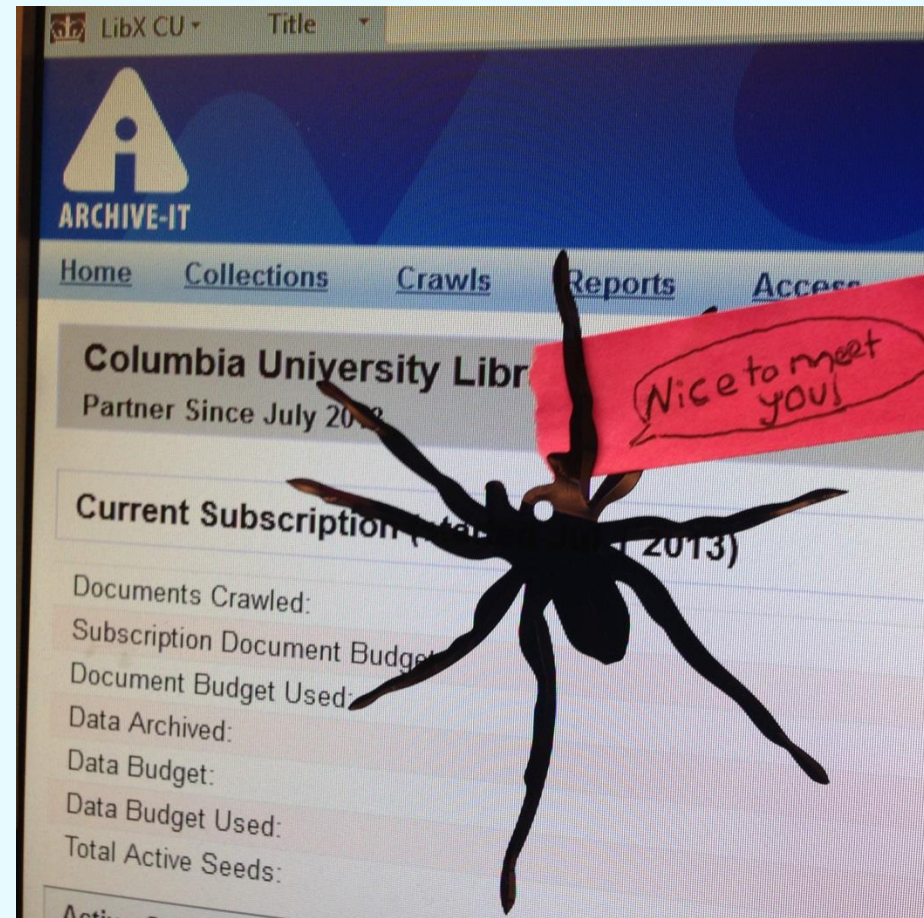
- First questions: what to collect, for whom and why?
- Collection development policy
- Defining themes and goals
 - For CCWA this was an extension of cooperative collecting of contemporary composers' scores
- Engaged selectors



Collection / harvest using software

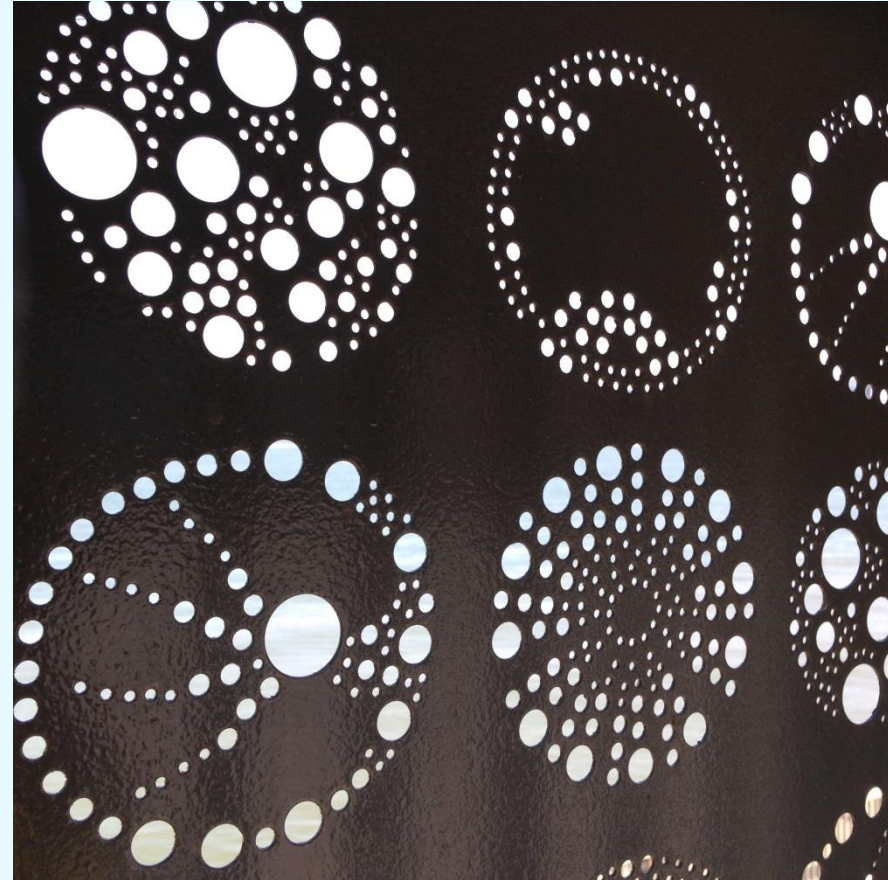
How do we get the content for our collections of archived websites?

For collecting and access we use Archive-It, a subscription-based software as a service of the Internet Archive



Cataloging & Quality Assurance

- Cataloging / Metadata assignment essential to discoverability
 - Recognition of OSMC & NYARC colleagues leading efforts to create a processes for cataloging & increasing the discoverability of archived websites
- Quality assurance testing
 - Many thanks to music librarians who did QA testing last year
 - Sometimes errors can be corrected through patch crawls/troubleshooting
 - Check out NYARC documentation



Cataloging expertise

- Alex Thurman's expertise in cataloging architecture & urban planning sites, and Russell Merritt's years experience of making great records for music resources enables them to make more specific MARC records

The screenshot shows the OCLC WorldCat search results page. The search query is "Collaborative Architecture, Urbanism, and Sustainability Web Archive (CAUSEWAY)". The results are displayed in a list format, showing the first three items. Each item includes a title, author/publisher, language, and database information. The items are:

- Congress for the New Urbanism.** by Congress for the New Urbanism.; Website - Document : Updating website; Continually Updated Resource; Computer File; Language: English; Publisher: Chicago, IL : Congress for the New Urbanism, [1998-]; Database: WorldCat
- Preservation Chicago : citizens advocating for the preservation of Chicago's historic architecture.** by Preservation Chicago.; Website - Document : Updating website; Continually Updated Resource; Computer File; Language: English; Publisher: Chicago, IL : Preservation Chicago, [2002-]; Database: WorldCat
- Massachusetts Smart Growth Alliance.** by Massachusetts Smart Growth Alliance.; Website - Document : Updating website; Continually Updated Resource; Computer File; Language: English; Publisher: Boston, MA : Massachusetts Smart Growth Alliance, [2004-]; Database: WorldCat

- Alex is working with our Bibliographic Assistant, Naeema Akter to put appropriate metadata for better browsing in the Archive-It interface for CAUSEWAY

The screenshot shows the Columbia University Libraries / Information Services CLIO search results page. The search query is "Collaborative Architecture, Urbanism, and Sustainability Web Archive (CAUSEWAY)". The results are displayed in a list format, showing the first four items. Each item includes a title, series title, published location, online status, and format information. The items are:

- Horsefeathers : market, residences**
Series Title: Collaborative Architecture, Urbanism, and Sustainability Web Archive (CAUSEWAY)
Published: Buffalo, NY : Horsefeathers Market & Residences, LLC, [2011-]
Online: Current site Archived site
Format: Online
- Historic Elmira, Inc**
Series Title: Collaborative Architecture, Urbanism, and Sustainability Web Archive (CAUSEWAY)
Published: Elmira, NY : Historic Elmira, Inc. [2010-]
Online: Current site Archived site
Format: Online
- Sustainable Saratoga**
Series Title: Collaborative Architecture, Urbanism, and Sustainability Web Archive (CAUSEWAY)
Published: Saratoga Springs, NY : Sustainable Saratoga, [2010-]
Online: Current site Archived site
Format: Online
- Urban Resources Initiative**
Series Title: Collaborative Architecture, Urbanism, and Sustainability Web Archive (CAUSEWAY)

Getting the CAUSEWAY records in your OPAC via OCLC

- As mentioned, records have been released to WorldCat
- A query can be built for OCLC WorldShare to obtain the MARC records for CAUSEWAY. The records can be delivered in a batch one time or periodically on an ongoing basis

A screenshot of the OCLC WorldShare interface. The left sidebar shows navigation options: Metadata, Admin, Manage Records, Manage Collections (with sub-options: Settings, Library Holdings, Create Standard Collection, Create WorldCat Query Collection, Activity History, Approve Changes to Global Collections), My Library Holdings Only, Collection (with a dropdown), Search, WebArchiving, and My Apps. The main content area is titled 'WebArchiving' and shows a 'Properties' section with the following fields: Collection Name (WebArchiving), Collection ID (customer.4861.7), Provider, Last Updated, Description, Notes (with a 'Show Details' link), WorldCat Selection Criteria (CCWA AND mt: URL, with a note '(36 records selected)' and an example: 'am:blackwellreference AND yr:2011 AND mt:etc'), and Collection Type (One-time Delivery selected, Ongoing Delivery unselected). At the bottom, there are expandable sections for Linking, MARC Records, and Sharing.

WorldCat records for CAUSEWAY sites

The screenshot shows a web browser window displaying a WorldCat record for the Massachusetts Smart Growth Alliance. The browser's address bar shows the URL: www.worldcat.org/title/massachusetts-smart-growth-alliance/oclc/893118510&referer=brief_results. The browser tabs include "StaffWeb Home Page", "Archive-It: Collections - Ac...", and "Massachusetts Smart Growth...".

The record details are as follows:

- Title:** Massachusetts Smart Growth Alliance.
- Author:** [Massachusetts Smart Growth Alliance.](#)
- Publisher:** Boston, MA : Massachusetts Smart Growth Alliance , [2004-]
- Series:** [Collaborative Architecture, Urbanism, and Sustainability Web Archive \(CAUSEWAY\)](#)
- Edition/Format:** Website : Document : Updating website Continually Updated Resource Computer File : English
- Database:** WorldCat
- Summary:** "The Massachusetts Smart Growth Alliance (MSGa) promotes healthy and diverse communities, protects critical environmental resources and working landscapes, advocates for housing and transportation choices, and supports equitable community development and urban reinvestment."
- Rating:** (not yet rated) [0 with reviews - Be the first.](#)
- Subjects:** [Massachusetts Smart Growth Alliance.](#)
[Community development -- Massachusetts.](#)
[Regional planning -- Massachusetts.](#)
[View all subjects](#)
- More like this:** [Similar Items](#)

On the right side of the record, there is a "Nearby libraries" section for ZIP code 02906, listing:

- [Rochambeau Branch](#)
Providence, Rhode Island
02906-3535, United States
< 1 m / km
- [John Carter Brown Library](#)
Providence, Rhode Island
02912, United States
< 1 m / km
- [Brown University](#)
Providence, Rhode Island
02906, United States

Below the record details, there is a "Find a copy online" section with "Links to this item":

- ma-smartgrowth.org
Current site
- wayback.archive-it.org
Archived site

At the bottom, there is a "Find a copy in the library" section with a search bar for "Enter your location: 02906" and a "Find libraries" button. The Windows taskbar at the bottom shows the system tray with the date and time: 9:23 AM, 5/29/2015.

Links in
WorldCat record
will take you to
the archived site
or the live site

The screenshot shows the Internet Archive Wayback Machine search results for the URL <http://ma-smartgrowth.org/>. The page title is "Collaborative Architecture, Urbanism, and Sustainability Web Archive (CAUSEWAY) Web Archive (Columbia University Libraries Consortium Collections)". The search results show 2 captures between Oct 13, 2014 and Mar 24, 2015. The table below summarizes the captures:

Found 2 Captures between Oct 13, 2014 - Mar 24, 2015		
Year	Page Count	Update Date
2014	1 page	Oct 13, 2014 *
2015	1 page	Mar 24, 2015 *

* denotes when page was updated

Links: [Home](#) | [Internet Archive](#)

The screenshot shows the homepage of the Massachusetts Smart Growth Alliance. The main heading is "MASSACHUSETTS SMART GROWTH ALLIANCE" with a sub-heading "GREAT NEIGHBORHOODS". The navigation menu includes "About", "Issues", "Action & Events", "Resources", and "Contact". The main content area features a green banner with the text "Moving Massachusetts Forward" and "A Policy Briefing on Smart Growth and Transportation". Below this is a "DOWNLOAD PDF" button. To the right is a "Stay Connected" form with fields for "First Name", "Last Name", "Email Address", and "Home Zip Code", and a "Submit" button. Below the form are social media icons for Facebook, Google+, Twitter, and RSS. At the bottom, there is a "What's New" section with a "Twitter Timeline" on the right. The "What's New" section lists three items:

- April 9th, 2015**
Transportation
Uncategorized
The MBTA Report: Reform and Revenue
4-9-2015 The Governor's Special Panel to Review the MBTA is exactly right about a fundamental point. It rejects...
- April 2nd, 2015**
Transportation
Uncategorized
The Route to Growth
April 2, 2015 Yesterday, business leaders and local officials stood behind a new report called The Route to...
- December 2nd, 2014**
Placemaking and Zoning
Banker and Tradesman on Zoning Reform

The Twitter Timeline shows two tweets from MA Smart Growth (@MASmartGrowth):

- 19 May: Request for session panels: New Partners for Smart Growth Conference [ow/jy1N0tM](#)
- 4 May: Community benefit districts are governance tool used elsewhere & fit our changing downtowns; CBD bill pending in MA [tinyurl.com/ljme29o](#)

**Please ask your colleagues
in technical services
to contact us!**

**We would love to get these records
into more catalogs**

**We need to know what supporting info
for WorldShare is required**

Access

- Who will use what we have collected and how will they access it?
 - We need more use cases
 - We need to make web archives more accessible to get use cases
- Archive-It is an access system with limitations



This is a server at the Internet Archive!

When? Now! Plus long-term stewardship...

- We need to collect websites before they disappear but we also must ensure their long-term survival and maintain access to them over time
- So far we are saving websites in the WARC file format (the preservation standard) and temporarily relying on the Internet Archive to store the files until a repository framework can be established/chosen

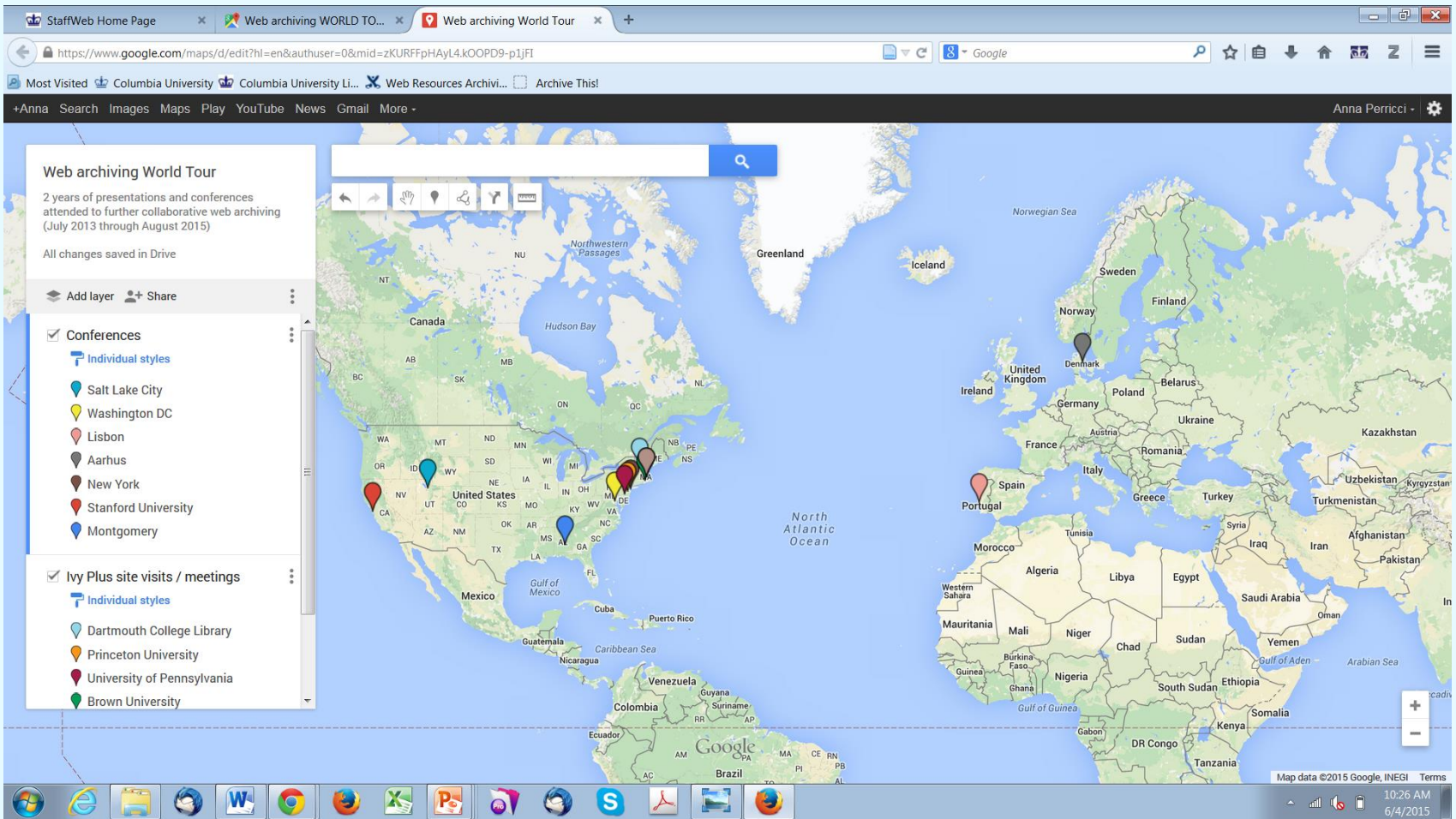


Advocating for curated collections

- Curated collections are
 - focused rather than haphazard
 - guided by a collection development policy
 - informed by skilled selectors
- Re-crawling sites at regular intervals can show patterns and maintain a consistent flow of information
- Because we ask permission, we ignore a file that blocks crawlers (so we get things that the Internet Archive otherwise would pass over as a matter of policy)



Advocating for collaborative web archiving



Some presentations, papers, panels & posters during grant

- Moderated: “Web Archiving: Experiences, Perspectives and Possibilities” held at METRO on 10/20/14
- Presentation (lightning talk): “MARC Records for the Contemporary Composers Web Archive” for the Society of American Archivists annual conference on 8/16/14
URL (via Academic Commons): <http://dx.doi.org/10.7916/D8028Q3S>
- Presentation: “SAA Web Archiving Roundtable Education Needs Assessment Survey Results” for the SAA Web Archiving Roundtable meeting at Society of American Archivists annual conference (co-presented with John Bence) on 8/14/14
- Presentation: “How Collaboration Can Save [More of] the Web: Recent Progress in Collaborative Web Archiving Initiatives” for the METRO Conference 2014 on 1/15/14
- Poster session: “Assessment of the Effectiveness of the Human Rights Web Archive @Columbia University” (co-presented with Pamela Graham) at the ACRL/NY Symposium on 12/6/13
URL (via Academic Commons): <http://dx.doi.org/10.7916/D8BG2KZ9>
- Presentation: “How Collaboration Can Save [More of] the Web: Recent Progress in Collaborative Web Archiving Initiatives” for the Best Practices Exchange on 11/14/13 (with Scott Reed)
URL (via Academic Commons): <http://dx.doi.org/10.7916/D8G73BNK>
- Presentation: “Web Archiving Resource Collaboration” at CrawlCamp held at METRO on 7/17/13

What we've learned about workflows and scale

- Collaborative effort builds the project and new tasks promote professional growth
- Quality Assurance and cataloging integral to process of creating high quality collections of web archives
- Distributing work does not reduce costs

The next 6 months

- Complete remainder of work called for in grant
- Refine and hopefully agree to a model for collaborative collection building through Ivy Plus
 - Shared cost
 - Shared expertise
 - Shared vision for scaling (maintenance and growth)
 - Shared governance
- Contribute to professional organizations to strengthen web archiving efforts nationally and internationally

Credits to as many collaborators as I can fit on this slide

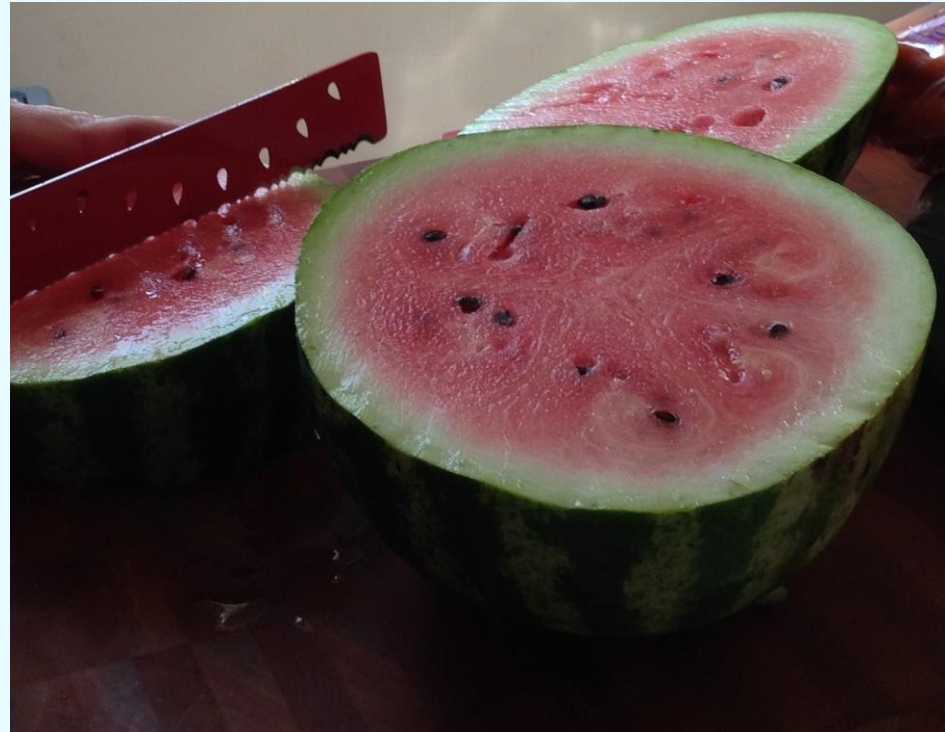
- Bob Wolven, Alex Thurman, Naeema Akter
- Pamela Graham, Kate Harcourt, Melanie Wacker, Christina Harlow
- Talia Jimenez, Stephen Davis, incentives awards oversight panel:
Kris Carpenter, Mark Phillips, Rob Sanderson, Perry Willett, Martin Klein, Jefferson Bailey
- Elizabeth Davis, Russell Merritt & Borrow Direct music librarians
- Carole Ann Fabian, Chris Sala, Ivies Plus Art & Architecture Group
- Borrow Direct / Ivy Plus Associate University Librarians for
Collection Development group
- Climate change selectors at Borrow Direct institutions
- Archive-It staff
- Community for discussion and participation
Including: NYARC, METRO, International Internet Preservation Consortium (IIPC), SAA Web Archiving Roundtable, ARLIS/NA Artist Files SIG

Furthering growth of seedlings



The limit of the comparison to gardening is reaching an end before comparing how various types of fertilizers could or couldn't be a metaphor

Growing web archives



Thanks!

Anna Perricci

alp2198@columbia.edu

@AnnaPerricci

Columbia University Libraries

