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

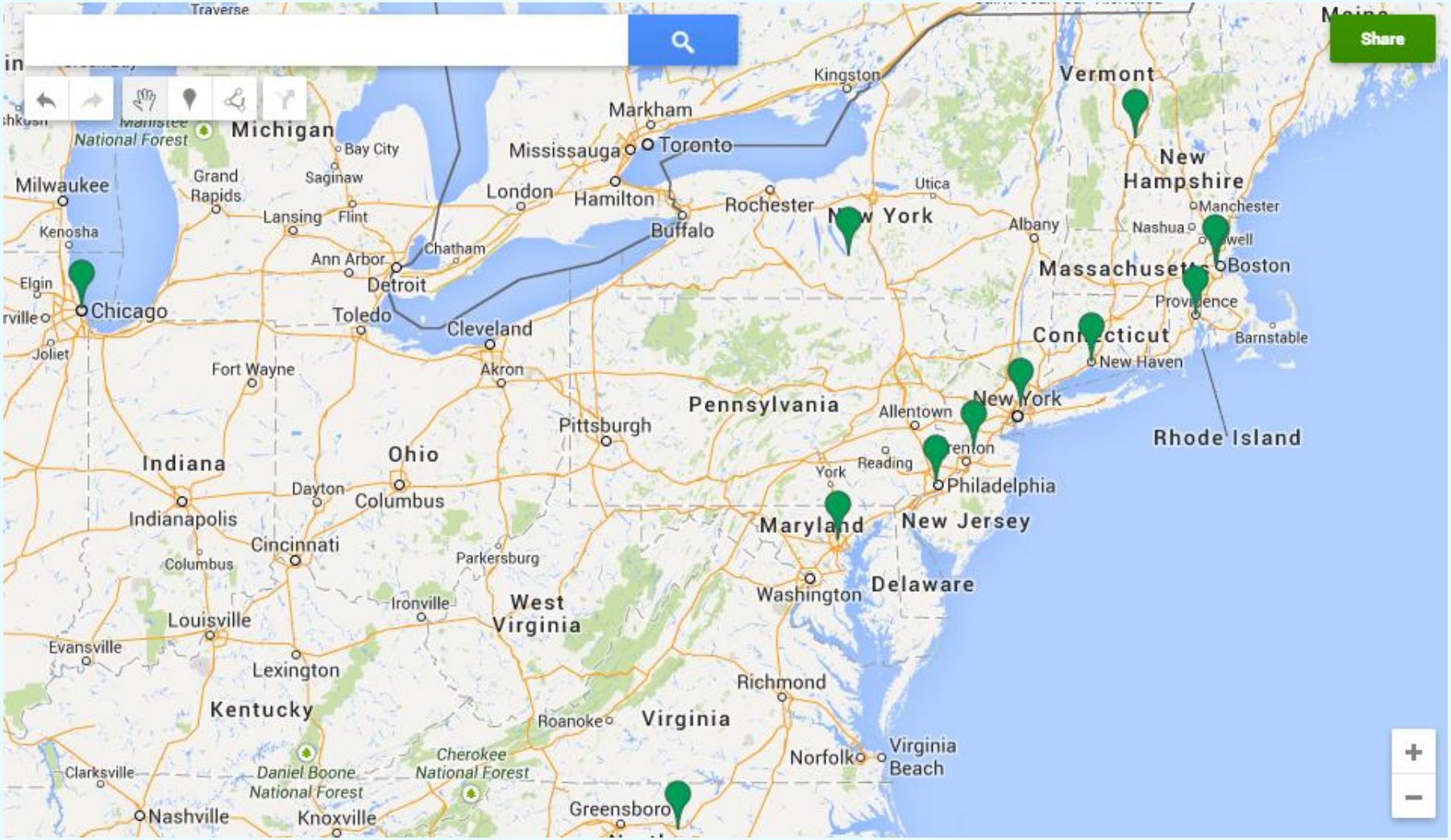
Project elements

- 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

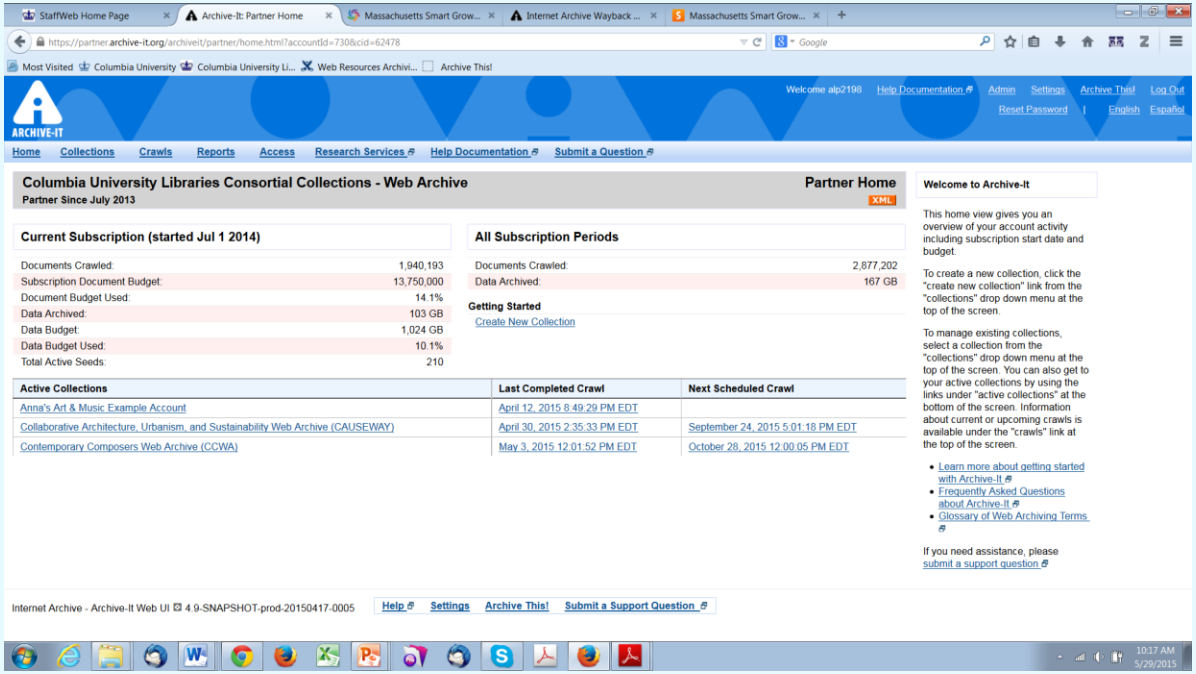
Ivy Plus


Share
Traverse
Kingston
Vermont
Markham
Michigan
National Forest
Bay City
Mississauga
Toronto
New Hampshire
Milwaukee
Grand Rapids
Saginaw
London
Hamilton
Utica
Rochester
Manchester
Lansing
Flint
Buffalo
Albany
Nashua
Kenosha
Ann Arbor
Chatham
Massachusetts
Boston
Elgin
Detroit
Providence
Chicago
Toledo
Cleveland
Connecticut
Barnstable
Joliet
Fort Wayne
Akron
New Haven
Pennsylvania
New York
Allentown
Pittsburgh
Rhode Island
Indiana
Ohio
Reading
York
Philadelphia
Dayton
Columbus
Indianapolis
Maryland
New Jersey
Cincinnati
Parkersburg
Columbus
Washington
Delaware
Ironville
West Virginia
Louisville
Evansville
Lexington
Richmond
Kentucky
Roanoke
Virginia
Cherokee National Forest
Virginia Beach
Norfolk
Clarksville
Daniel Boone National Forest
Nashville
Knoxville
Greensboro

Where we start

Borrow Direct / Ivy Plus

Archive-It

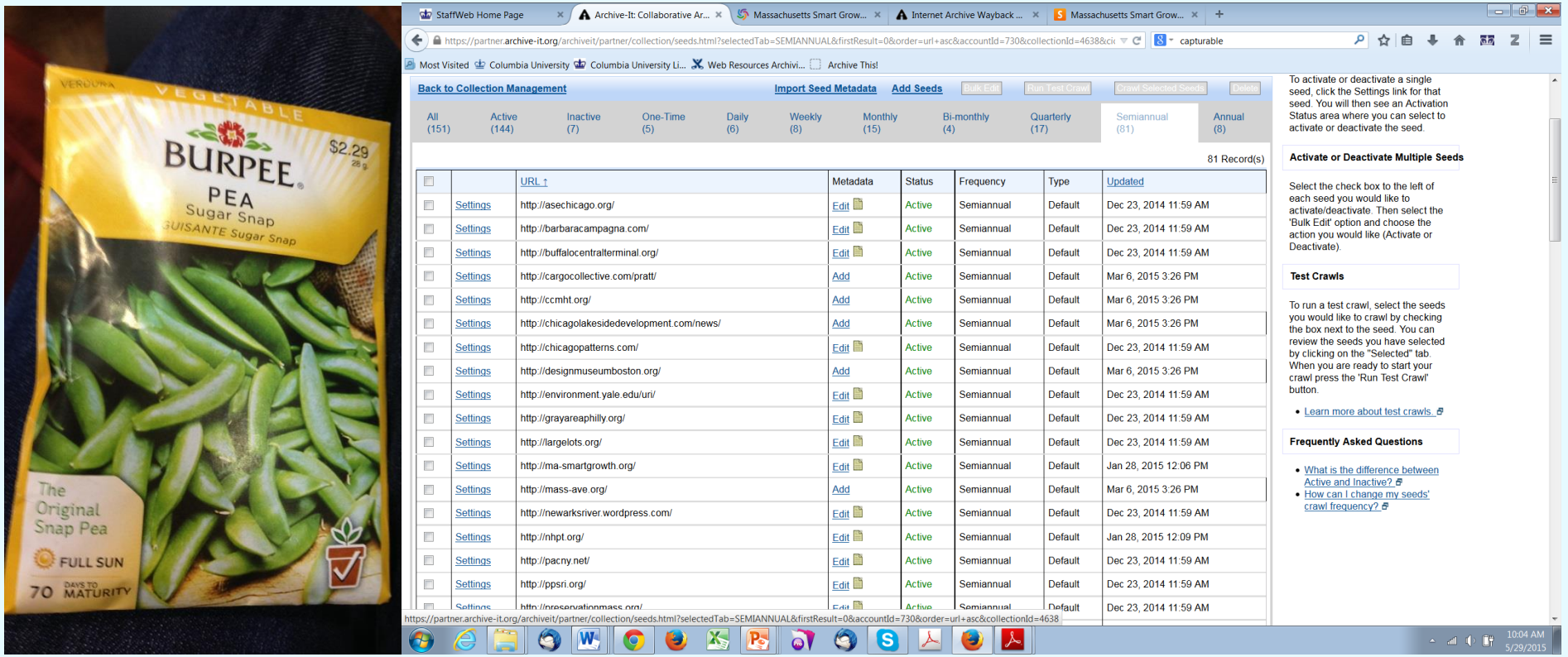


Basis for comparison

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


VEGETABLE
BURPEE
$2.29
PEA
Sugar Snap
The Original Snap Pea
FULL SUN
70 DAYS TO MATURITY
StaffWeb Home Page
Archive-It: Collaborative Ar...
Massachusetts Smart Grow...
Internet Archive Wayback ...
Massachusetts Smart Grow...
Back to Collection Management
Import Seed Metadata
Add Seeds
Bulk Edit
Run Test Crawl
Crawl 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)
URL
Metadata
Status
Frequency
Type
Updated
http://asechicago.org/
http://barbaracampagna.com/
http://buffalocentralterminal.org/
http://cargocollective.com/pratt/
http://ccmht.org/
http://chicagolakesidedevelopment.com/news/
http://chicagopatterns.com/
http://designmuseumboston.org/
http://environment.yale.edu/uri/
http://grayareaphilly.org/
http://largelots.org/
http://ma-smartgrowth.org/
http://mass-ave.org/
http://newarksriver.wordpress.com/
http://nhpt.org/
http://pacny.net/
http://ppsri.org/
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

Photo credit: Anna Perricci

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

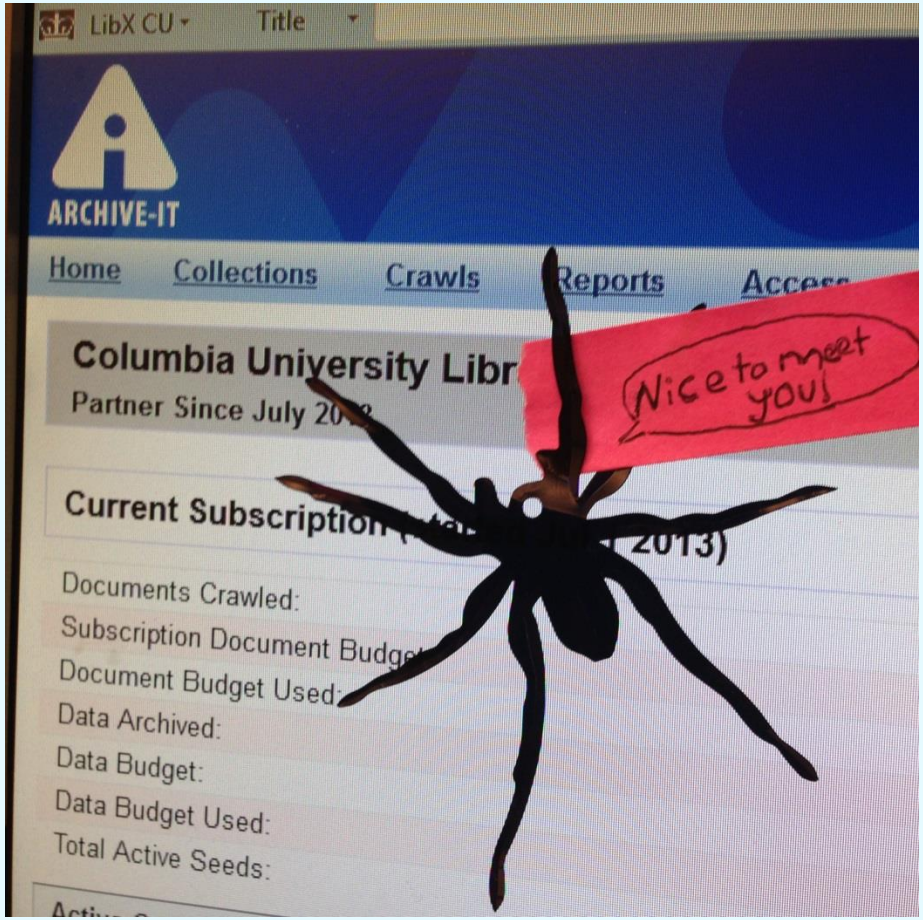


Photo credit: Anna Perricci

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

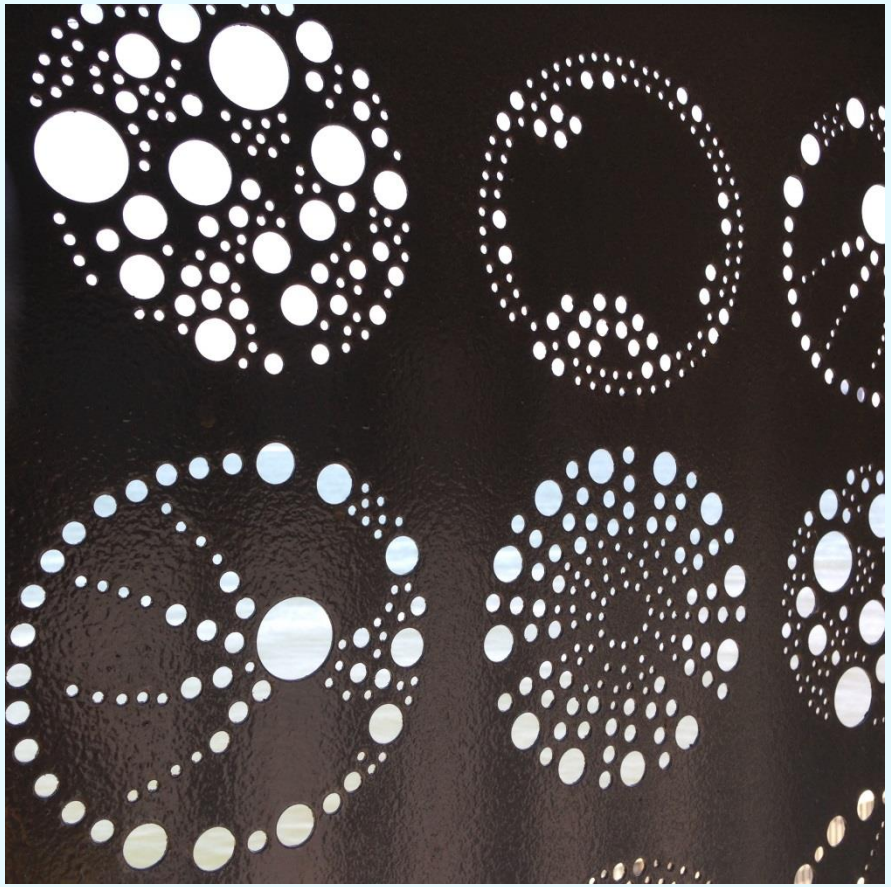
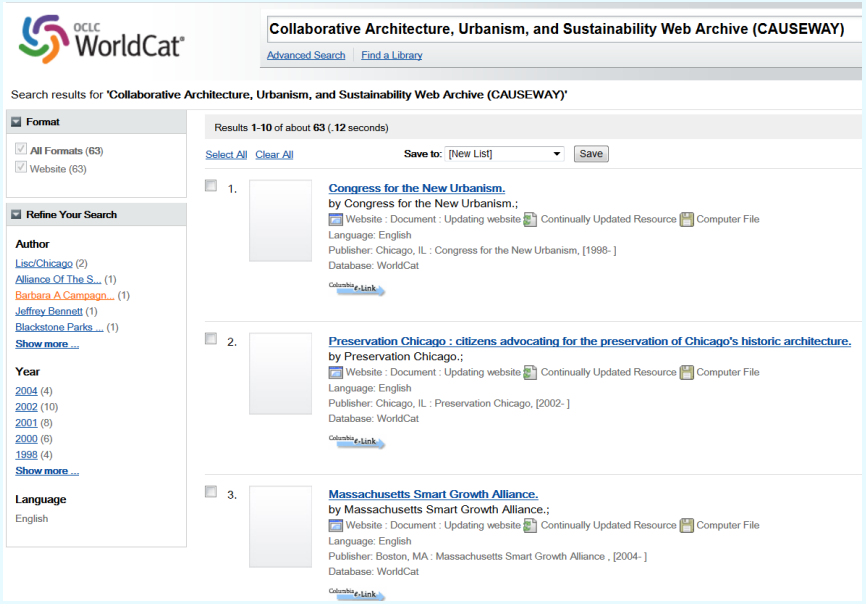


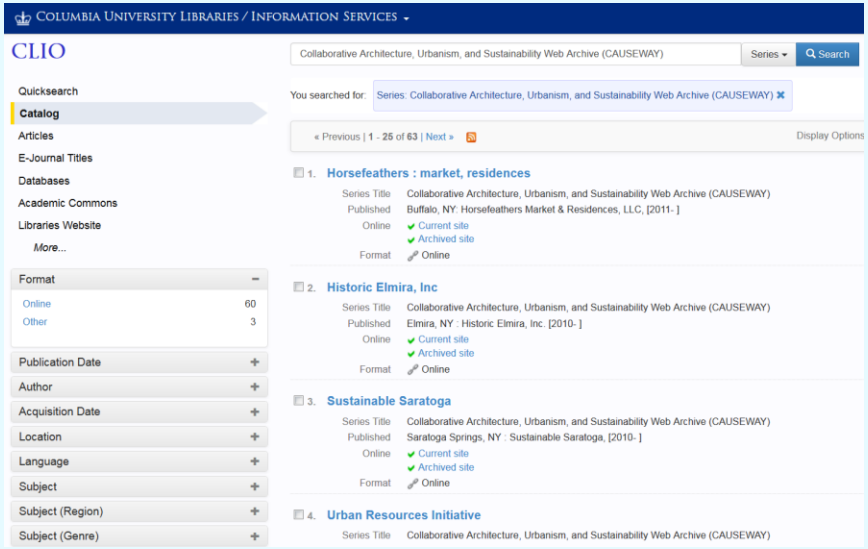
Photo credit: Anna Perricci

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

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

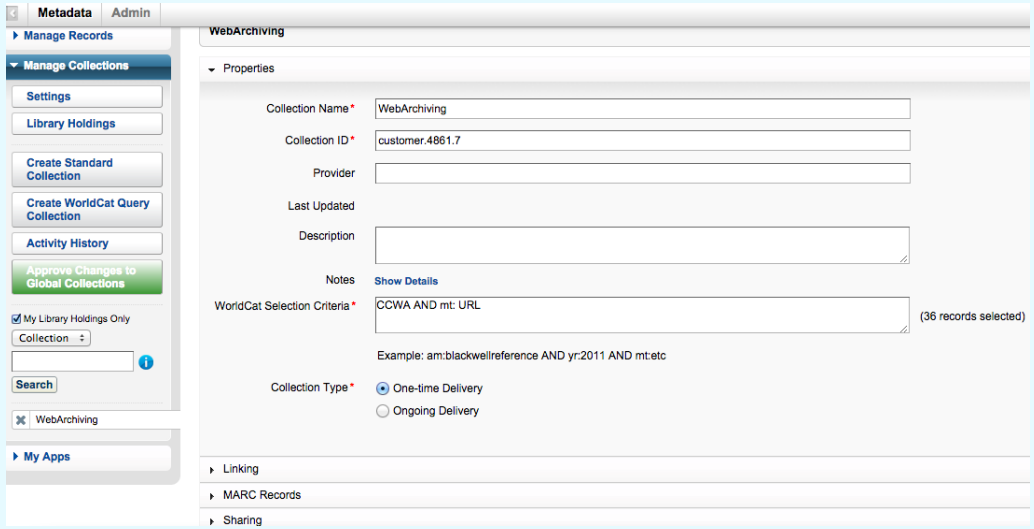




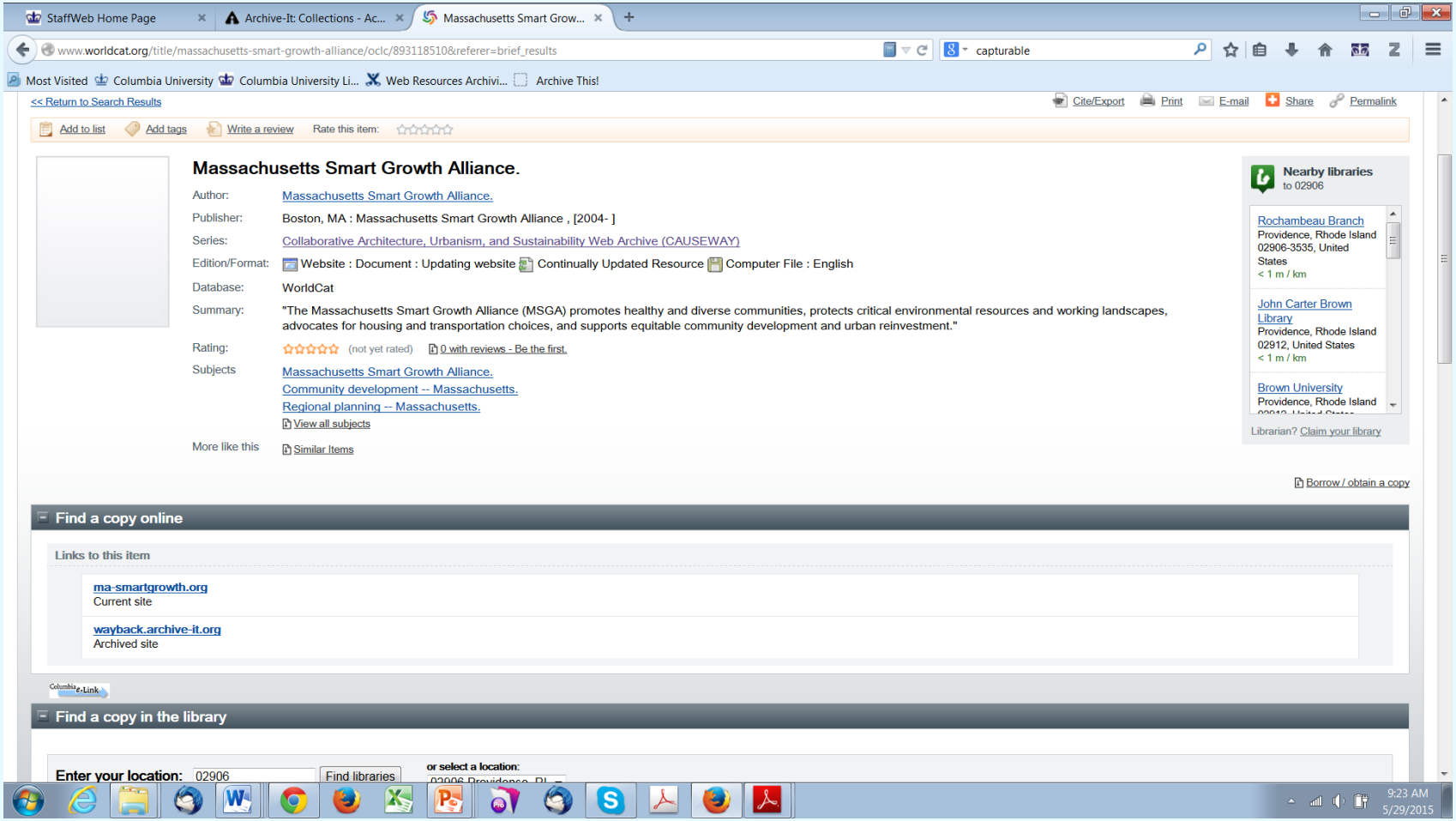
Getting the CAUSEWAY records in your OPAC via OCLC

- As mentioned, records have been released to WorldCat

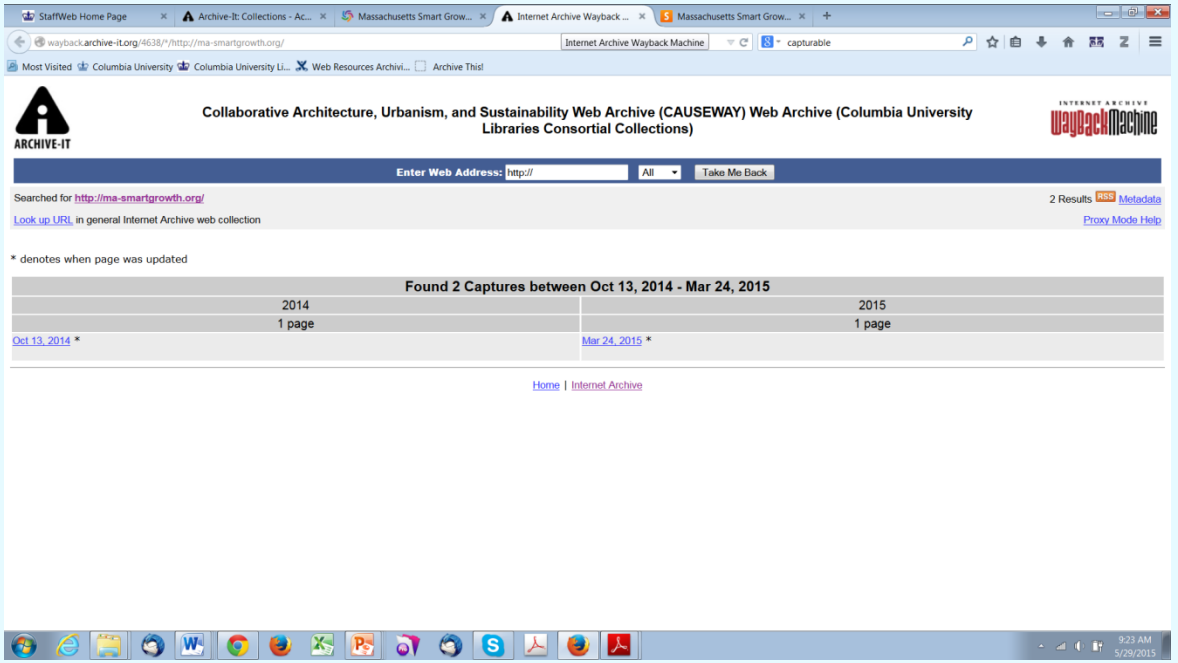
- A query can be built for OCLC WorldShare to obtain the MARC records for CAUSEWAY. The records can be delivered in a batch one time or periodically on an ongoing basis

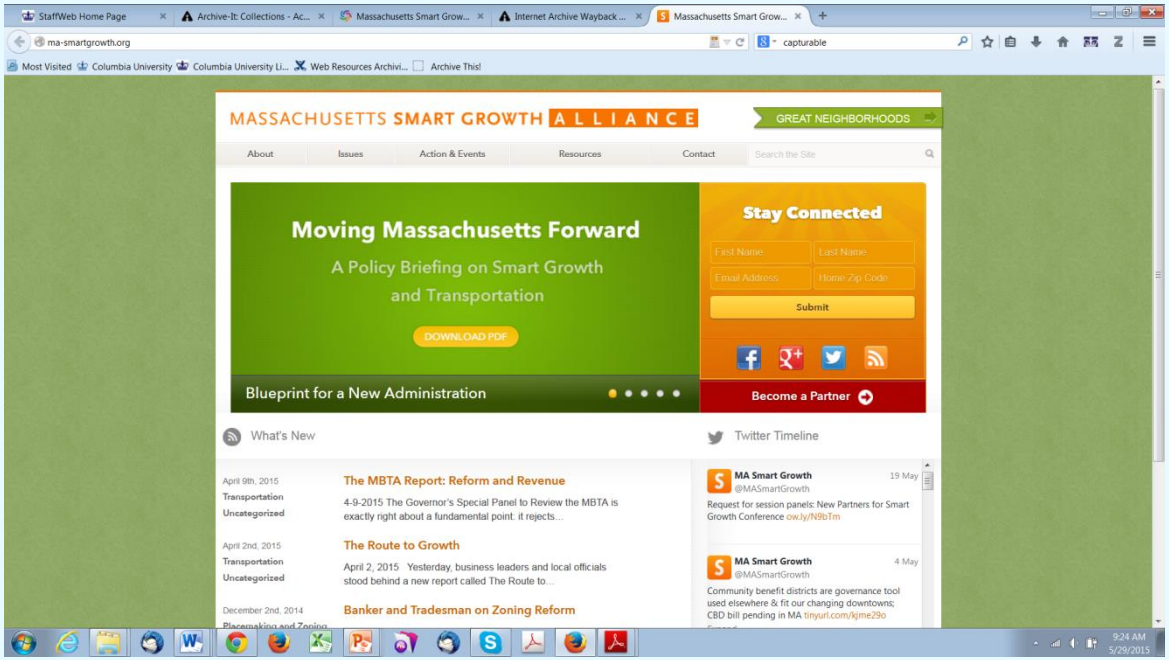


WorldCat records for CAUSEWAY sites



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





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!

Photo credit: Anna Perricci

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

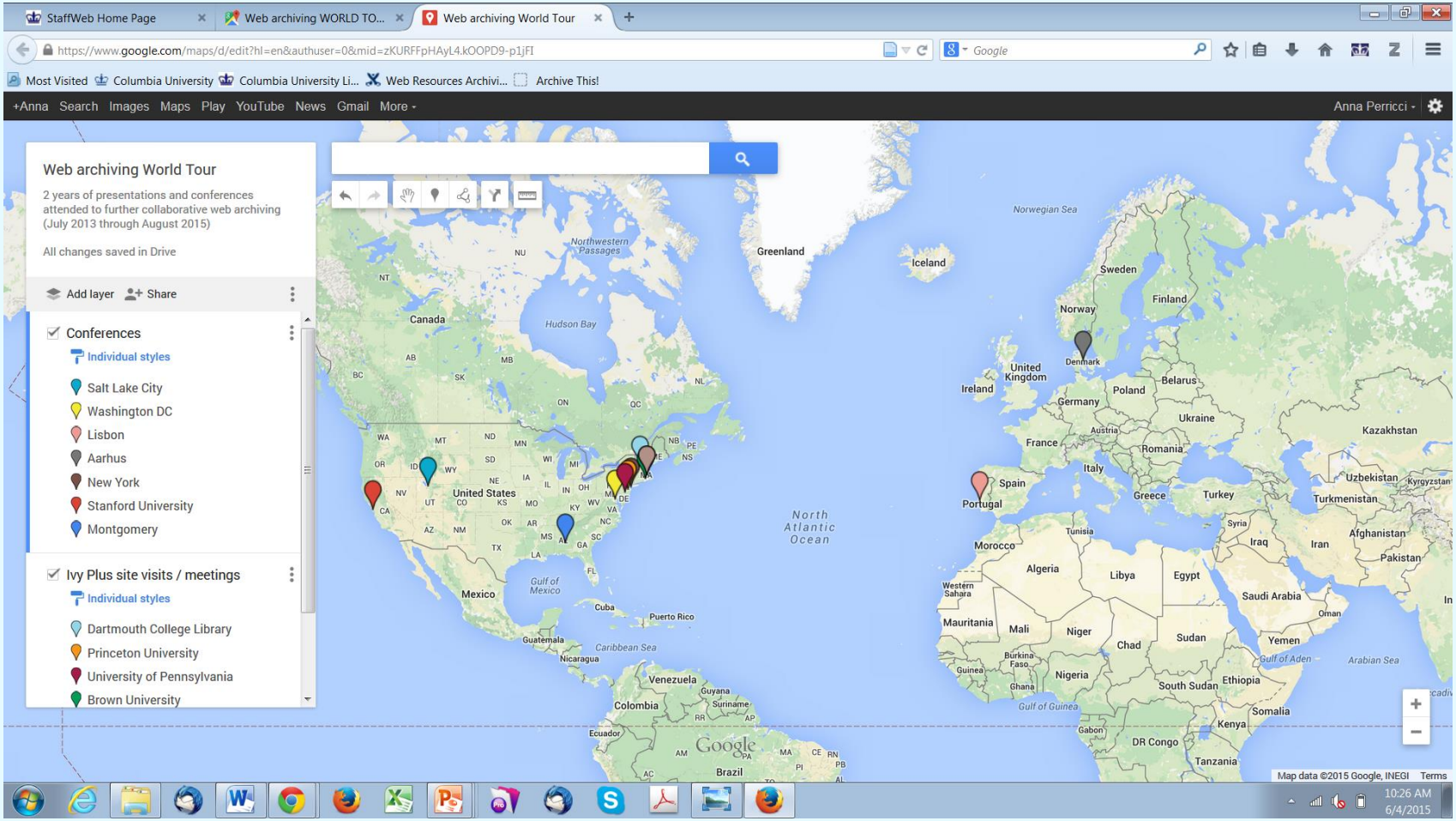
Photo credit: Anna Perricci

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)

Photo credit: Anna Perricci

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

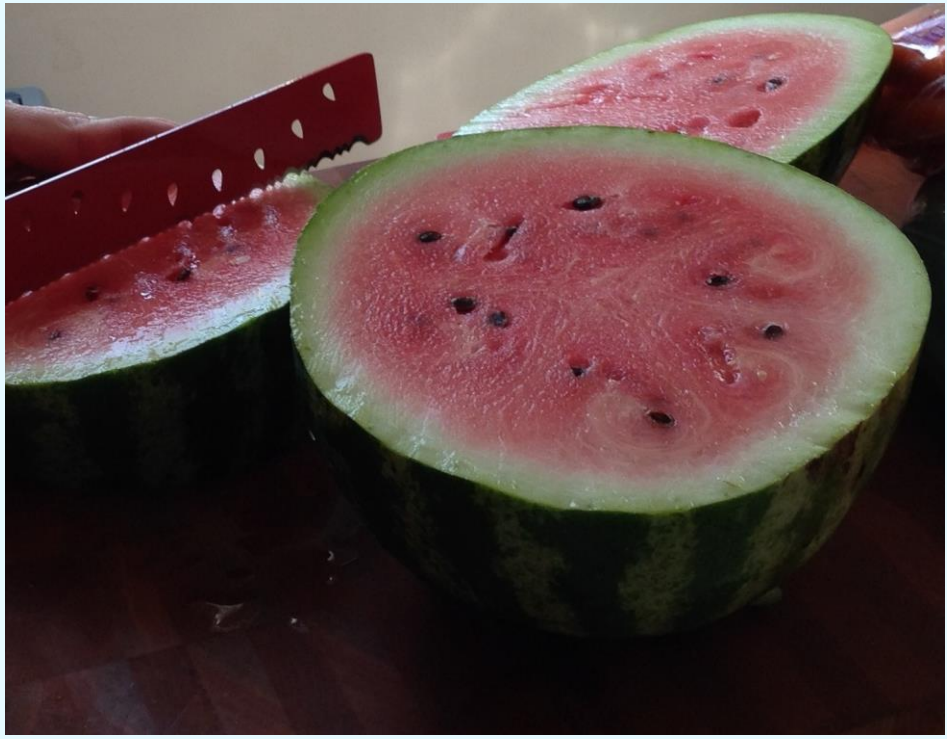
- Bob Wolven, Alex Thurman, Naeema Akter
- Pamela Graham, Kate Harcourt, Melanie Wacker, Christina Harlow
- Talia Jimenez, Stephen Davis, incentives awards oversight panel: Kris Carpenter, Mark Phillips, Rob Sanderson, Perry Willett, Martin Klein, Jefferson Bailey
- Elizabeth Davis, Russell Merritt & 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