

How collaboration can save [more of] the web: recent progress in collaborative web archiving initiatives

Anna Perricci

Columbia University Libraries

Best Practices Exchange

November 14, 2013

Overview

- Web archiving context (Who)
- Benefits of curated web archives (Why)
- Columbia University Libraries Web Resources Collection Program (What/How)
 - Background
 - Selection
 - Permissions
 - Harvesting
 - Description
 - Access

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 asst



CUL Web Archive Collections

- Avery Library Historic Preservation and Urban Planning
60+ websites, semiannual
- Burke Library New York City Religions
225 websites, semiannual
- Human Rights
547 websites, quarterly
- Rare Book and Manuscript Library
30 websites, semiannual
- University Archives
most of columbia.edu domain, plus 75+ affiliated sites
with "external" URLs
- General

CUL's Archive-It space: <http://www.archive-it.org/organizations/304>

Archive-It - Columbia Universit...

www.archive-it.org/organizations/304

Requested: 1996 2011 07/16/2013

Columbia University Libraries

Archive-It Partner Since: May, 2008
Organization Type: Colleges & Universities
Organization URL: <http://library.columbia.edu/>

Description: The Columbia University Libraries (CUL) web resources collection program archives selected websites in thematic areas corresponding to existing CUL collection strengths, websites produced by affiliates of Columbia University, and websites from organizations or individuals whose papers or records are held in CUL's physical archives.

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.

Subject Sort By: Count | (A-Z)

- Arts & Humanities (3)
- Society & Culture (3)
- Universities & Libraries (3)
- Architecture (1)
- City planning (1)

More ▾

Creator Sort By: Count | (A-Z)

- Avery Architectural and Fine Arts Library, Columbia University Libraries (1)
- Burke Library (Union Theological Seminary) (1)

Enter search terms here Search Clear

Collections Sites Search Page Text

Page 1 of 1 (6 Total Results)

Sort By: Collection Name (A-Z) | Collection Name (Z-A)

Avery Library Historic Preservation and Urban Planning

Archived since: Jan, 2010
Description: A collection of websites chosen by subject specialists

Creator Sort By: Count | (A-Z)

- Avery Architectural and Fine Arts Library, Columbia University Libraries (1)
- Burke Library (Union Theological Seminary) (1)
- Columbia University (1)
- Columbia University Libraries, Center for Human Rights Documentation and Research (1)

Coverage Sort By: Count | (A-Z)

- New York (N.Y.) (1)

Collector Sort By: Count | (A-Z)

- Avery Architectural and Fine Arts Library, Columbia University Libraries (1)
- Columbia University Libraries, Burke Theological Library (1)
- Columbia University Libraries, Center for Human Rights Documentation and Research (1)
- Columbia University Libraries, Rare Books and Manuscripts Library (1)

Sort By: Collection Name (A-Z) | Collection Name (Z-A)

Avery Library Historic Preservation and Urban Planning

Archived since: Jan, 2010
Description: A collection of websites chosen by subject specialists from the Avery Architectural and Fine Arts Library at Columbia University. The collection's principal thematic focus is documenting the evolution of the built environment and public spaces through the interaction of historic preservation efforts and new development projects within urban planning debates, particularly in and around New York City.
Subject: Arts & Humanities, Government - Cities, Architecture, Historic preservation, City planning, Parks, Public spaces
Creator: Avery Architectural and Fine Arts Library, Columbia University Libraries
Collector: Avery Architectural and Fine Arts Library, Columbia University Libraries

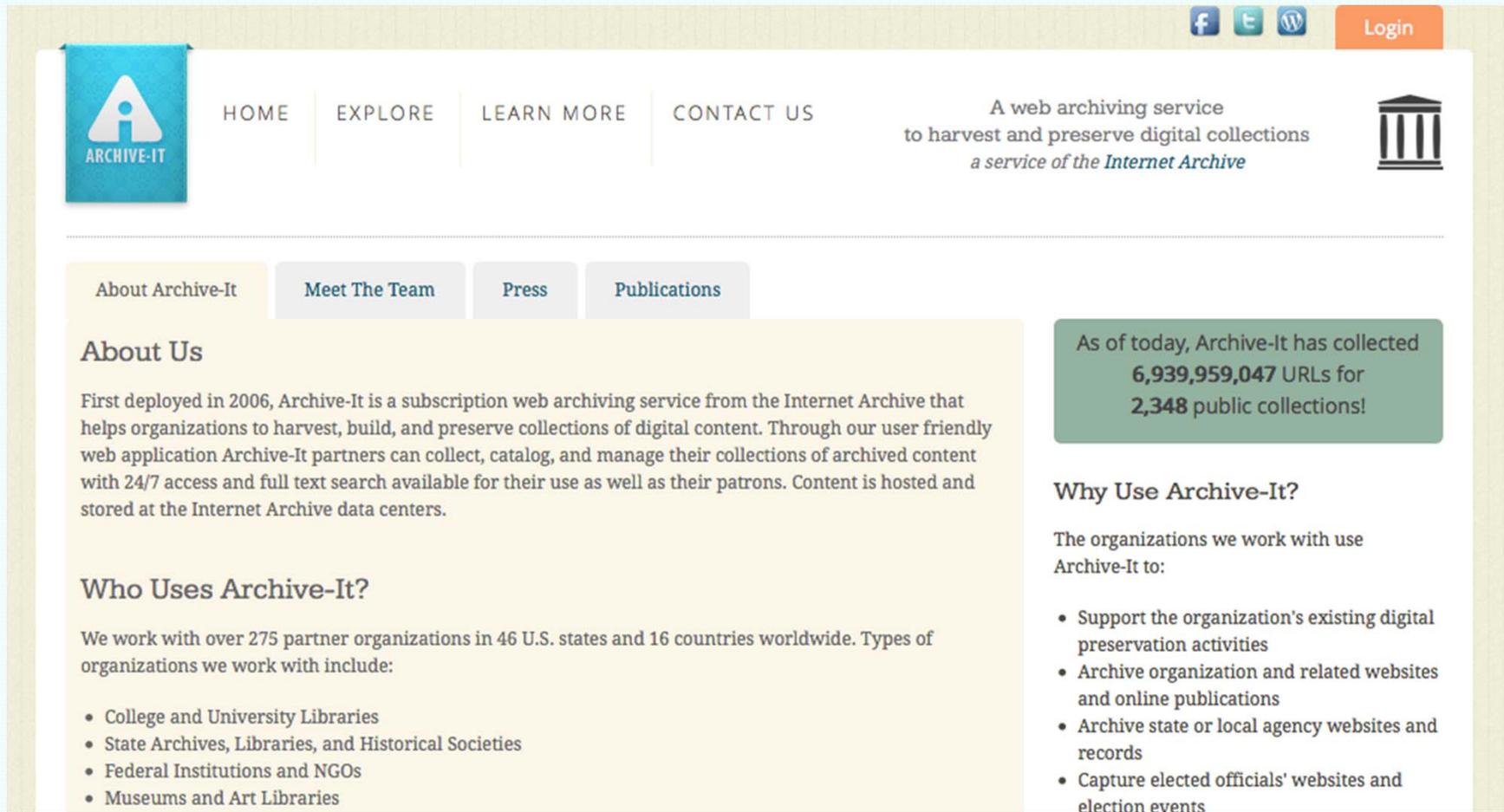
Burke Library New York City Religions

Archived since: May, 2010
Description: The Burke Library seeks to identify and preserve for the future information in the fields of religion, theology, and contextually related areas of study. As part of this mission, we seek to archive websites related to the Burke Library's existing collections, with a focus on religious communities and not-for-profit organizations in New York City.
Subject: Arts & Humanities, Society & Culture, Universities & Libraries, Religion, Theology, New York City religions
Creator: Burke Library (Union Theological Seminary)
Coverage: New York (N.Y.)
Collector: Columbia University Libraries, Burke Theological Library

Elements of CUL's web archiving workflows

- Selection or nomination
- Archive-It for scoping, crawls, access, storage & QA
- Cataloging workflows/output
- Firm policy on permissions
- Project management tools for growing number of collaborations
- Workflows can vary a little between collections

Archive-It provides some essential services and thereby lets CUL focus on curation, cataloging and collaboration building



The screenshot shows the Archive-It website homepage. At the top right, there are social media icons for Facebook, Twitter, and LinkedIn, along with a 'Login' button. The main navigation menu includes 'HOME', 'EXPLORE', 'LEARN MORE', and 'CONTACT US'. Below the navigation, there are tabs for 'About Archive-It', 'Meet The Team', 'Press', and 'Publications'. The 'About Us' section describes the service as a subscription web archiving service from the Internet Archive, first deployed in 2006. A green callout box highlights that Archive-It has collected 6,939,959,047 URLs for 2,348 public collections. The 'Who Uses Archive-It?' section lists partner organizations such as College and University Libraries, State Archives, Libraries, and Historical Societies, Federal Institutions and NGOs, and Museums and Art Libraries. The 'Why Use Archive-It?' section lists reasons for using the service, including supporting existing digital preservation activities, archiving organization and related websites, and capturing elected officials' websites and election events.

ARCHIVE-IT

HOME | EXPLORE | LEARN MORE | CONTACT US

A web archiving service
to harvest and preserve digital collections
a service of the Internet Archive

About Archive-It | Meet The Team | Press | Publications

About Us

First deployed in 2006, Archive-It is a subscription web archiving service from the Internet Archive that helps organizations to harvest, build, and preserve collections of digital content. Through our user friendly web application Archive-It partners can collect, catalog, and manage their collections of archived content with 24/7 access and full text search available for their use as well as their patrons. Content is hosted and stored at the Internet Archive data centers.

Who Uses Archive-It?

We work with over 275 partner organizations in 46 U.S. states and 16 countries worldwide. Types of organizations we work with include:

- College and University Libraries
- State Archives, Libraries, and Historical Societies
- Federal Institutions and NGOs
- Museums and Art Libraries

As of today, Archive-It has collected
6,939,959,047 URLs for
2,348 public collections!

Why Use Archive-It?

The organizations we work with use Archive-It to:

- Support the organization's existing digital preservation activities
- Archive organization and related websites and online publications
- Archive state or local agency websites and records
- Capture elected officials' websites and election events

Archive-It's responsiveness to CUL's requests

- Metadata: custom field functionality added to metadata template & all fields made repeatable
- Pilot projects: options related to ignoring robots.txt, downloading WARC files and export metadata as XML, improvements to capture for architecture/urban planning sites via Wayback QA tools

Challenges: using QA tools (media files & images)

HIGH LINE 

The official Web site of the High Line and Friends of the High Line

[DONATE](#) | [SEARCH](#) | [EMAIL UPDATES](#)

[ABOUT](#) | [NEWS & BLOG](#) | [EVENTS](#) | [GET INVOLVED](#) | [IMAGES & VIDEOS](#) | [DESIGN](#) | [SHOP](#)

IMAGES & VIDEO

Image Galleries
Videos

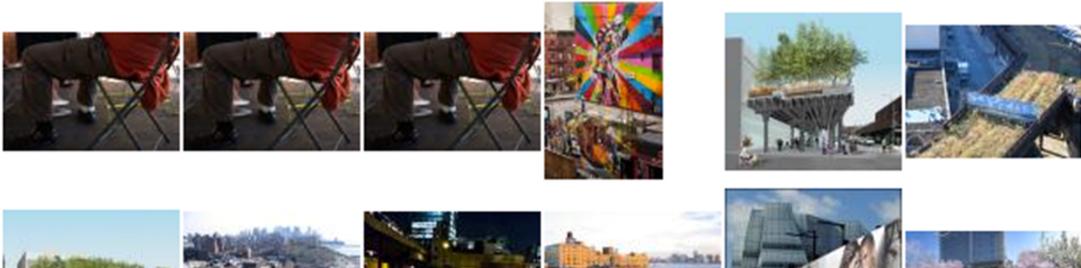
POPULAR PICTURES

1 2 3 4 5 6 7 8 9 ... next › last »

Show All Galleries
View Most Popular Images

Image Search

10thavenuesquare 2012 aerial
architecture art



Description and access for archived websites

- **Archive-it.org site-level metadata** (All thematic collections, DCMI, copied from MARC records if possible)
- **CLIO collection-level MARC records** (Human rights, Avery, Burke)
- **CLIO site-level MARC records** (Human rights, Avery)
- **Document-level MARC records** (selected longish Avery collection reports, pre-existing IRCR records)
- **Human Rights Web Archive portal on CUL website** (using metadata extracted from MARC records)

Avery Library historic preservation and urban planning web archive.

Title: Avery Library historic preservation and urban planning web archive.
Published: 2010-
Online Link(s): [Full collection access via Archive-It.org](#)
Restrictions: Browsing and full-text search of archived websites is available without restriction.
LC Subjects: [Atlantic Yards \(Project\)](#)
[Historic preservation--New York \(State\)--New York.](#)
[Historic preservation.](#)
[City planning--New York \(State\)--New York.](#)
[Historic districts--Conservation and restoration--New York \(State\)--New York.](#)
[Public spaces--New York \(State\)--New York.](#)
[Parks--New York \(State\)--New York.](#)
[Web archives.](#)
[New York \(N.Y.\)--Buildings, structures, etc.](#)

Also Listed Under: [Avery Library.](#)
[Columbia University. Libraries. Web Resources Collection Program.](#)

Holdings Information:

Location (guide):	Online
Call Number:	Avery Library Historic Preservation and Urban Planning
Status:	No information available

Other Subject Terms: [Web archives](#)

Summary:

A growing collection of websites selected by the Avery Architectural and Fine Arts Library staff for web archiving preservation by the Columbia University Libraries' Web Resources Collection Program. Website captures began in 2010 and are ongoing. The collection's principal thematic focus is documenting the evolution of the built environment and public spaces through the interaction of historic preservation efforts and new development projects within urban planning debates. Selected websites are mostly published by non-profit groups or individuals based in the New York City area, including historic preservation groups, neighborhood associations, public policy organizations, parks conservancies, and both sponsors and critics of ongoing development projects. Websites on related themes from other

CLIO public view of MARC collection-level record

Permissions

- No explicit US Copyright Act libraries exception for web archiving
- CUL policy is to request permission from website owners to harvest their websites and provide access to archived versions
 - Permission request email sent to contact info from website
 - If no response after 3 weeks, follow-up request with notification of intent to archive website
- Statistics
 - 918 requests sent
 - 437 responded Yes
 - 5 responded No
 - 415 did not respond

Web Resources Archiving Collaboration

- Many thanks to the Mellon Foundation
- Building collaborations among
 - The web archiving community
 - Other research libraries
 - Users and potential users of web archives
 - Site creators

Incentives grants to advance web archiving tools

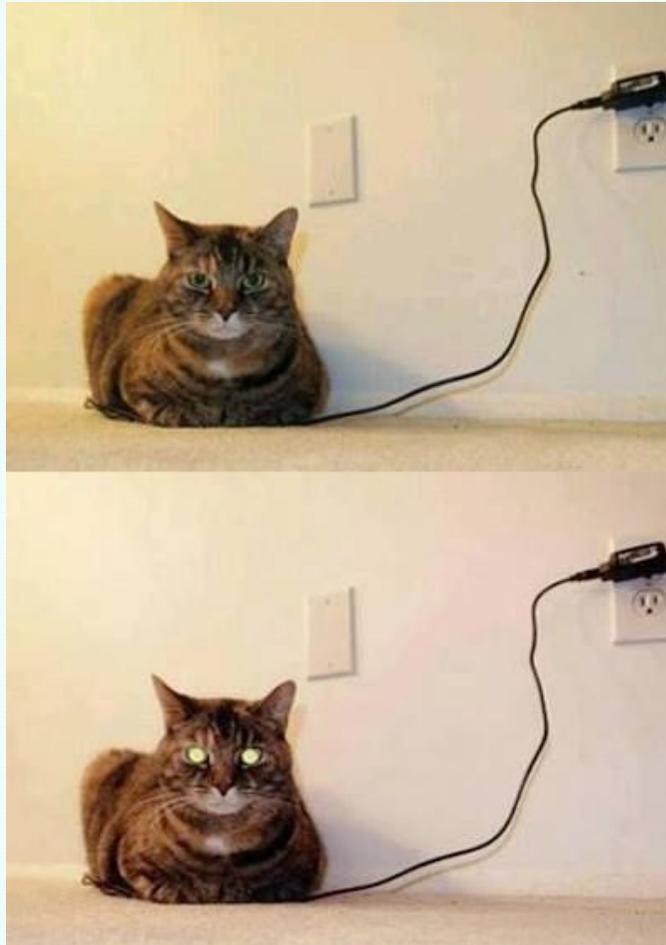


Image source: <http://imgur.com/gallery/vG7KE48>

Building an efficient, coherent, and scalable national framework for collecting web content



Designated space for collaborative collecting



HOME

EXPLORE

LEARN MORE

CONTACT US

A web archiving service
to harvest and preserve digital collections
a service of the Internet Archive



Explore >> Columbia University Libraries Consortial Collections



Columbia University Libraries Consortial Collections

Archive-It Partner Since: Jul, 2013

Organization Type: [Colleges & Universities](#)

Organization URL: <http://library.columbia.edu/>

Narrow Your Results

Subject

Sort By: **Count** | [\(A-Z\)](#)

Arts & Humanities (1)

Society & Culture (1)

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.

Search

Clear

Collections

Sites

Search Page Text

Collection

Collaborative collection experiment in progress via partner institutions in Borrow Direct

The screenshot shows the Archive-IT website interface. At the top right, there are social media icons for Facebook, Twitter, and WordPress, along with a 'Login' button. The main navigation bar includes 'HOME', 'EXPLORE', 'LEARN MORE', and 'CONTACT US'. A tagline reads: 'A web archiving service to harvest and preserve digital collections a service of the Internet Archive'. A classical building icon is also present.

Breadcrumbs: [Explore](#) >> [Columbia University Libraries Consortial Collections](#) >> [Contemporary Composers Web Archive](#)



ARCHIVE-IT

Contemporary Composers Web Archive

Collected by: [Columbia University Libraries Consortial Collections](#)

Archived since: Oct, 2013

Description: The Contemporary Composers Web Archive is a newly launched initiative by the music librarians at Brown, Columbia, Cornell, Dartmouth, Harvard, Princeton, and Yale universities, MIT, and the universities of Chicago and Pennsylvania (collectively known as the Borrow Direct Music Librarians Group) and operates under the auspices of Columbia University Libraries and Information Services.

Subject: [Arts & Humanities](#), [Society & Culture](#)

Collector: [Borrow Direct Music Librarians Group](#)

Narrow Your Results

There are no further ways to narrow your

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.

Collaboration with music librarians



Contemporary composers—the perfect storm?

The screenshot shows the John Cage Complete Works website in a Firefox browser. The page is mostly blank with a 'Please Wait...' message in the center. On the left, there is a search and filter interface. The search bar contains 'Please Wait...'. Below it, there are several filter categories: Duration, Premiere Date, Manuscript Location, Ensemble Type, Instrument Type, and Specific Instrument. There is also a 'Number of Players' section with radio buttons for 'Exactly this Number', 'Up to This Number', and 'At Least this Number'. At the bottom of the filter section, there are dropdown menus for 'Collaborator', 'Cunningham Related', and 'Recordings'. On the right side of the page, there is a list of links: 'Sonatas and Interludes new', 'Prepared Piano App new', 'Autobiographical Statement', 'The John Cage Trust', 'Indeterminacy', 'Folksonomy', 'Kuhn's Blog', and 'Cage Database new'.

The screenshot shows the same website with search results displayed. The search bar contains '0'00" (4'33" No. 2)'. The results table has four columns: Search, Title, Work Summary, and Details/Permalink. The first result is '0'00" (4'33" No. 2)'. The 'Work Summary' column for this result contains the following information: 'Work Title: 0'00" (4'33" No. 2)', 'Alt. Title: 4'33" No. 2, Zero Minutes Zero Seconds, 0:00', '2nd Alt. Title: 0'00"', 'Date: Composed in 1962. Premiered in Tokyo, October 24, 1962.', 'Instrumentation: Solo to be performed in any way by anyone.', 'Length: indeterminate', and 'Comments: The original score for this work consists of one sentence: "In a situation provided with maximum amplification (no feedback), perform a disciplined action." A day later, Cage added further instructions, for example allowing interruptions of the action, not

Search	Title	Work Summary	Details/Permalink
0'00" (4'33" No. 2)	0'00" (4'33" No. 2)	Work Title 0'00" (4'33" No. 2)	
	103	Alt. Title 4'33" No. 2, Zero Minutes Zero Seconds, 0:00	
	108	2nd Alt. Title 0'00"	
	101	Date Composed in 1962. Premiered in Tokyo, October 24, 1962.	
	26'1.1499"	Instrumentation Solo to be performed in any way by anyone.	
	07' 10.554"	Length indeterminate	
	2nd Construction	Comments The original score for this work consists of one sentence: "In a situation provided with maximum amplification (no feedback), perform a disciplined action." A day later, Cage added further instructions, for example allowing interruptions of the action, not	
	34'57.9864"		
	33 1/3		
	34'46.776"		
	4'33"		
	49 Waltzes for the Five Boroughs		
	59 1/2" for a String Player		
	A Book of Music		
	A Chant with Claps		

Collection Development (test case)

Project tracking

The screenshot shows a Firefox browser window displaying a Basecamp project page. The browser's address bar shows the URL: <https://basecamp.com/2322108/projects/3292759-ccwa-contemporary>. The page title is "CCWA - Contemporary Composers Web Archive" and it is categorized as a "Borrow Direct - Composers project".

The page features a navigation bar with "Projects", "Calendar", "Everything", "Progress", "Everyone", and "Me". A search bar is present with the text "Jump to a project, person, label, or search...".

Key elements on the page include:

- Project Overview:** "CCWA - Contemporary Composers Web Archive" with a star icon. It shows "20 people on this project" and options to "Invite more people" and "Catch up on recent changes".
- Navigation:** Links for "1 Discussion", "346 To-dos", "1 File", "84 Forwarded emails", and "Dates". There is also an "Add the first: Text document" button.
- Latest project updates:** A section with three entries:
 - 3:03pm:** "You forwarded an email: Contemporary Composers Web Archive"
 - 2:34pm:** "Hoda H. completed a to-do: Record the permissions response (or lack thereof) in master spreads..."
 - 2:34pm:** "Hoda H. completed a to-do: Send 2nd permissions request message to composer (if applicable--re..."
- Upcoming Events:** A yellow box on the right side lists "October 21" with the event "Borrow Direct AUL meeting".
- Discussions:** A section with a "Post a new message" button. A discussion by Anna P. is titled "Composers' permissions obtained (in order..." and dated "Oct 17".
- To-do lists:** A section with an "Add a to-do list" button. A to-do list for "Christine Southworth" includes:
 - Send 2nd permissions request message to composer (if applicable--resolve this item if you get a response to the first message and confirm that the message is in Basecamp)
 - Record the permissions response (or lack thereof) in master spreadsheet (Hoda Hassanein)
 - Test crawl site (if permission granted) (Anna Perricci)
 - Harvest site (if permission granted) (Anna Perricci)

Use cases



Who are the web archives for? Are they being used? Could we encourage more effective use?

The screenshot shows the Columbia University website header with the university logo and name. A navigation bar includes links for ABOUT, ADMISSIONS, ACADEMICS, RESEARCH, LIBRARIES, MEDICAL CENTER, and Resources for. A search bar is located in the top right corner. Below the navigation bar, the main heading 'AREAS OF STUDY' is displayed, followed by a breadcrumb trail: Home > Admissions > Areas of Study. A link 'Can't find what you're looking for? Let us know!' is also present. The 'AREAS OF STUDY' section features a horizontal list of letters from A to Y, with 'A' highlighted. Below this list, a scrollable area displays a list of study programs under the letter 'A', each followed by a right-pointing arrow.

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Email | Quick Links | Main Menu | A-Z Index

Search for people, departments & websites

ABOUT | **ADMISSIONS** | ACADEMICS | RESEARCH | LIBRARIES | MEDICAL CENTER | Resources for

AREAS OF STUDY

Home > Admissions > Areas of Study Can't find what you're looking for? Let us know!

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

- Accelerated Program in Interdisciplinary Legal Education (Barnard College) »
- Actuarial Science (School of Continuing Education) »
- Adult Learning and Leadership (Teachers College) »
- Advanced Architectural Design (Graduate School of Architecture, Planning and Preservation) »
- African Studies (Columbia College) »
- African Studies (School of General Studies) »
- African Studies, Certificate in (Graduate School of Arts & Sciences) »
- African-American Studies (Columbia College) »
- African-American Studies (Graduate School of Arts & Sciences) »
- African-American Studies (School of General Studies) »
- Africana Studies (Barnard College) »



HUMAN RIGHTS WEB ARCHIVE

CENTER FOR HUMAN RIGHTS DOCUMENTATION & RESEARCH AT COLUMBIA UNIVERSITY

BETA



LEARN MORE

NOMINATE A SITE

CONTACT

SEARCH HISTORY

The Human Rights Web Archive

@ Columbia University is a searchable collection of archived copies of human rights websites created by non-governmental organizations, national human rights institutions, tribunals and individuals. Collecting began in 2008 and has been ongoing for active websites. New websites are added to the collection regularly.

The HRWA is an initiative of the Center for Human Rights Documentation & Research and is a key focus of the Columbia University Libraries' Web Resources Collection Program. The HRWA was made possible by generous support from the Andrew W. Mellon Foundation. [Learn More »](#)

FEATURED SITES

TITLES

URLS

SUBJECTS

PLACES

LANGUAGES

A selection of sites relating to the subject Transitional justice.



Featured Sites



International Criminal Tribunal f...

Site Details



Institute for Justice and Recon...

Site Details



Hay'at al-Inṣāf wa-al-Muṣālaḥah

Site Details



CSVr, Centre for the Study of Vi...

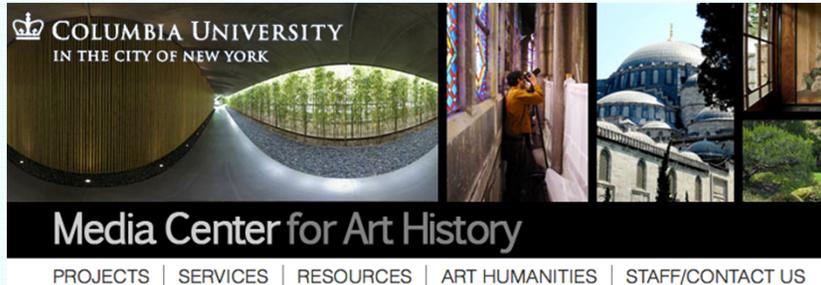


Truth and Reconciliation Commi...



South Africa Truth and Reconcil...

Columbia University web resources: creating best practices for site creators



The Media Center for Art History, part of the [Department of Art History and Archaeology](#) at Columbia University, explores material culture, vision, media, and pedagogy in the broadest sense to connect faculty research and student learning through the creative application of technology. A motivated group of faculty principal investigators work with the Media Center to develop, conduct, and administer their projects in the study, interpretation, and conservation of works of art, monuments, or heritage sites.

Our goal is to examine and extend the ways of interpreting images, objects, buildings, and sites and to reinforce Columbia's historic strengths in core



Columbia Center for New Media Teaching and Learning

CCNMTL partners with Columbia University faculty to enhance

Spotlight

NEWS & UPDATES | November 11, 2013

[Frank A. Moretti Memorial Symposium to Be Held at Columbia Nov. 18](#)

A symposium honoring the work and life of

Best Practices for site creators: working with website creators



Open issue: division and maintenance of cooperative efforts (communication, software and more)

The screenshot displays the Basecamp web interface. At the top, the Basecamp logo is on the left, followed by navigation tabs: **Projects**, Calendar, Everything, Progress, Everyone, and Me. A search bar on the right contains the text "Jump to a project, person, label, or search...".

On the left sidebar, there is a "New Project" button with a green plus icon, and a "Templates" link below it. The main content area features a grid of six project cards, each with a star icon in the top right corner:

- Citation Tracking and Use Case Project**: Last updated on Nov 6. Includes a row of five circular icons representing team members.
- CCWA - Contemporary Composers Web Archive**: Borrow Direct - Composers project. Last updated 25 minutes ago. Includes a row of seven circular icons representing team members.
- Web Archiving Incentives Grant Program**: Last updated Friday at 2:44pm. Includes a row of seven circular icons representing team members.
- Collaborative networks**: e.g. Borrow Direct (macro), NYARC, NYU, etc. Last updated on Nov 1.
- Best Practices - Collaboration with site creators**: CU, CCNMTL, external. Last updated on Oct 18.
- Ivies + Art & Architecture**: Last updated on Nov 1.

Process over next 18 months

- Planning, needs assessment (interviews)
- Group communication
- Ongoing growth (scale of collecting and distribution of effort)
- Planning for sharing costs
- 5 year plan for Borrow Direct collaborations

Leveraging expertise & devising ways to share responsibilities

- Collecting strategy and establishment of priorities for collection development
 - Suggest seed URLs
 - Liaise with site owners to solicit permission to archive websites
- Conduct detailed quality assurance by browsing the archived website as a user would (e.g. try to access media files to ensure they have been successfully captured)
- Assessment of efficacy for users?

Leveraging expertise: web archivists

- Capture web archives & do initial quality assurance of a crawl
- Coordinate efforts and field questions regarding
 - technical elements of web archiving
 - existing CUL web archiving policy
 - permissions processing
 - needs assessment
 - user profiles and use cases
 - value and usage assessment? (someday)

Larger questions to return to periodically (and maybe discuss today!)

- Do you have any ideas about what users would want to find in web archives or use cases?
- Which websites do you and your colleagues consider indispensable for your day to day work?
- Are there any websites in your institution's domain that should be harvested and saved (e.g. web pages created by or about faculty members, academic programs or student groups)?
- Are there any types of sites that would certainly be out of scope or have proven to be out of scope?
- Is there anything impeding the use of web archives we are creating?

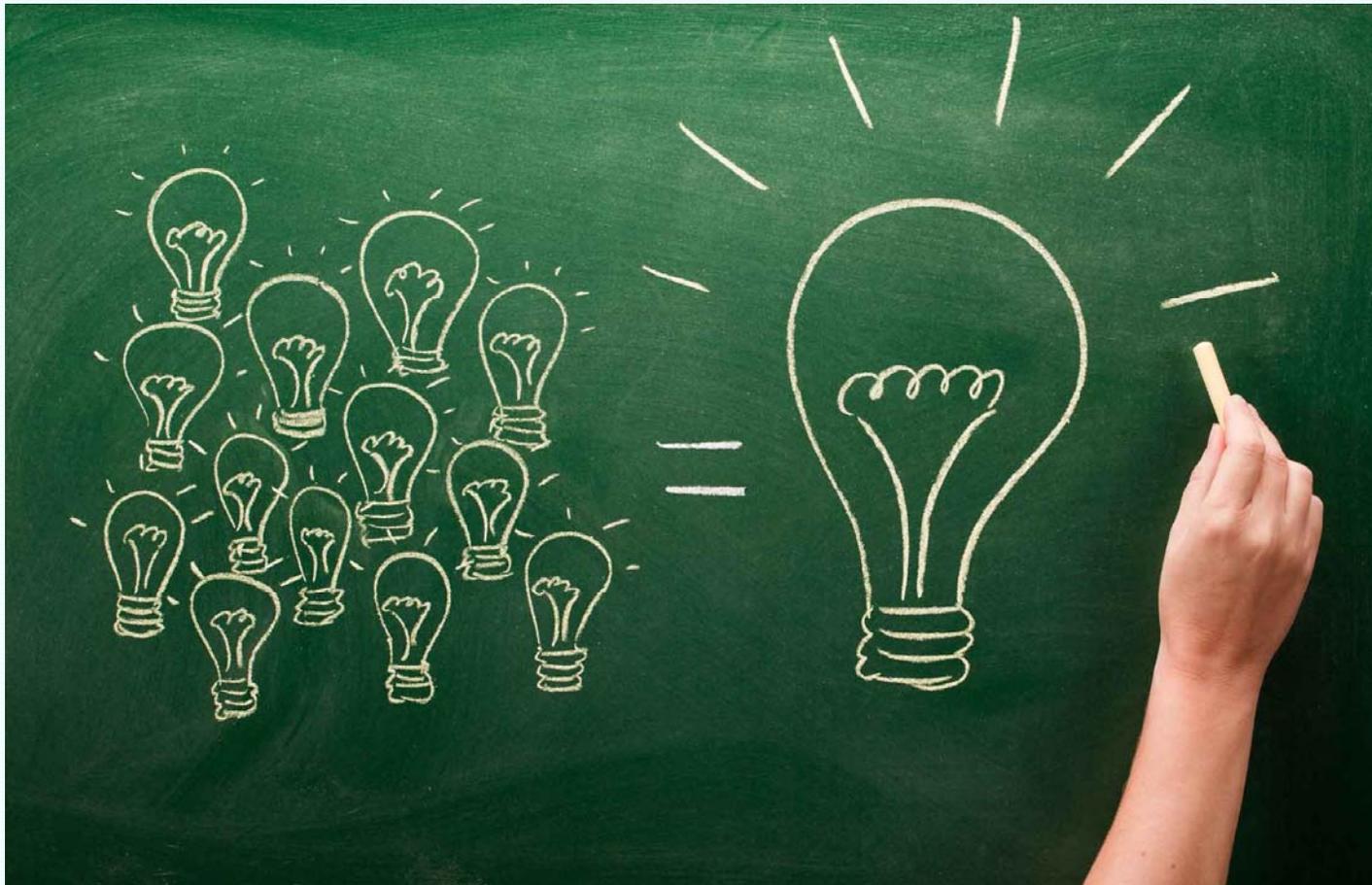
Takeaways



The future is bright!



Opportunities for discussion: Today and on an ongoing basis



Thank you!

Anna Perricci

alp2198@columbia.edu