# Sustainable Web Archiving at Scale: An Introduction



June 21-22, 2021

### Workshop Instructors





Sumitra Duncan

Amye McCarther

# Agenda

Day 1: Intro to Web Archiving

Concepts, Policy and Practice:

- What is web archiving?
- Collecting objectives
- Copyright, permissions & ethics
- Capacity considerations & description/metadata

Day 2: Tools + Technology

Overview, demos and hands-on exercises:

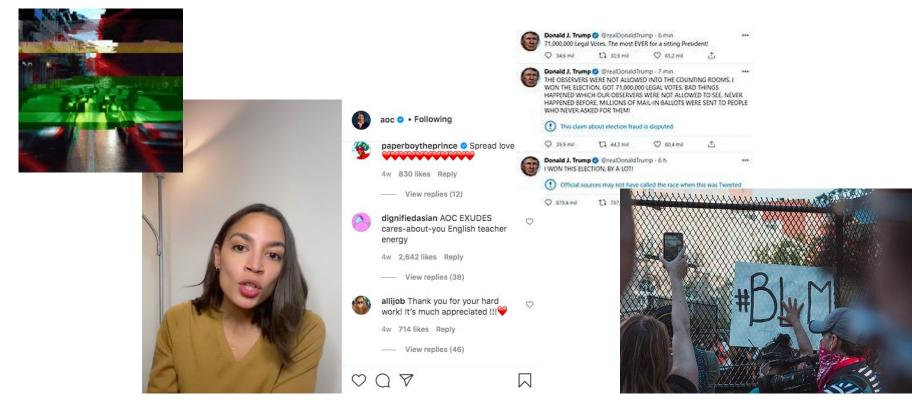
- Tools & replay mechanisms
- Internet Archive + Archive-It
- Conifer/Webrecorder
- Other tools

### What is Web Archiving?

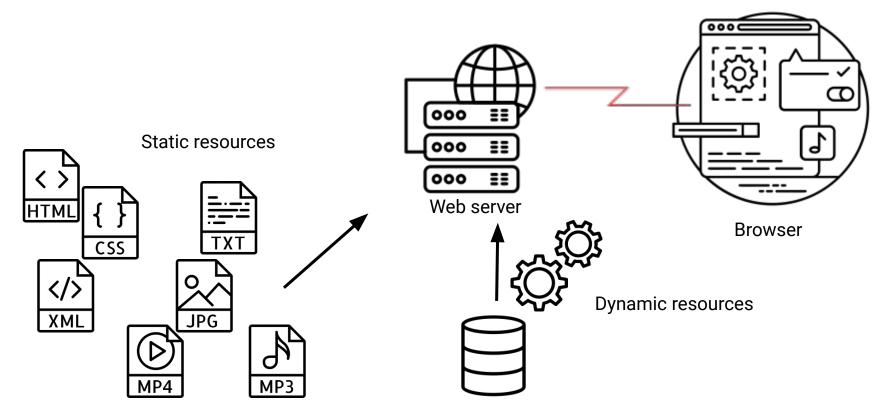
Web archiving is the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use.

- International Internet Preservation Consortium

### What is Web Archiving?



### Static vs. Dynamic Web



## Archival Processing: Similarities & Differences

- Appraisal & Selection
- Permissions Clearance
- Ethics Review
- Collecting

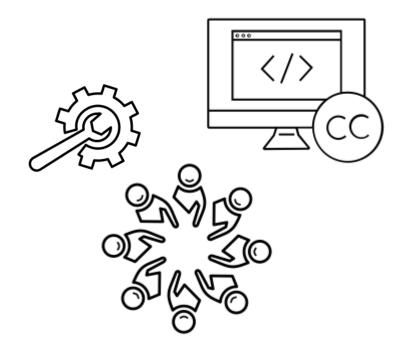
- Description
- Access
- Long-term Preservation

### Software

Open source vs. Proprietary / Subscription

Community adoption

Interoperability



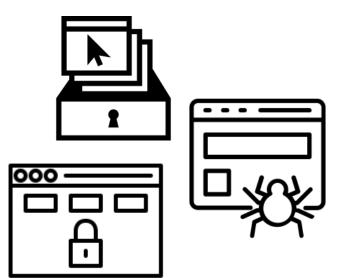
### File Formats

- WARC
   ISO Standard
   International community
- WACZ
  - In development

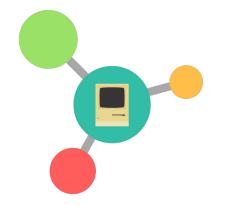


## Key Terms & Concepts

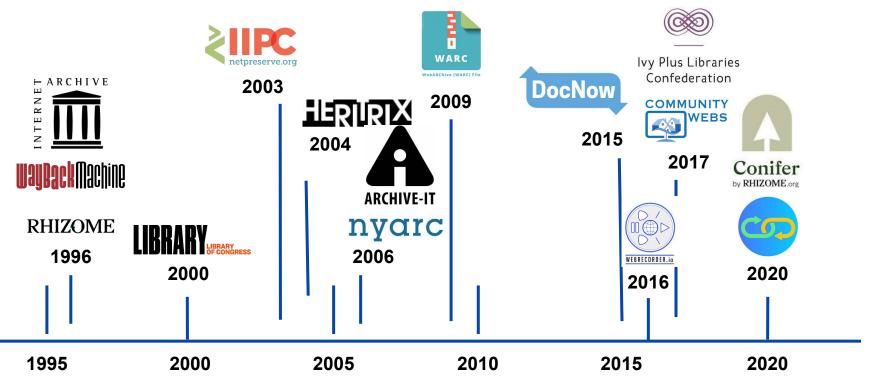
- Bound Archives
- Crawl/Capture/Harvest
- Crawler
- Seed URLs
- robots.txt



### Questions?



### **Community & Practitioners**



### Community & Practitioners Internet Archive

- Founded in 1996, the first large-scale web archiving project.
  - Wayback Machine
  - Heritrix
  - Archive-It



### Community & Practitioners National Libraries



Community & Practitioners International Internet Preservation Consortium

Founded in 2003 to promote best practices, dialogue, and advocacy.

- Consortium of 35+ countries
- International standards, tools
- WARC ISO Standard





Community & Practitioners Universities & Consortia

- Ivy Plus Libraries Confederation (IPLC)
- New York Art Resources Consortium (NYARC)
- Community Webs



Ivy Plus Libraries Confederation

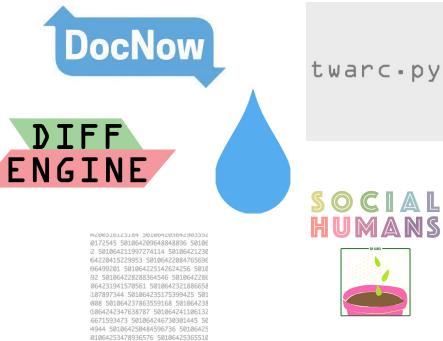


**nyarc** New York Art Resources Consortium Community & Practitioners Social Media & the Dynamic Web



### Community & Practitioners Documenting the Now

Founded in 2015, DocNow develops tools and guidance for collecting, analyzing, and preserving Twitter that prioritize ethical practices.



56775663616 501064257115402240 5

### Community & Practitioners Webrecorder/Conifer

- Colloq
- Oldweb.today
- Webrecorder
  - Conifer
  - Replay.webpage



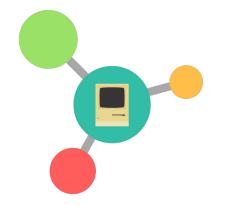
RHIZOME



Community & Practitioners Professional Organizations & Groups

- The Society of American Archivists (SAA) Web Archiving Section
- ART|WARC, the Art Libraries Society of North America (ARLIS/NA) Web Archiving Special Interest Group
- The National Digital Stewardship Alliance (NDSA) Web Archiving Surveys

### Questions?

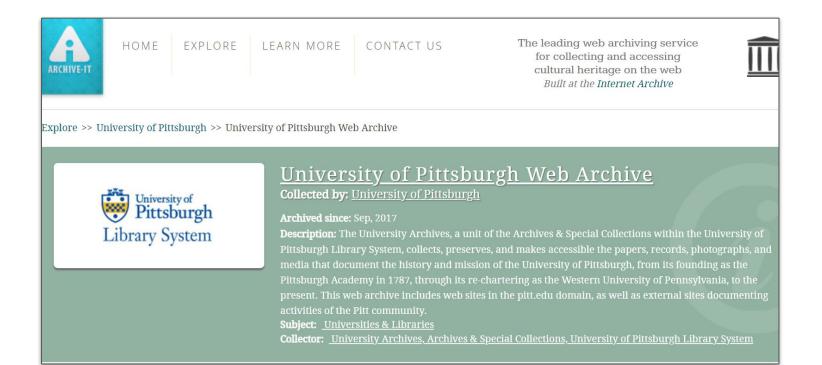


## Collecting Objectives & Mission



Image: RaHul Rodriguez https://creativecommons.org/licenses/by-sa/2.0/

### Institutional Web Presence



### **Collaborative Collecting**

HOME EXPLORE	LEARN MORE	CONTACT US	The leading web archiving service for collecting and accessing cultural heritage on the web <i>Built at the Internet Archive</i>			
Explore >> New York Art Resources Consorti NYCIIC New York Art Resources Consortium	New Yo Archive-It Partu Organization Ty	P <b>rk Art Resou</b> ner Since: Oct, 2010 /pe: <u>Museums &amp; Art Librar</u> RL: <u>http://www.nyarc.org</u>	u <mark>rces Consortium (NYARC)</mark> <sup>ies</sup>			
	Description: The New York Art Resources Consortium (NYARC) consists of the research libraries of three leading art museums in New York City: The Brooklyn Museum, The Frick Collection, and The Museum of Modern Art. The NYARC web resources program archives curated collections of websites in areas which correspond to the scope and strengths of the print collections at each research library, as well as NYARC project websites and the institutional websites of the three museums.To nominate websites for inclusion NYARC's collections, please email: webarchive[at]frick[dot]org					

### Subject-based Collecting

### Venice Biennale 2017 on the Web

### Archived since: May, 2017

**Description:** This collection of blogs, social media sites, video, and organizational websites documents the international art exhibition, La Biennale di Venezia - 57th International Art Exhibition. The crawl began May 10, 2017 and continues through to the end of the exhibition in November. An initiative undertaken by the Sterling and Francine Clark Art Institute Library and its partner Archive-It, the Venice Biennale web project complements the Clark's Venice Biennale Collection of exhibition catalogues, press materials, and ephemera beginning with the 52nd Biennale in 2007. **Subject:** Arts & Humanities, Society & Culture

### Venice Biennale 2019 on the Web

#### Archived since: Mar, 2019

**Description:** This collection of blogs, social media sites, video, and organizational websites documents the international art exhibition, La Biennale di Venezia - 58th International Art Exhibition. The crawl will begin May 8, 2019 and continues through to the end of the exhibition in November. An initiative undertaken by the Sterling and Francine Clark Art Institute Library and its partner Archive-It, the Venice Biennale web project complements the Clark's Venice Biennale Collection of exhibition catalogues, press materials, and ephemera beginning with the 52nd Biennale in 2007. **Subject:** Arts & Humanities, Society & Culture, Blogs & Social Media

## #Syllabus



### **Event-based** Collecting

ARCHIVE-IT	HOME	EXPLORE	LEARN MORE	CONTACT US	The leading web archiving service for collecting and accessing cultural heritage on the web <i>Built at the Internet Archive</i>	Î		
Explore >> International Internet Preservation Consortium >> Novel Coronavirus (COVID-19)								



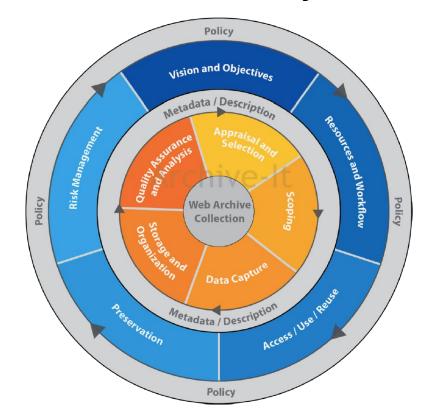
### Novel Coronavirus (COVID-19)

#### Collected by: International Internet Preservation Consortium

#### Archived since: Feb, 2020

Description: A collection created by the Content Development Group of the International Internet Preservation Consortium in collaboration with Archive-It to preserve web content related to the ongoing Novel Coronavirus (Covid-19) outbreak. Identification of seed websites and initial web crawling began in February 2020, and the collection will continue to add new content as needed during the course of the pandemic and its containment. High priority subtopics include: coronavirus origins; information about the spread of infection; regional or local containment efforts, including vaccine development and implementation of vaccinations; medical and scientific aspects; social aspects; economic aspects; and political aspects. Websites from anywhere in the world and in any language are in scope. Subject: Science & Health, Spontaneous Events, Epidemics, Coronavirus infections, COVID-19 (Disease), COVID-19 Pandemic, 2020 Creator: International Internet Preservation Consortium Collector: International Internet Preservation Consortium

### Web Archive Life Cycle Model



### Purpose & Goals

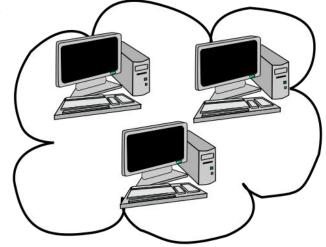
- What is the scope of this web archive and why?
- What types of web-published materials could be included?
   What should *not* be included/considered as in-scope?
- What (if any) relationship does this web archive have to other collecting and/or preservation areas on this subject or for these collectors?
- Is there a mandate to collect comprehensively around this topic, to preserve institutional records, or to fill a gap in the historical record?

## Purpose & Goals

- What is the institutional, regional, thematic scope/subject area of this collection?
  - Does this collection enhance an existing one?
- Are there areas of strength in your collections that have shifted entirely or partially to the web?
- Do you already have collection policies for digital content? For the web specifically?
  - Existing collecting policies may not need to change if they're format agnostic or already apply to the web.

# Avoiding Duplication

- Survey and assess whether the content is already being adequately archived by another institution
- Review existing collection policies as benchmarks
- Nominate appropriate sites to existing collections (if accepted) and focus your resources on other topics



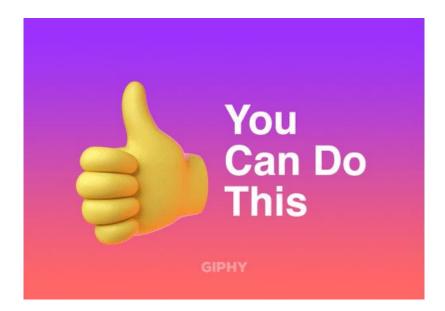
### Stakeholders & Users

- Who exactly is responsible for defining the scope? For appraisal and selection of its contents moving forward?
- Who are the direct patrons of this web archive? What are their goals?
- Consider internal stakeholders in your institutions or consortia, such as curators, faculty, publishers and IT.
- What agency do artists, scholars, publishers, or other creators have in their inclusion in this archive?

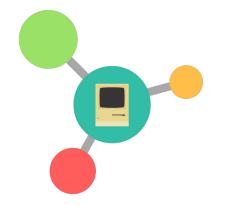
## Access & Rights in Policies

- Are there any restrictions or limitations on the use of the archive?
- How does a typical primary patron find and explore the archive?
- Is it related to or integrated into any existing finding aid, catalog, discovery layer, or other access point?
- Is there a model or standard for the lifecycle of these collected materials as born-digital files?

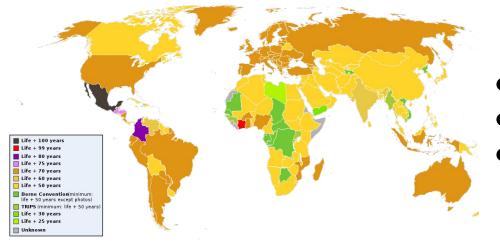
# Starting Small & Building Incrementally



### Questions?



## **Copyright Considerations**



- Domain (ie. ".fr" ".de" ".nz")
- Legal deposit
- Permissions requirements

## Approaches

- Permissions
- Fair use
- Notification
- Opt-out
- Robots.txt
- Embargo periods



## A Question of Scale

Large Scale

- Individual permissions not feasible
- Obey robots.txt

Small Scale

- May circumvent robots.txt
- Able to handle notices, take-downs





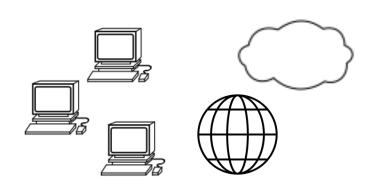
#### Access

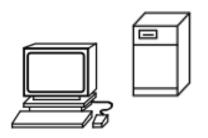
#### Local Access

- Less risk
- Registered users

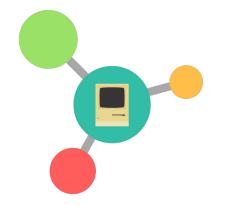
**Online Access** 

- Obtain permissions
- Robot exclusions





#### Questions?

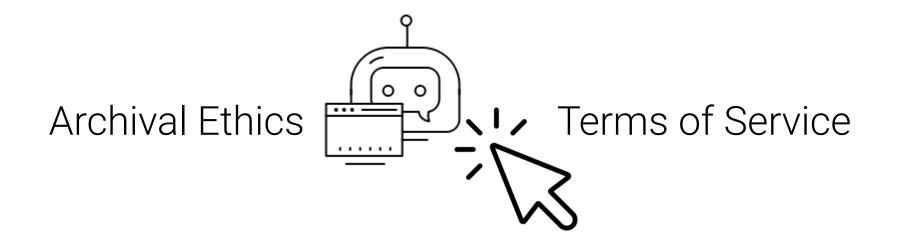


#### **Ethical Considerations**

The emphasis on legal accountability, while useful in the case of atrocities, ignores more subtle forms of discrimination and harm.

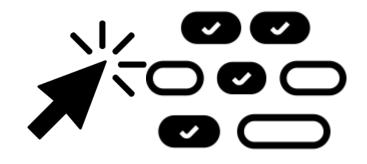
- Michelle Caswell and Marika Cifor

#### **Ethical Considerations**



## Web Archiving as Appraisal

Web archiving is fundamentally an appraisal process. - Amy Wickner



#### Representation

For digital collections, who gets represented is closely tied to who writes the software, who builds the tools, who produces the technical standards, and who provides the funding or other resources for that work.

- Bergis Jules

## Representation



## Potential for Re/Producing Harm

Knowledge of how marginalized communities exist on the web is one of the most important ethical considerations and a powerful check against causing harm to those communities through archiving practices.

- Bergis Jules, Ed Summers, Dr. Vernon Mitchell

## Potential for Re/Producing Harm

- Surveillance
- Hate speech
- Online Harassment



## Feminist Ethics of Care

Radical empathy as a means of shifting affective responsibilities within archival practice may be mapped to the specificities of web archiving:

- Archivist + Record Creator
- Archivist + Subjects of Records
- Archivist + User
- Archivist + Community

## Feminist Ethics of Care

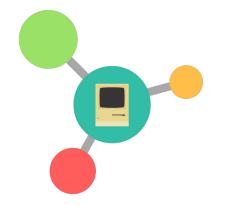
"Who gets to become an archivist, how archives get organized, and even what counts as an archive have a profound racial impact on what endures as valued historical research. Expansive, digital archives can still be locked behind paywalls or library turnstiles at elite universities. Brick and mortar archives stand in racially segregated parts of town. In the most concrete ways possible, racial politics determine how we locate the past."

- N.D.B. Connolly

## Web Archiving Ethics

The work we do as archivists, as librarians, as digital preservationists, have real consequences for marginalized people because who is remembered and how they're remembered dictates who gets violence perpetrated against them. - Bergis Jules

#### Questions?



### Capacity Considerations



13. What are the top considerations for the development of your web arc	hiving program? Choose
three.	
Access and use (e.g., researcher interactions, web analytics, use cases)	
Cost (e.g., budgeting, service allowance utilization, staffing level requirements)	
Data volume (e.g., data volume collected, objects collected, acquisitions statistics)	
Institutional buy-in (e.g., programmatic growth, stakeholder testimonials, resource co	mmitments)
Loss (e.g., link and/or reference rot of archived resources)	
Quality (e.g., accuracy, completeness, replay fidelity)	80%
Risk management (e.g., permission responses, takedown requests, policy confe	_
Other (please specify)	<b>70%</b>
	60% 60% 59%



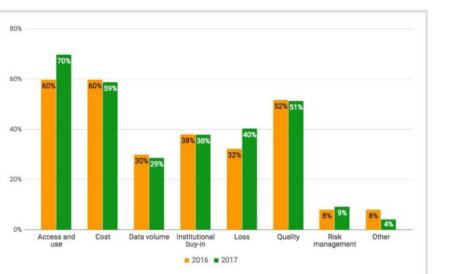
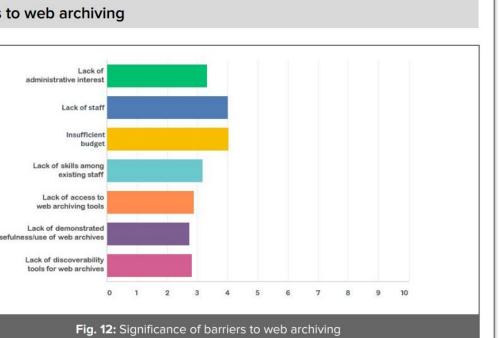


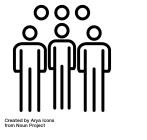
FIGURE 9: Top considerations for the development of respondents' programs

⊢ A R C H I V E For publication Ē NTERN nyarc **Advancing Art Libraries and Curated** Web Archives: A Community Survey November 14, 2018 Barriers to web archiving Lack of administrative interest Internet Archive Lack of staff New York Art Resources Consortium Insufficient budget Persistent URL: https://archive.org/details/AALCWACommunitySurvey Lack of skills among existing staff Lack of access to web archiving tools Lack of demonstrated usefulness/use of web archives Lack of discoverability tools for web archives 0 2 3 4



## Capacity Considerations

- Identify barriers to entry:
  - Buy-in or administrative support
  - Knowledge and experience with web archiving practice
  - Resources and what you can budget for:
    - Funds
    - Available staff





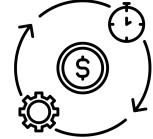
Created by Smalllike from Noun Project



Emoji One, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0 >, via Wikimedia Commons

## Capacity Considerations

- Focus on ways to overcome those barriers:
  - Repurposing of budget
  - Training
- Evaluate existing strengths within the institution/current workflows:
  - Systems and tools presently in use
  - Extension of workflows to incorporate web-based materials



Created by Normansyah from Noun Project

Created by Adrien Coquet from Noun Project

# "Making the case" to your administration

- What are you asking for?
- Why do we need to do it?
- Who else is doing this (peer institutions)?
- How much will it cost/resources required?





Created by Normansyah from Noun Project

## "Making the case" to your administration

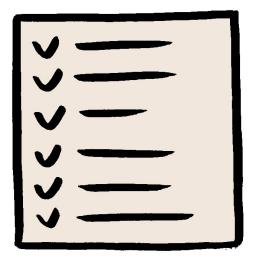


Image: Jurgen Appelo https://creativecommons.org/licenses/by/2.0/

- Compile scenarios for establishing programs
- Tie into mission & existing collecting practice
- Share "good" practice
- Develop cost modeling

# Resources for "making the case"

- Narrative and use cases to convince stakeholders/admin that web-based information is highly ephemeral and valuable
- Pivotal times to frame a significant argument for web archiving:
  - Website redesign
  - Shuttering of an organization
  - Events/local and community issues
- Metrics from prior surveys
- Benchmarking and estimates



Created by Adrien Coquet from Noun Project

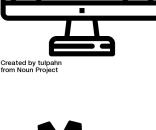
## Resources for "making the case"

- Examine areas where you presently collaborate with other organizations that may transition well to addressing web archiving
- Determine what resources you already have that might be made more available (financial, staff, and systems/tools)
- Look at areas of flexibility within your budget
- Consider strategies for allocating acquisitions funds to web archive collection building

## Tools & Software Subscriptions

- Examine your needs and capacity in relation to available tools:
  - Paid versus free or open source
  - Scale & frequency of capture
  - Quality assurance

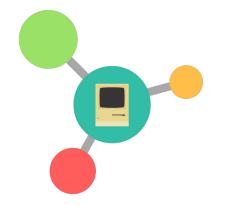




. . .

Created by i cons from Noun Project

#### Questions?



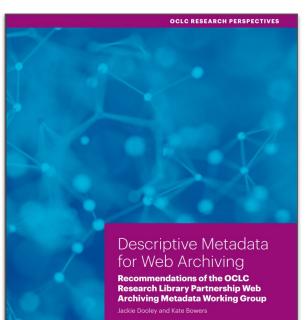
#### Description, Metadata & Access



Bruce The deus, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

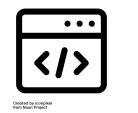
## Archival vs. Bibliographic Description

- Bibliographic description approaches:
  - Collection level vs. seed level records
  - Item level records
- Finding Aids



### Metadata & Systems

- Archive-It:
  - Built-in template for Dublin Core fields (standard + custom fields)
  - OAI-PMH metadata:
    - OCLC WorldCat for collection-level records
  - ArchivesSpace integration



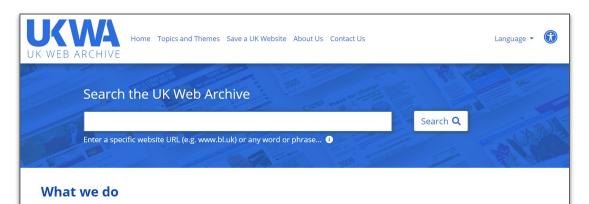


Created by Ahmad Roaayala from Noun Project



## Search Interfaces & API Integration

- UK Web Archive search portal
- Princeton Theological Seminary: CDX/C API
- NYARC Discovery: OpenSearch API



The UK Web Archive (UKWA) collects millions of websites each year, preserving them for future generations. Use this site to discover old or obsolete versions of UK websites, search the text of the websites and browse websites curated on different topics and themes.

## UK Web Archive: Topics and Themes

Topics and Themes are groups of websites brought together on a particular theme by librarians, curators and other specialists, often working in collaboration with key organisations in the field.



#### **British Stand-up Comedy Archive**

Collection owned and administered by Elspeth Millar.



#### **French in London**

This collection of websites has been selected by Saskia Huc-Hepher.

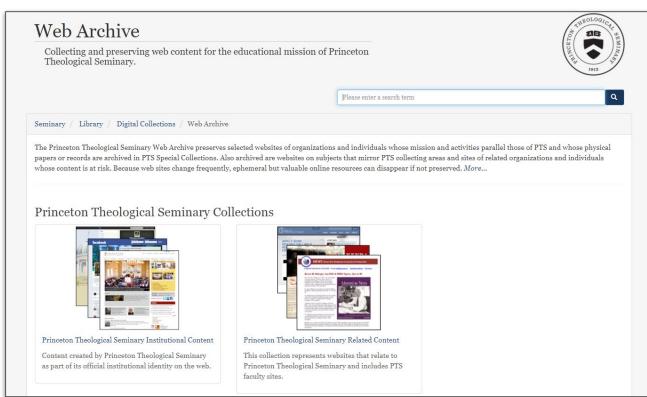
View all Topics and Themes



#### **News Sites**

558 titles are included in this collection.

## Princeton Theological Seminary



#### Princeton: CDX/C API

Title	new jersey — Tags — Browse — Presbyterian Mission Agency			
Site	www.presbyterianmission.org			
URL	https://www.presbyterianmission.org/browse/ministries-ministrycontent/tag/new%20 jersey/linew%20 jersey/line			
Туре	Web page			
Collection	Religious Organizations			
Continue reading saturday, August 68 The Presbytery of New Brunswick New Jersey The Rev. Anand Veeraraj and Sanjay Alexander provide leadership for the New Jersey Indian Church. This document was captured on April 27, 2014 So more results from www.presbyterianmic.con.org				
Title	Princeton Historic District (Princeton, New Jersey) - Wikipedia			
Site	en.wikipedia.org			
URL	https://en.wikipedia.org/wiki/Princeton_Historic_District_(Princeton,_New_Jersey)			
Туре	Web page			
Collection	Princeton Theological Seminary Related Content			
Registers of H This document	ter of Historic Place Vistings in Mercer County, New Jersey List of National Historic Landmarks in New Jersey References [edit] ^ "New Jersey and National istoric Places t was captured on <u>January 22, 2019</u> Its from en.wikipedia.org			

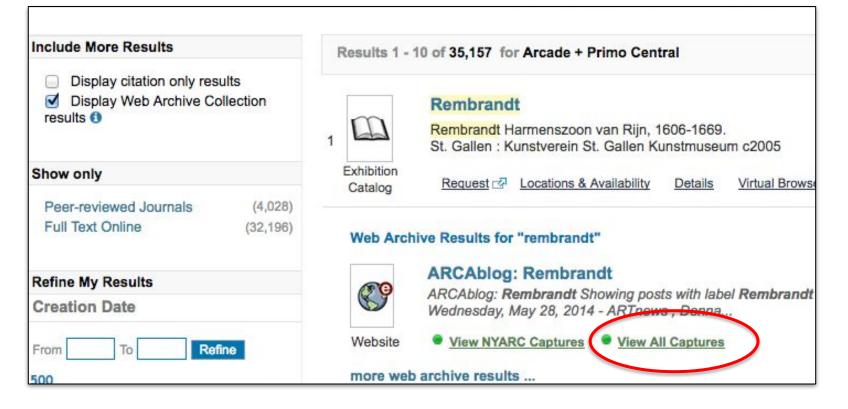
## NYARC Discovery

Books, Articles & More Arcade	Search Advanced Search Browse Search
items v keywords v	
NYARC Discovery is a new research tool from the libraries of the Brooklyn Museum, The Frick Collection, and The Museum of Modern Art. It is a gateway to a trove of rich and varied materials, much of it unique, on art and cultural history spanning the spectrum from ancient Egypt to contemporary art. Find books, exhibition catalogs, auction sale catalogs, artists' books, images, online journal articles, historical newspapers, archives, web archives, photoarchives, and more.	Books, Articles & More - Arcade journal articles e-books & newspaper articles journal titles
What is NYARC?	dissertations auction catalogs
The New York Art Resources Consortium (NYARC) is a collaboration among the libraries of three museums in New York City: Brooklyn Museum, The Frick Collection, and The Museum of Modern Art.	images archives
What is Arcade?	nyarc photoarchives web archives & more

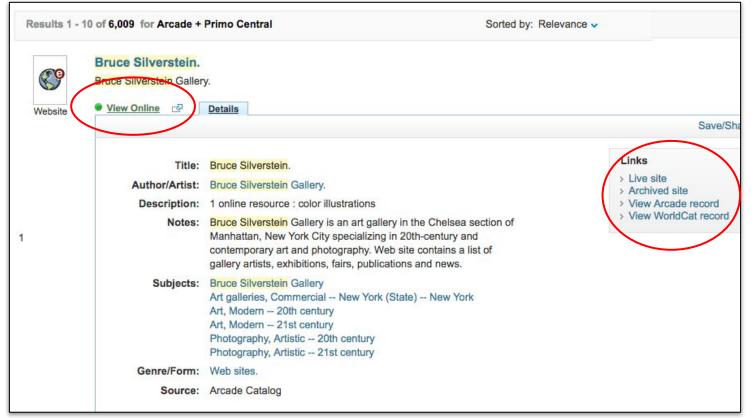
## NYARC Discovery: OpenSearch API

New York Art Resources Conso			
Books, Articles & More Arcade rembrandt		Search	
		Help us improve NYARC Discovery by sending	
Include More Results ☐ Display citation only results ☑ Display Web Archive Collection results ③	Results 1 -	10 of 35,157 for Arcade + Primo Central Rembrandt Rembrandt Harmenszoon van Rijn, 1606-1669.	
Show only	Exhibition Catalog	St. Gallen : Kunstverein St. Gallen Kunstmuseum c2005 Request 대과 Locations & Availability Details <u>Virtual Browse</u>	
Peer-reviewed Journals(4,028)Full Text Online(32,196)	Web Archive Results for "rembrandt"		
Refine My Results Creation Date	S	ARCAblog: Rembrandt ARCAblog: Rembrandt Showing posts with label Rembrandt Wednesday, May 28, 2014 - ARTnews , Donna	
From To Refine	Website more web	<u>View NYARC Captures</u> <u>View All Captures</u> archive results	

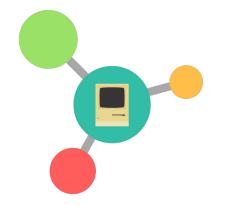
## NYARC Discovery: OpenSearch API



#### NYARC: Seed Level Records



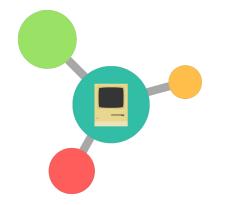
#### Questions?



#### Break: 15 minutes

#### Worksheets

#### Discussion/Questions



# Thank you!

