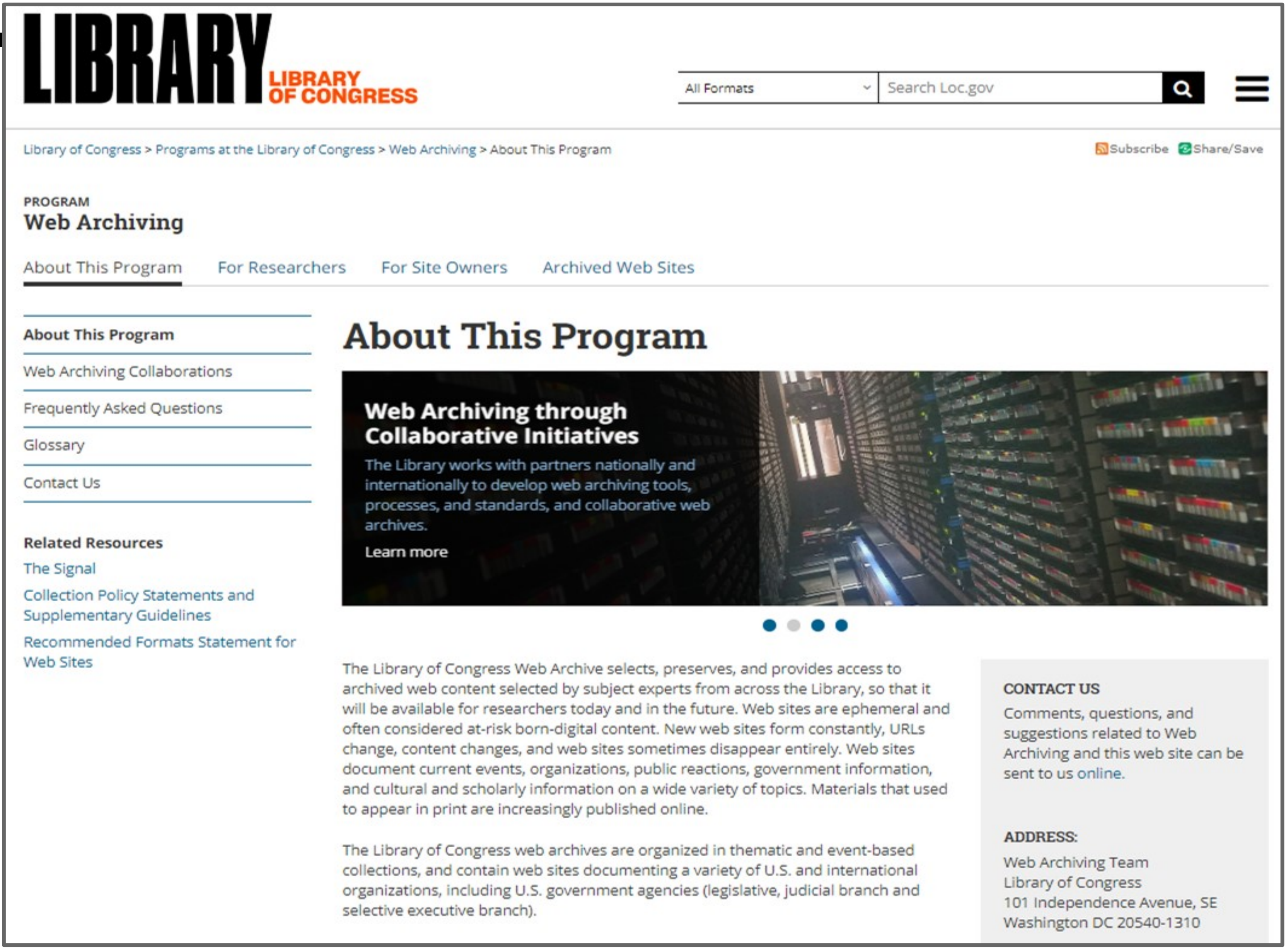
Library of Congress Web Archiving: Selective Archiving at Scale

DPC Web Archiving & Preservation Webinar Januarv 30. 2020

Abbie Grotke
Lead Librarian, Web Archiving Team
abgr@loc.gov
@agrotke

Program Overview



Library of Congress > Programs at the Library of Congress > Web Archiving > About This Program

PROGRAM

Web Archiving

About This Program | For Researchers | For Site Owners | Archived Web Sites

- **About This Program**
- Web Archiving Collaborations
- Frequently Asked Questions
- Glossary
- Contact Us

Related Resources

The Signal

Collection Policy Statements and Supplementary Guidelines

Recommended Formats Statement for Web Sites

About This Program

Web Archiving through Collaborative Initiatives

The Library works with partners nationally and internationally to develop web archiving tools, processes, and standards, and collaborative web archives.

Learn more

The Library of Congress Web Archive selects, preserves, and provides access to archived web content selected by subject experts from across the Library, so that it will be available for researchers today and in the future. Web sites are ephemeral and often considered at-risk born-digital content. New web sites form constantly, URLs change, content changes, and web sites sometimes disappear entirely. Web sites document current events, organizations, public reactions, government information, and cultural and scholarly information on a wide variety of topics. Materials that used to appear in print are increasingly published online.

The Library of Congress web archives are organized in thematic and event-based collections, and contain web sites documenting a variety of U.S. and international organizations, including U.S. government agencies (legislative, judicial branch and selective executive branch).

CONTACT US

Comments, questions, and suggestions related to Web Archiving and this web site can be sent to us online.

ADDRESS:

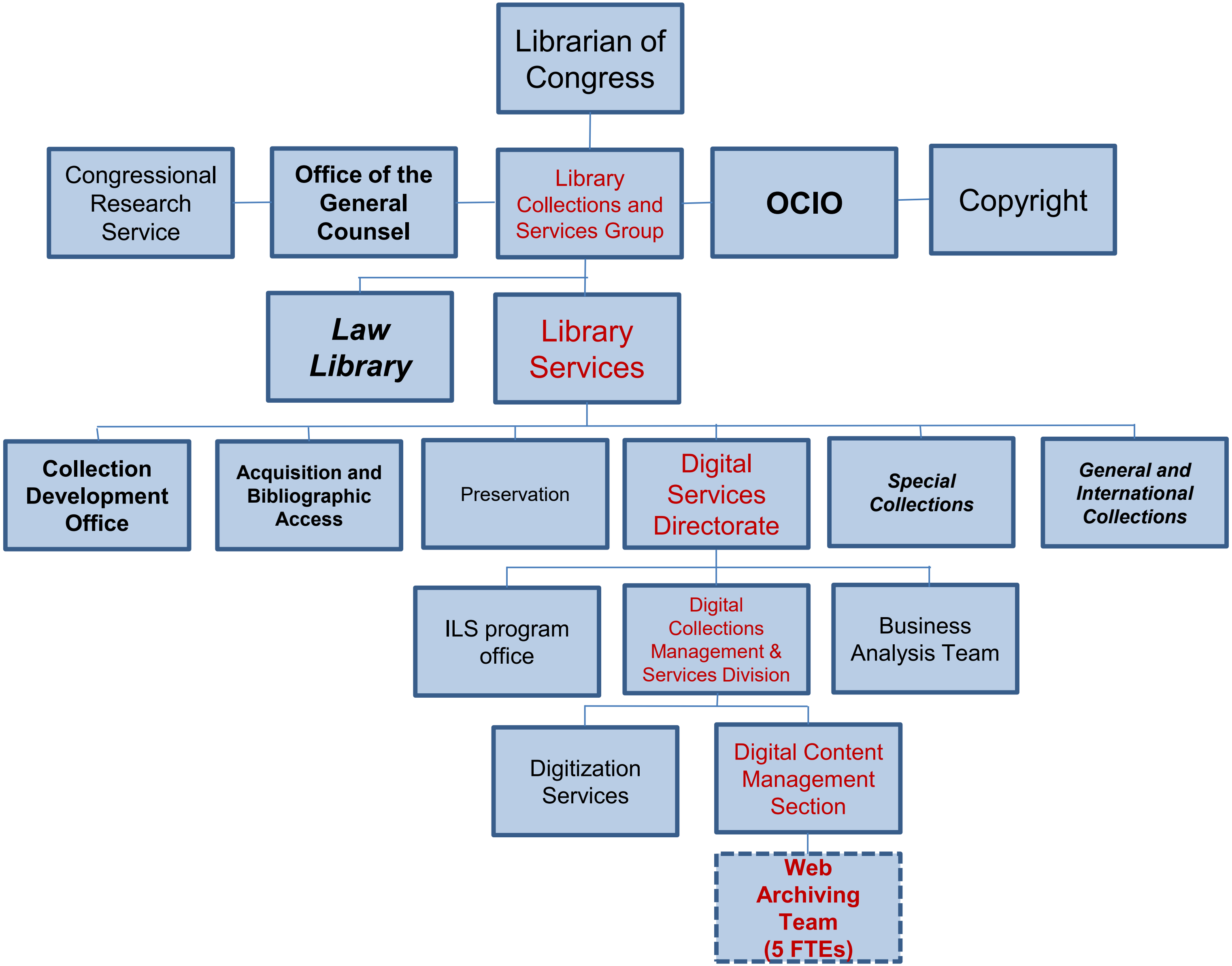
Web Archiving Team
Library of Congress
101 Independence Avenue, SE
Washington DC 20540-1310

https://www.loc.gov/programs/web-archiving/about-this-program/

https://www.loc.gov/websites/

http://webarchive.loc.gov/

Where is Web Archiving organizationally within LC?



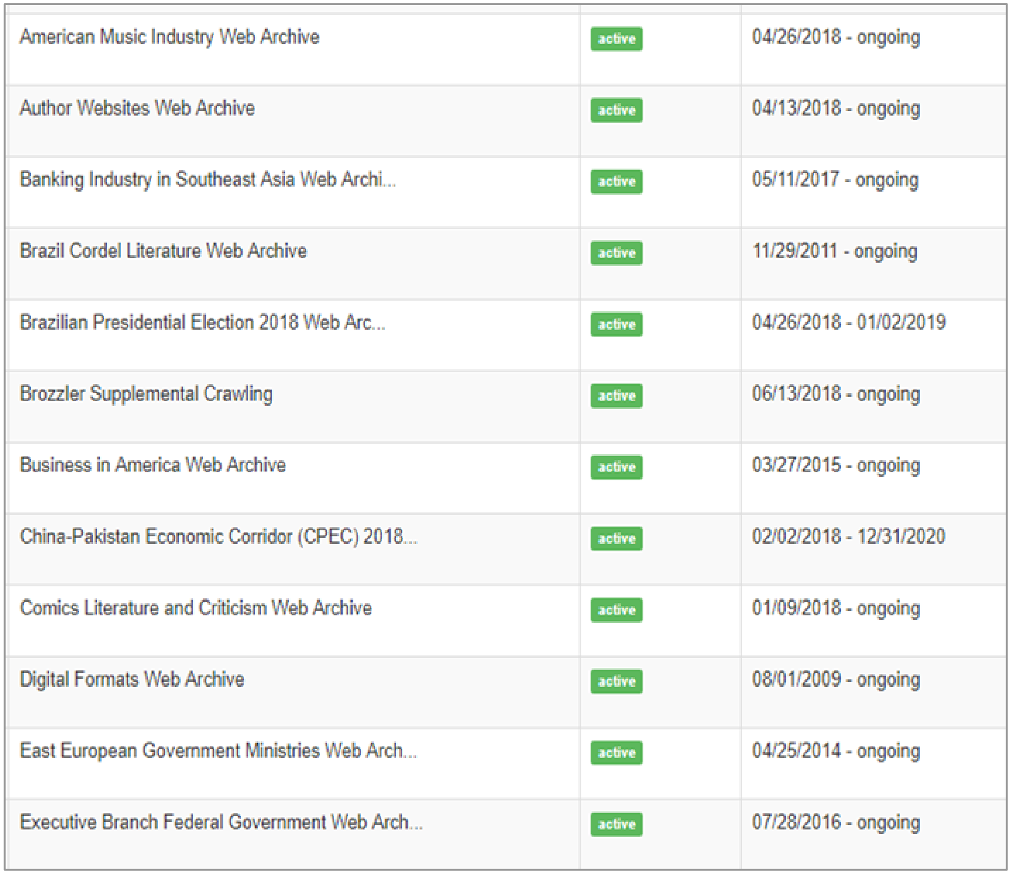
Librarian of Congress
Congressional Research Service
Office of the General Counsel
Library Collections and Services Group
OCIO
Copyright
Law Library
Library Services
Collection Development Office
Acquisition and Bibliographic Access
Preservation
Digital Services Directorate
Special Collections
General and International Collections
ILS program office
Digital Collections Management & Services Division
Business Analysis Team
Digitization Services
Digital Content Management Section
Web Archiving Team (5 FTEs)

https://www.loc.gov/static/portals/about/documents/library-congress-orgchart-043019.pdf

Our Technical Approach and Tools

- Acquisition
    - Primarily outsourced crawling under contract with Internet Archive
        - IA manages and runs crawls (Heritrix and Brozzler)
        - We do QA before transferring to the Library for preservation and access
    - Limited crawling in-house
        - Heritrix
        - Some experimentation/testing of WebRecorder
    - .gov content from 1996-2001 acquired via backfile purchase from IA
    - Copies of data collected through collaborative efforts (such as End of Term project)
- Access
    - Open Wayback
    - Some testing of Pywb
    - Collections integrated into loc.gov – MODS records are searchable
    - No full-text indexing
    - Some datasets available through https://labs.loc.gov/experiments/webarchive-datasets/
- Storage and Processing
    - LC infrastructure and tools used:
        - Content Transfer Services copy to long term tape storage and a copy for access
        - Participated in recent experimentations with cloud processing of collections

The LC Approach: Event and Thematic Collections



American Music Industry Web Archive	active	04/26/2018 - ongoing
Author Websites Web Archive	active	04/13/2018 - ongoing
Banking Industry in Southeast Asia Web Archi...	active	05/11/2017 - ongoing
Brazil Cordel Literature Web Archive	active	11/29/2011 - ongoing
Brazilian Presidential Election 2018 Web Arc...	active	04/26/2018 - 01/02/2019
Brozzler Supplemental Crawling	active	06/13/2018 - ongoing
Business in America Web Archive	active	03/27/2015 - ongoing
China-Pakistan Economic Corridor (CPEC) 2018...	active	02/02/2018 - 12/31/2020
Comics Literature and Criticism Web Archive	active	01/09/2018 - ongoing
Digital Formats Web Archive	active	08/01/2009 - ongoing
East European Government Ministries Web Arch...	active	04/25/2014 - ongoing
Executive Branch Federal Government Web Arch...	active	07/28/2016 - ongoing

Library of Congress Collection Policy Statements:
http://www.loc.gov/acq/devpol/cpsstate.html
Web Archiving Supplemental Guidelines:
http://www.loc.gov/acq/devpol/webarchive.pdf

Our web archive collections are typically:

- **Thematic or subject-focused** (e.g., Authors Web Archive, LGBTQ Studies Web Archive, Web Cultures Web Archive)

OR

- **Event-focused** (e.g.; national or foreign elections)

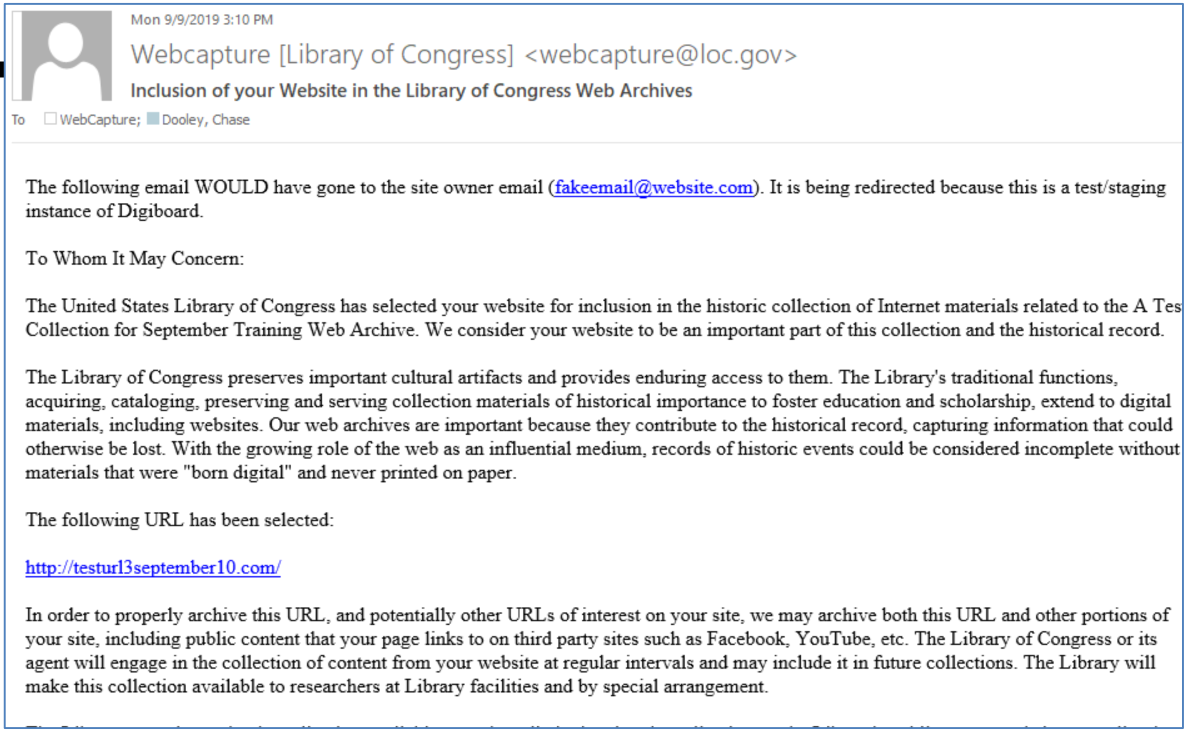
We manage over **150** collections

63 active are currently ACTIVE, event-based collections

9 are “administrative” collections to help us manage our crawls

We follow a permissions-based approach

- Permissions/notices typically have to be sent for anything selected for web archiving
- Permissions are currently based on the **COUNTRY of PUBLICATION** and the **CATEGORY** of site.



Mon 9/9/2019 3:10 PM

Webcapture [Library of Congress] <webcapture@loc.gov>

Inclusion of your Website in the Library of Congress Web Archives

To □ WebCapture; Dooley, Chase

The following email WOULD have gone to the site owner email (fakeemail@website.com). It is being redirected because this is a test/staging instance of Digiboard.

To Whom It May Concern:

The United States Library of Congress has selected your website for inclusion in the historic collection of Internet materials related to the A Tes Collection for September Training Web Archive. We consider your website to be an important part of this collection and the historical record.

The Library of Congress preserves important cultural artifacts and provides enduring access to them. The Library's traditional functions, acquiring, cataloging, preserving and serving collection materials of historical importance to foster education and scholarship, extend to digital materials, including websites. Our web archives are important because they contribute to the historical record, capturing information that could otherwise be lost. With the growing role of the web as an influential medium, records of historic events could be considered incomplete without materials that were "born digital" and never printed on paper.

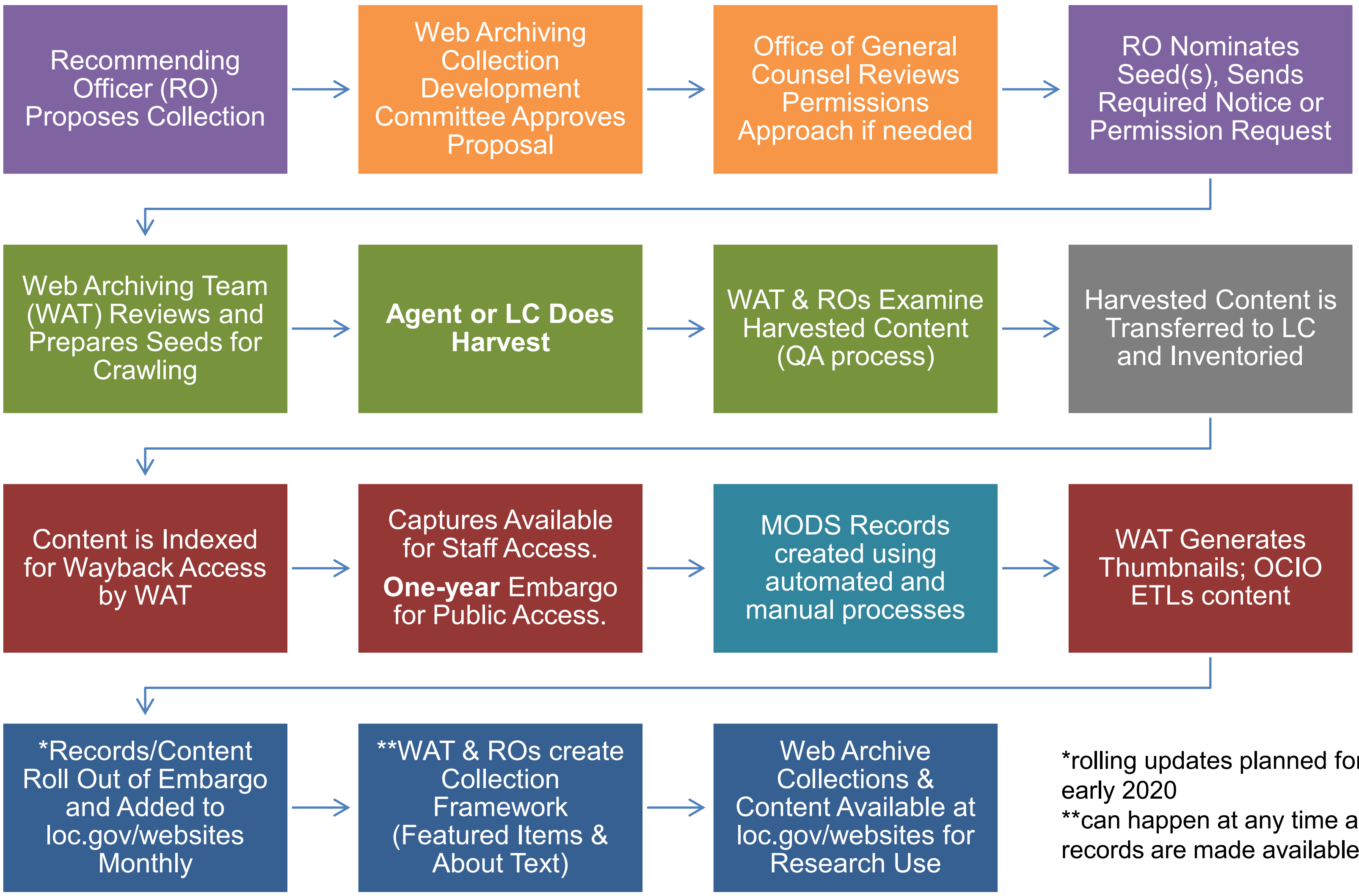
The following URL has been selected:

http://testurl3september10.com/

In order to properly archive this URL, and potentially other URLs of interest on your site, we may archive both this URL and other portions of your site, including public content that your page links to on third party sites such as Facebook, YouTube, etc. The Library of Congress or its agent will engage in the collection of content from your website at regular intervals and may include it in future collections. The Library will make this collection available to researchers at Library facilities and by special arrangement.

- **Notice/Notice:** We can crawl once notice is sent and make accessible after the embargo period
- **Notice/Permission:** We can crawl once the notice is sent, and make available offsite if the site owner grants permission, otherwise it is made available onsite after the embargo period.
- **Permission/Permission:** We must have explicit permission to crawl and make accessible offsite.
- **No Notice:** US Government sites and those with creative commons

The LC Web Archiving Process



Recommending Officer (RO) Proposes Collection

Web Archiving Collection Development Committee Approves Proposal

Office of General Counsel Reviews Permissions Approach if needed

RO Nominates Seed(s), Sends Required Notice or Permission Request

Web Archiving Team (WAT) Reviews and Prepares Seeds for Crawling

Agent or LC Does Harvest

WAT & ROs Examine Harvested Content (QA process)

Harvested Content is Transferred to LC and Inventoried

Content is Indexed for Wayback Access by WAT

Captures Available for Staff Access.

One-year Embargo for Public Access.

MODS Records created using automated and manual processes

WAT Generates Thumbnails; OCIO ETLs content

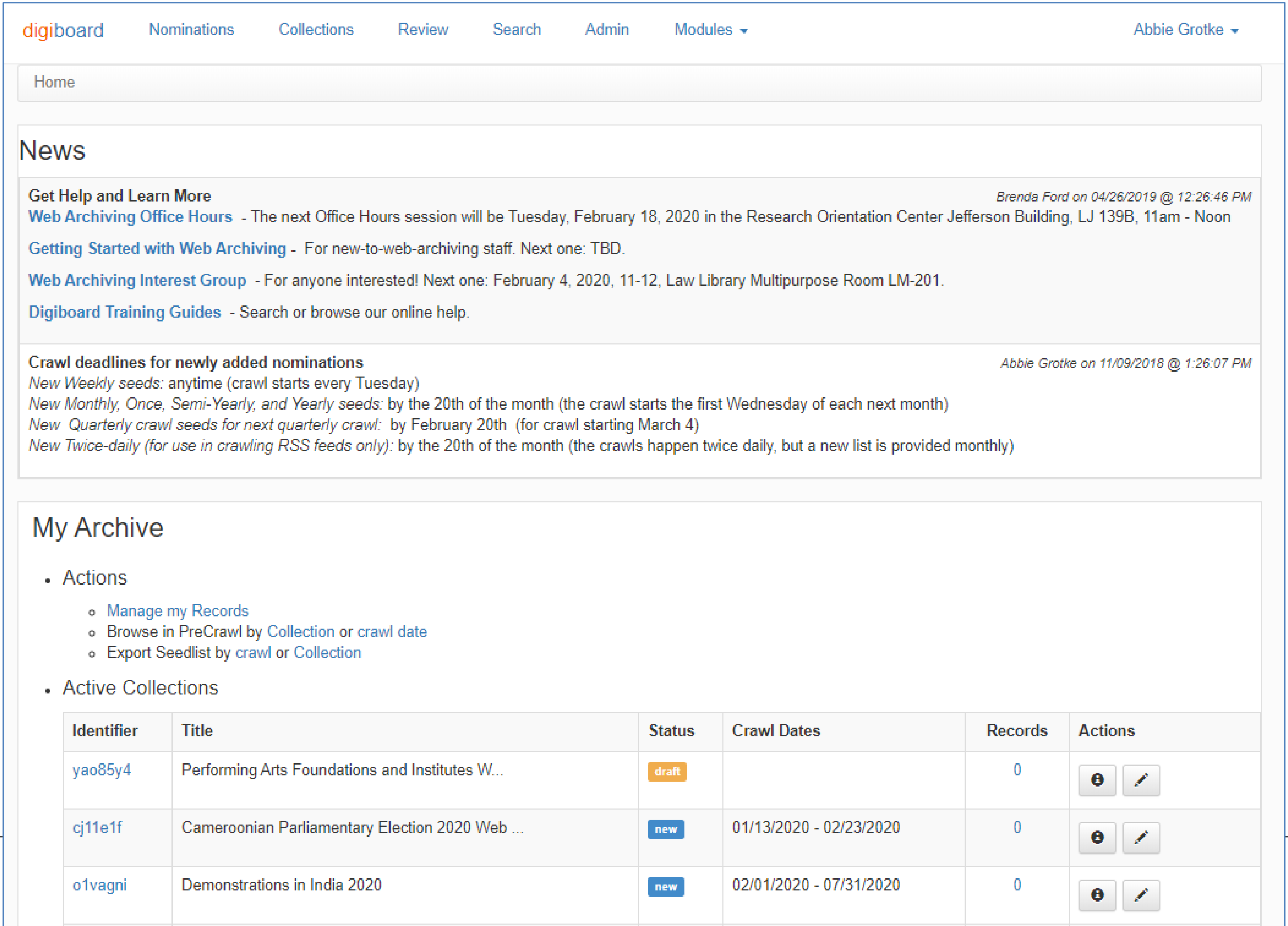
*Records/Content Roll Out of Embargo and Added to loc.gov/websites Monthly

**WAT & ROs create Collection Framework (Featured Items & About Text)

Web Archive Collections & Content Available at loc.gov/websites for Research Use

*rolling updates planned for early 2020
**can happen at any time after records are made available

Digiboard: Our Web Archiving Workflow Tool



News

Get Help and Learn More *Brenda Ford on 04/26/2019 @ 12:26:46 PM*

Web Archiving Office Hours - The next Office Hours session will be Tuesday, February 18, 2020 in the Research Orientation Center Jefferson Building, LJ 139B, 11am - Noon

Getting Started with Web Archiving - For new-to-web-archiving staff. Next one: TBD.

Web Archiving Interest Group - For anyone interested! Next one: February 4, 2020, 11-12, Law Library Multipurpose Room LM-201.

Digiboard Training Guides - Search or browse our online help.

Crawl deadlines for newly added nominations *Abbie Grotke on 11/09/2018 @ 1:26:07 PM*

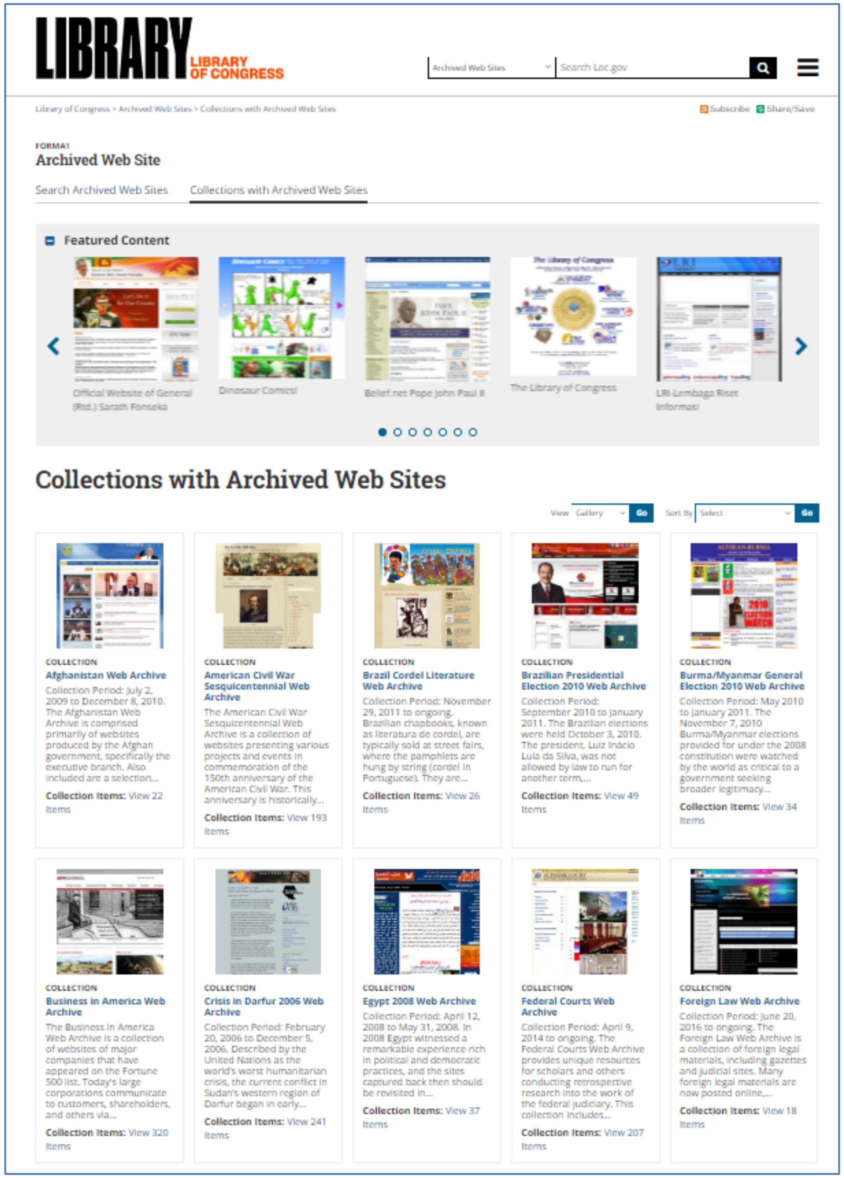
New Weekly seeds: anytime (crawl starts every Tuesday)
New Monthly, Once, Semi-Yearly, and Yearly seeds: by the 20th of the month (the crawl starts the first Wednesday of each next month)
New Quarterly crawl seeds for next quarterly crawl: by February 20th (for crawl starting March 4)
New Twice-daily (for use in crawling RSS feeds only): by the 20th of the month (the crawls happen twice daily, but a new list is provided monthly)

My Archive

- Actions
  - Manage my Records
  - Browse in PreCrawl by Collection or crawl date
  - Export Seedlist by crawl or Collection
- Active Collections

Identifier	Title	Status	Crawl Dates	Records	Actions
yao85y4	Performing Arts Foundations and Institutes W...	draft		0	
cj11e1f	Cameroonian Parliamentary Election 2020 Web ...	new	01/13/2020 - 02/23/2020	0	
o1vagni	Demonstrations in India 2020	new	02/01/2020 - 07/31/2020	0	

Access: loc.gov/websites/collections/



Library of Congress > Archived Web Sites > Collections with Archived Web Sites

FORMAT

Archived Web Site

Search Archived Web Sites | Collections with Archived Web Sites

Featured Content

- Official Website of General (Rtd.) Sarath Fonseka
- Dinosaur Comics!
- Belief.net Pope John Paul II
- The Library of Congress
- LRI-Lembaga Riset Informasi

Collections with Archived Web Sites

COLLECTION
Afghanistan Web Archive
Collection Period: July 2, 2009 to December 8, 2010. The Afghanistan Web Archive is comprised primarily of websites produced by the Afghan government, specifically the executive branch. Also included are a selection...
Collection Items: View 22 Items

COLLECTION
American Civil War Sesquicentennial Web Archive
The American Civil War Sesquicentennial Web Archive is a collection of websites presenting various projects and events in commemoration of the 150th anniversary of the American Civil War. This anniversary is historically...
Collection Items: View 193 Items

COLLECTION
Brazil Cordel Literature Web Archive
Collection Period: November 29, 2011 to ongoing. Brazilian chapbooks, known as literatura de cordel, are typically sold at street fairs, where the pamphlets are hung by string (cordel in Portuguese). They are...
Collection Items: View 26 Items

COLLECTION
Brazilian Presidential Election 2010 Web Archive
Collection Period: September 2010 to January 2011. The Brazilian elections were held October 3, 2010. The president, Luiz Inácio Lula da Silva, was not allowed by law to run for another term,...
Collection Items: View 49 Items

COLLECTION
Burma/Myanmar General Election 2010 Web Archive
Collection Period: May 2010 to January 2011. The November 7, 2010 Burma/Myanmar elections provided for under the 2008 constitution were watched by the world as critical to a government seeking broader legitimacy...
Collection Items: View 34 Items

COLLECTION
Business in America Web Archive
The Business in America Web Archive is a collection of websites of major companies that have appeared on the Fortune 500 list. Today's large corporations communicate to customers, shareholders, and others via...
Collection Items: View 320 Items

COLLECTION
Crisis in Darfur 2006 Web Archive
Collection Period: February 20, 2006 to December 5, 2006. Described by the United Nations as the world's worst humanitarian crisis, the current conflict in Sudan's western region of Darfur began in early...
Collection Items: View 241 Items

COLLECTION
Egypt 2008 Web Archive
Collection Period: April 12, 2008 to May 31, 2008. In 2008 Egypt witnessed a remarkable experience rich in political and democratic practices, and the sites captured back then should be revisited in...
Collection Items: View 37 Items

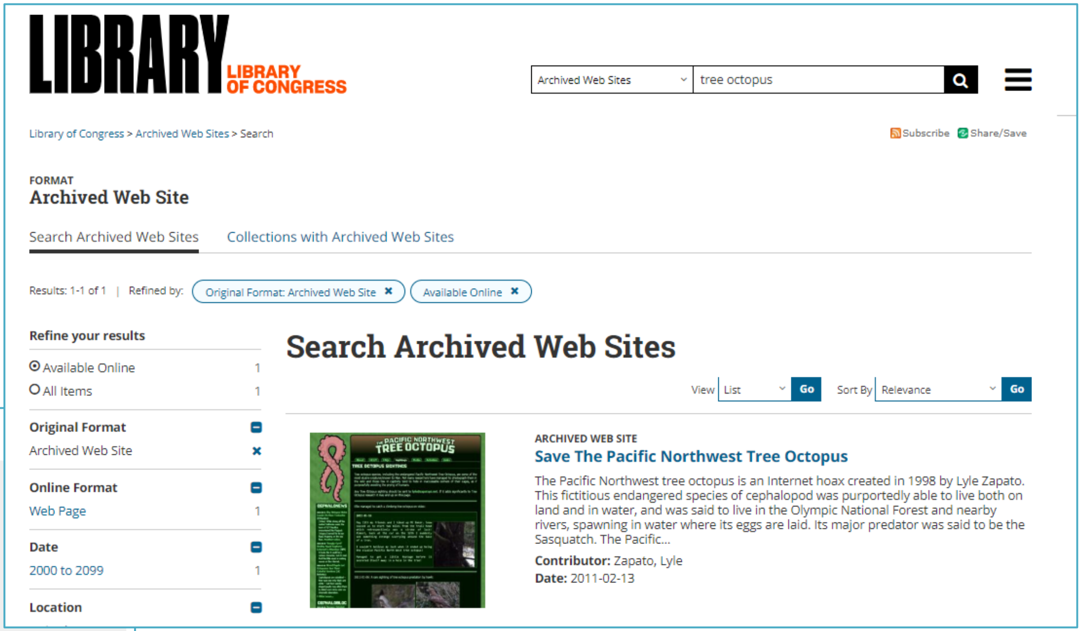
COLLECTION
Federal Courts Web Archive
Collection Period: April 9, 2014 to ongoing. The Federal Courts Web Archive provides unique resources for scholars and others conducting retrospective research into the work of the federal judiciary. This collection includes...
Collection Items: View 207 Items

COLLECTION
Foreign Law Web Archive
Collection Period: June 20, 2016 to ongoing. The Foreign Law Web Archive is a collection of foreign legal materials, including gazettes and judicial sites. Many foreign legal materials are now posted online,...
Collection Items: View 18 Items

- 21,729 web archives available
- Records from over 97 collections available
- 63 collections with contextual (aka framework) material, more on the way!
- All content is embargoed for (at least) one year after capture
- Some content restricted to onsite only; the rest is available from anywhere
- Rolling updates starting soon

Record search at loc.gov

records describing each archived item are searchable alongside other digital collections at the Library



Archived Web Sites
tree octopus

Library of Congress > Archived Web Sites > Search

Subscribe Share/Save

FORMAT
Archived Web Site

Search Archived Web Sites | Collections with Archived Web Sites

Results: 1-1 of 1 | Refined by: Original Format: Archived Web Site ✖ Available Online ✖

Refine your results

● Available Online	1
○ All Items	1
Original Format	
Archived Web Site	✖
Online Format	
Web Page	1
Date	
2000 to 2099	1
Location	

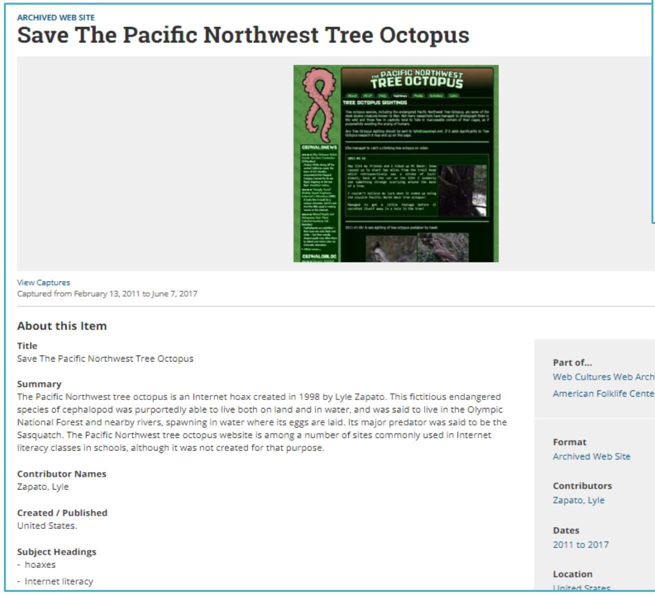
Search Archived Web Sites

View List Go Sort By Relevance Go

ARCHIVED WEB SITE
Save The Pacific Northwest Tree Octopus

The Pacific Northwest tree octopus is an Internet hoax created in 1998 by Lyle Zapato. This fictitious endangered species of cephalopod was purportedly able to live both on land and in water, and was said to live in the Olympic National Forest and nearby rivers, spawning in water where its eggs are laid. Its major predator was said to be the Sasquatch. The Pacific...

Contributor: Zapato, Lyle
Date: 2011-02-13



ARCHIVED WEB SITE
Save The Pacific Northwest Tree Octopus

View Captures
Captured from February 13, 2011 to June 7, 2017

About this Item

Title
Save The Pacific Northwest Tree Octopus

Summary
The Pacific Northwest tree octopus is an Internet hoax created in 1998 by Lyle Zapato. This fictitious endangered species of cephalopod was purportedly able to live both on land and in water, and was said to live in the Olympic National Forest and nearby rivers, spawning in water where its eggs are laid. Its major predator was said to be the Sasquatch. The Pacific Northwest tree octopus website is among a number of sites commonly used in Internet literacy classes in schools, although it was not created for that purpose.

Contributor Names
Zapato, Lyle

Created / Published
United States.

Subject Headings
- hoaxes
- Internet literacy

Part of...
Web Cultures Web Archive (34)
American Folklife Center (33,122)

Format
Archived Web Site

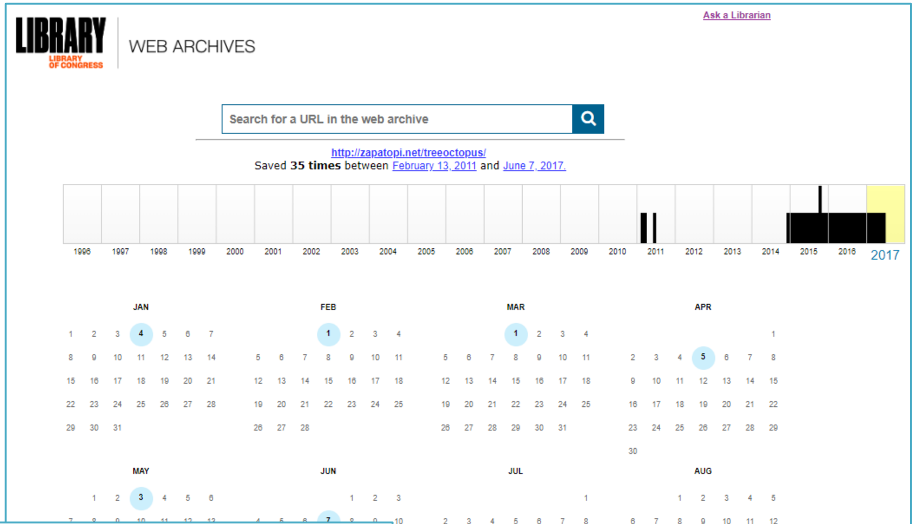
Contributors
Zapato, Lyle

Dates
2011 to 2017

Location
United States

URL search at webarchive.loc.gov

displays the archived content (including uncatalogued sites) up through embargo date



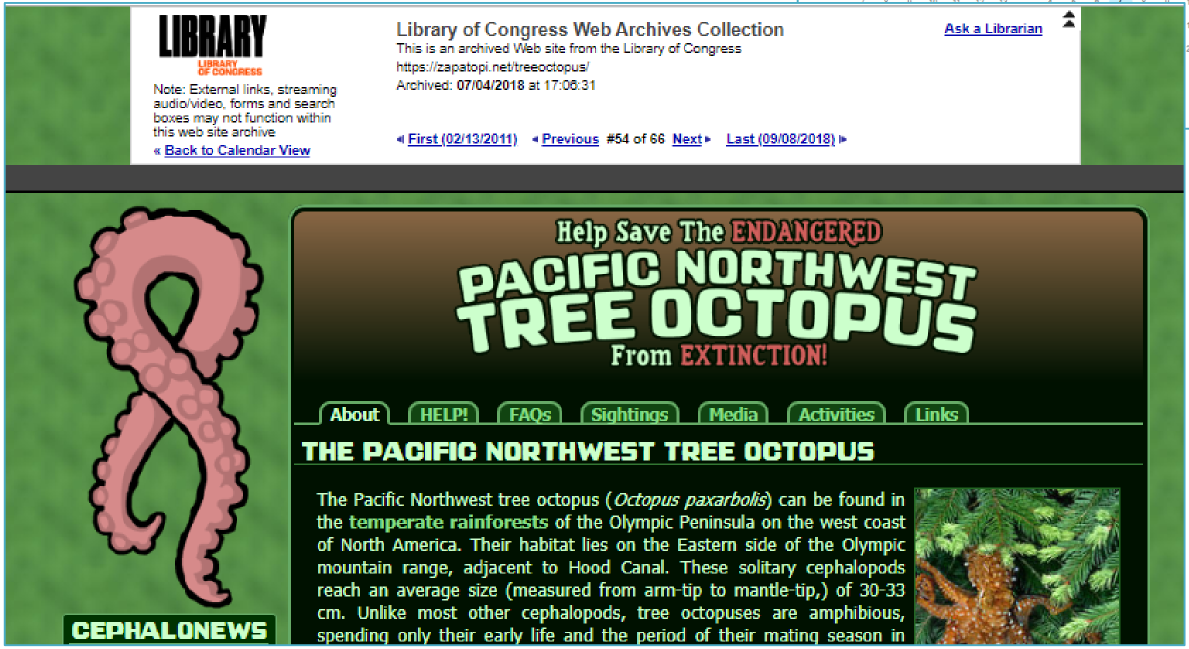
Ask a Librarian

WEB ARCHIVES

Search for a URL in the web archive

http://zapatopi.net/treeoctopus/

Saved **35 times** between February 13, 2011 and June 7, 2017.



Library of Congress Web Archives Collection

This is an archived Web site from the Library of Congress

https://zapatopi.net/treeoctopus/

Archived: 07/04/2018 at 17:06:31

Ask a Librarian

Note: External links, streaming audio/video, forms and search boxes may not function within this web site archive

« Back to Calendar View

◂ First (02/13/2011) ◂ Previous #54 of 66 Next ▸ Last (09/08/2018) ▸

Help Save The ENDANGERED PACIFIC NORTHWEST TREE OCTOPUS From EXTINCTION!

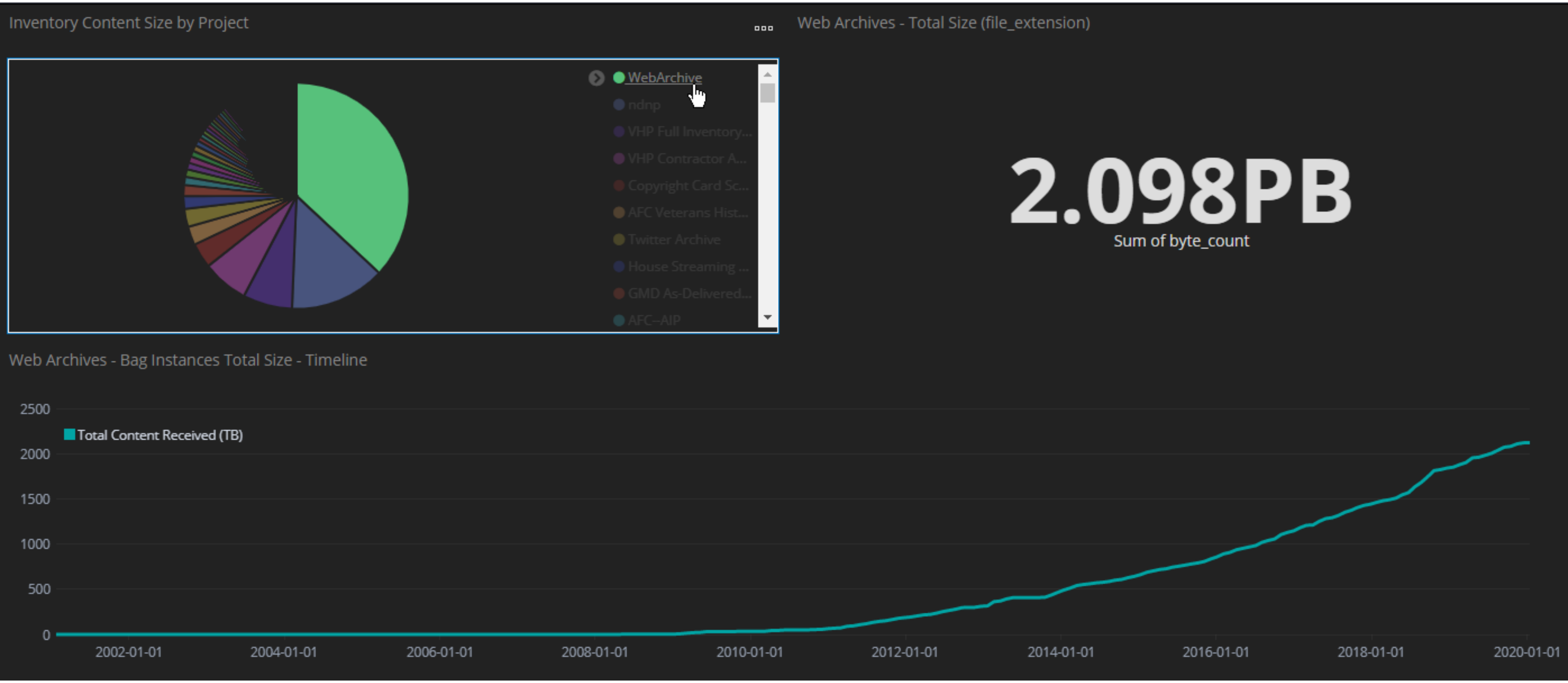
About | HELP! | FAQs | Sightings | Media | Activities | Links

THE PACIFIC NORTHWEST TREE OCTOPUS

The Pacific Northwest tree octopus (*Octopus paxarbolis*) can be found in the temperate rainforests of the Olympic Peninsula on the west coast of North America. Their habitat lies on the Eastern side of the Olympic mountain range, adjacent to Hood Canal. These solitary cephalopods reach an average size (measured from arm-tip to mantle-tip,) of 30-33 cm. Unlike most other cephalopods, tree octopuses are amphibious, spending only their early life and the period of their mating season in

Amount of data in the LC Web Archives

(as of 01/23/20)



Inventory Content Size by Project

Web Archives - Total Size (file_extension)

2.098PB

Sum of byte_count

Web Archives - Bag Instances Total Size - Timeline

Over 18 billion documents

After 2 PB and 20 Years: A Few Lessons Learned

Automate and Reuse Data

- Do things manually for awhile to help understand the process, then figure out where you can automate
- Work with all the data you have
- Automate permissions process as much as possible. Develop permissions letter templates.
- Spreadsheets to manage tasks like QA and permissions are great but only take you so far
- Crawl reports have a ton of data that is just waiting to be used better for QA at scale

Collaborate

- Internally - get to know all of the people who might be able to help you along the way
- Externally – DPC and IIPC are a great community, everyone is willing to help!
- Collaborative collection efforts are great when your own institutional policies make it difficult to collect rapidly or without permission
  - IIPC projects
  - IA-led efforts
  - Collaborations among other partner institutions
  - Everyone has something to contribute

Train and Retrain

- Many hires still don't have extensive web archiving skills, so training off staff takes time and investment (mostly on the job)
- Attend IIPC and other conferences (Archives Unleashed events)
  - Learn new skills –python/scripting have been critical in recent years
- Train staff as needed, and think about refresher training ALL the time
  - For Recommending officers: one-on-one, Office hours, Interest Group discussion, classroom training, etc.

Document Everything

- Documenting decisions is key: put all policy decisions in writing
- Having good, detailed FAQ on our website geared toward site owners
- Develop good training materials for infrequent web archiving staff and for volunteers/interns brought in to help
- Develop a collection proposal template that can be used to help shape a project before it gets started
- Document workflows and processes
- Log and document all decisions for actions taken on seeds

Accept "Good Enough"

- Defined a minimal amount of data that will allow research use but not require too much human effort to catalog
- Use experts/humans to enhance descriptive records
- Don't get too far behind, catching up on backlogs can take years and decades in some cases
- The faster we got our collections out, the more engaged the subject experts became
- We have to accept that our crawls won't be 100% perfect
- Be okay with not QAing everything

Reinvent

- Be flexible
- Always look for ways to reinvent processes or workflows to improve things
- Test new tools as you can, but know that you might be working with the old tools for longer than you'd like (until new ones improve or scale up)
- Help the international community in efforts that will help us all

Thank you !

Abbie Grotke
abgr@loc.gov
@agrotke
webcapture@loc.gov