LIBRARY OF CONGRESS

November 15, 2018

Expansion & Exploration in 2018:

Processing the Library of Congress Web Archive

Grace Thomas & Abbie Grotke
Library of Congress
grth@loc.gov
@gracehthom

Overview

1. Where We Stand: staff, size, collections
2. Increasing Internal Engagement
3. Going Public
4. Digging In: the Cloud, CDX Indexes
5. Future Projects

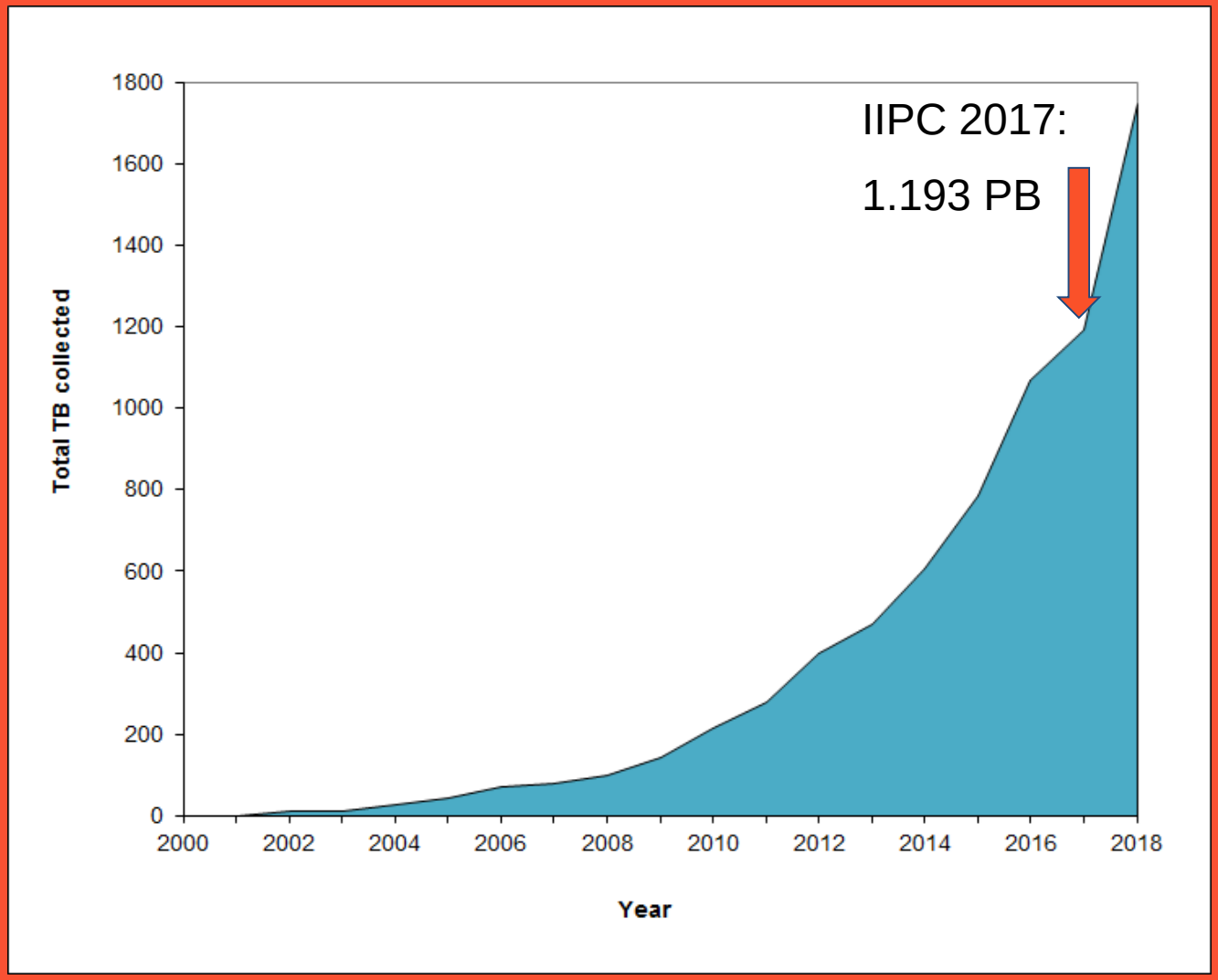
The Team

5 members of the LC Web Archiving Team (still)

We have a new home!

- Library Services
  - Office of the Chief Operating Officer for Library Services
    - Digital Collections Management & Services (DCMS)
      - Digital Content Management Section (DCM)
        - Web Archiving Team!
      - Digitization Services

Archive Size as of October 26, 2018: 1.745 PB



Digital Collecting Plan - February 2017

Strategic Objective 4:

Expand Use of Web Archiving to Acquire Digital Content

*Expand the use of web archiving as a standard acquisitions tool for digital content, such as openly available materials in various digital forms.

https://www.loc.gov/acq/devpol/CollectingDigitalContent.pdf

Event + Thematic Collections

19 new web archives collections started in 2018 (two time-bound):

- American Music Creators
- American Music Industry
- Author Websites
- Brazilian Presidential Election 2018
- Cambodian National Assembly Election 2018
- China-Pakistan Economic Corridor (CPEC) 2018-2020
- Comics Literature & Criticism
- Food & Foodways
- LGBTQ+ Politics & Political Candidates
- LGBTQ+ Studies
- Macedonian Politics
- Malaysian General Election 2018
- Open Content
- State Government Websites of the United States
- United States Election 2018
- Veterans History Project
- Women's and Gender Studies

+ Royal Cremation Ceremony King Bhumibol Adulyadej of Thailand

+ Indonesia Provincial Election 2018

+ 35 ongoing continuing collections started pre-2018

Increasing Internal Engagement

LC Internal Web Archiving Interest Group
Four meetings in 2018

Previous Topics:

- Introduction to Interest Group, reminders about *Digiboard* (proprietary curatorial tool), updates on web archiving news at LC
- Web Archiving Team's process from seed nomination to crawl
- Guests from external institutions to share collection development strategies & activities
- LC selection librarian shares development and evolution of Eastern European Government Ministries collection (3+ years)

Future Topics:

- Navigating Wayback and exploring usage statistics for web archives records on loc.gov + public Wayback
- Explorations of the Web Archiving Team for new modes of access
- Collaborative collection development within LC
- *Digiboard* training opportunities

Recent Web Archive Releases

COLLECTION

Web Cultures Web Archive

Collection Period: April 14, 2014 to ongoing. The Web Cultures Web Archive includes sites documenting the creation and sharing of emergent cultural traditions on the web. The mission of the American Folklife...

View 34 Items

COLLECTION

Webcomics Web Archive

Collection Period: June 4, 2014 to ongoing. This collection focuses on comics created specifically for the web and supplements the Library of Congress' extensive holdings in both comic books, graphic novels, and...

View 44 Items

COLLECTION

Science Blogs Web Archive

Collection Period: January 1, 2013 to ongoing The Science Blogs Web Archive provides resources for scholars and others conducting research on science writing, research, teaching and communication, as well as scientific discourse...

View 70 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,...

View 17 Items

July 2018: Backlog release:

4,240 new web archives across 43 event and thematic collections on loc.gov

https://blogs.loc.gov/thesignal/2018/08/more-web-archives-less-process/

LIBRARY LIBRARY OF CONGRESS

All Formats | Search Loc.gov

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

Subscribe Share/Save

loc.gov/programs/web-archiving/

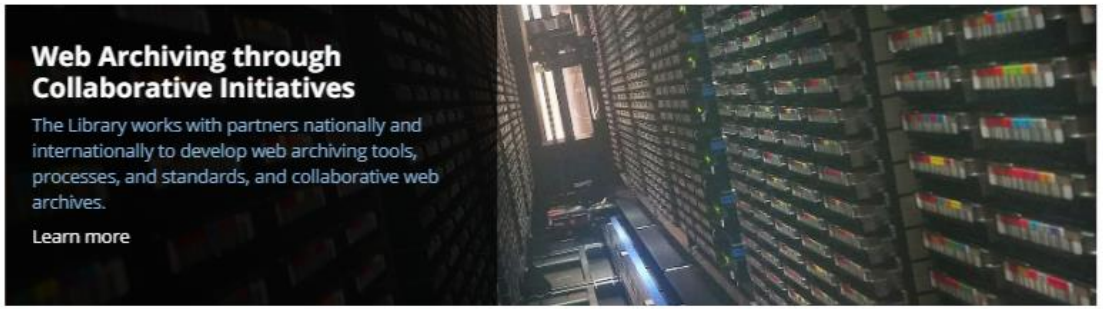
PROGRAM
Web Archiving

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



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).

This site provides general information about program activities, information for researchers who are interested in using the web archives, and information for site owners who might be included in the archives.

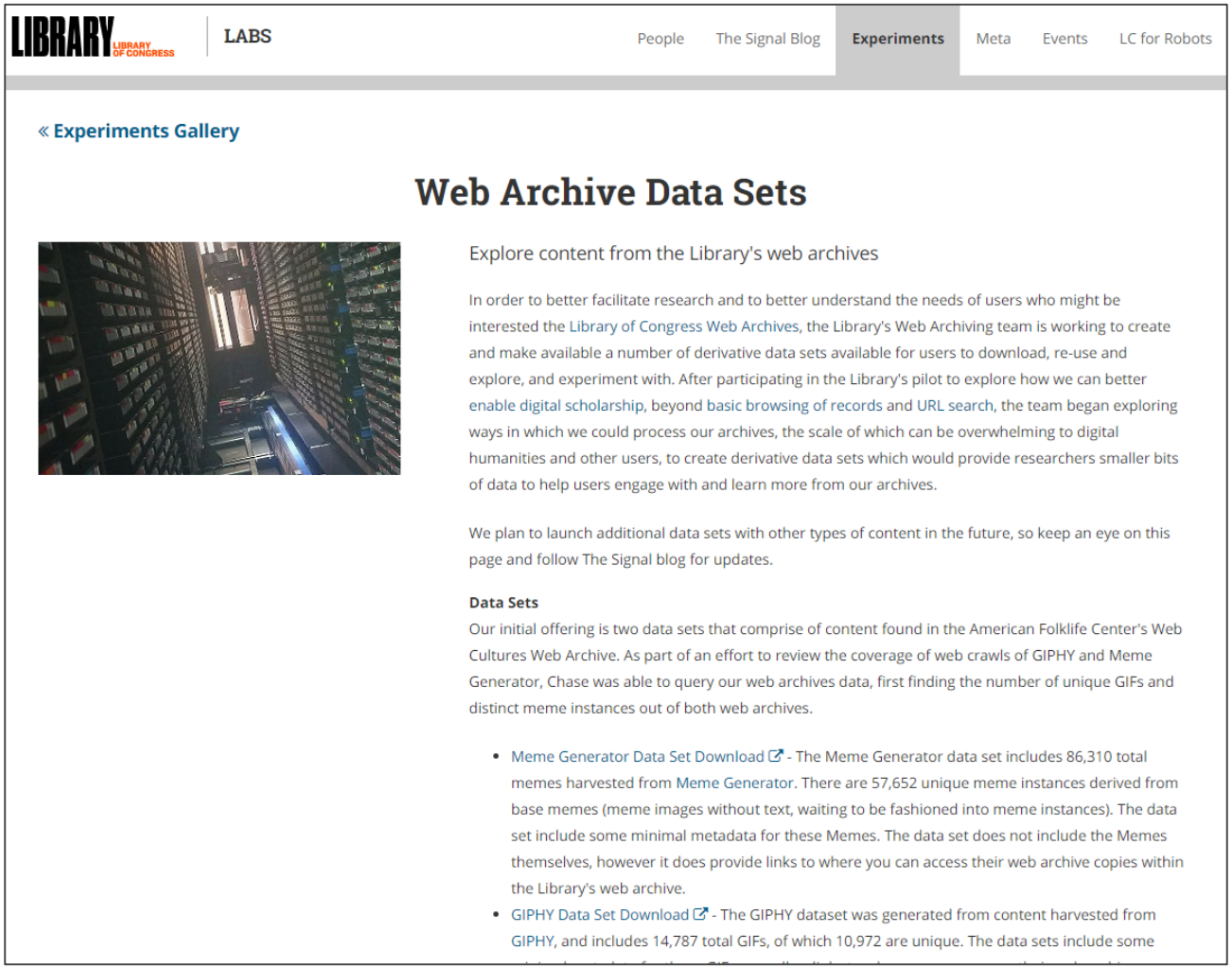
CONTACT US

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

Data sets now available!



https://labs.loc.gov/experiments/webarchive-datasets/

The Cloud

"In the coming months, the Library of Congress will ingest the web archive into the cloud and test new processes for managing the web archive at scale"

-Grace Thomas & Abbie Grotke, *Expansion and exploration in 2018: Processing the Library of Congress Web Archive* (Abstract)

The Cloud - Expectation

Milford Sound, New Zealand - CC BY-ND 2.0 https://www.flickr.com/photos/brianpowersphotography/16290109679

The Cloud - Reality

Milford Sound, New Zealand - CC BY-NC-ND 2.0 https://www.flickr.com/photos/georgeka/10128119644

The Cloud - Leveraging

Pop the 6+ TB of flat CDXs into the cloud storage space

New types of data models from CDX and crawl log data
* Parquet files, Athena, querying in different ways

Run Map-Reduce processes over the CDXs to find number of unique items in the archive (digest) per MIME type (assigned by the server at time of crawl, recorded in the WARCs)

From there, dig deeper into more common, more interesting, more “standalone” items in the archive...

Examples of Digging Deeper

Length	# of PDFs	%
1 page	359	35.90%
2-4 page	325	32.50%
5-9 page	110	11.00%
10-29 pages	113	11.30%
30-99 pages	65	6.50%
100-200 pages	12	1.20%
200-299 pages	4	0.40%
300 pages +	5	0.50%
missing page info	7	0.70%

- **PDFs:**
  - Gain insight into nature/extent of PDFs in the archive
  - Explore trends in use of PDF over time on the web
  - Explore potential use of computational characterization of files
  - Determine potential for developing standard, shareable sets of files for workflow and tool testing

- **"1,000 of X" data sets (starting with .gov)**
  - Generate random samplings, pull embedded metadata, and explore various tools as a means for characterization, for example:
    - **Images:** Explore OCRing text, facial detection, color identification, etc.
    - **Audio:** Explore sound processing tools, Speech to text, music identification, speaker changes, etc.
    - **Video:** Explore both computer vision and sound analysis techniques for characterization.

A note about our flat CDXs:

Variety of standards over time (in international community and LC practice) led to inconsistencies across the flat CDXs

When we start using indexes for back-end analysis, or anything outside of Wayback replay, we have to approximate results

Reindexing to CDXJ-11

- Will help get most accurate data for analysis
- Will lead to creation of zipnum clusters for faster access to public and onsite Waybacks

Future Projects

November 2018: Adding 113th and 114th Congresses

COLLECTION

United States Congressional Web Archive

Collection Period: December 12, 2002 to ongoing. The United States Congressional Web Archive includes member websites from the House of Representatives and Senate, from the end of the 107th Congress through the...

View 976 Items

February 2019: Adding Election 2016 and 2018

COLLECTION

United States Elections Web Archive

Collection Period: August 1, 2000 to ongoing. The United States Elections Web Archive includes sites archived weekly during the election seasons since 2000, documenting sites associated with the Presidential, Congressional, and gubernatorial...

View 5,406 Items

In the coming year:

New index models & functionality

Deeper dives into individual and sets of resources

More data sets at labs.loc.gov

Rolling updates planned for a variety of content and collections – more records, monthly updates to bring content up to embargo date at webarchive.loc.gov

Additional collection descriptions (aka "framework") that will appear at https://www.loc.gov/websites/collections

Thank you!

Grace Thomas
grth@loc.gov
@gracehthom