We are a non-profit Digital Library & Archive founded in 1996
35+PB unique data: 20PB web, 14m text, 3.6m vid, 3.5m aud, 100K soft, etc
Developed: Open source web archiving tools, formats and standards
Engineers, librarians/archivists, program staff
### Archive-It Demo Account

3.6 TB archived since Sep 22, 2010

#### Current Subscription

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Data (this period)</th>
<th>Docs (this period)</th>
<th>Active Seeds</th>
<th>Last Crawl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archive-It Websites and Press</td>
<td>72.4 GB</td>
<td>1,187,978</td>
<td>57</td>
<td>Jan 6, 2016</td>
</tr>
<tr>
<td>2015 United Nations Climate Change</td>
<td>42.6 GB</td>
<td>726,589</td>
<td>5</td>
<td>Dec 10, 2015</td>
</tr>
<tr>
<td>Climate Change</td>
<td>1.9 GB</td>
<td>92,932</td>
<td>13</td>
<td>Dec 27, 2015</td>
</tr>
</tbody>
</table>

628.9 GB Archived

Data Budget Usage
Details on Archive-It

- Used by over 650 organizations around the world from small public libraries to large federal institutions and national libraries to collect, describe, provide access to, and preserve web-published born-digital records

- Harvesting institutional, political, thematic, domain-scale, events, social media, government records, and other web data
Use Cases

Topical Collections
> Often related to traditional, existing similar collecting activities
> Capture spontaneous events
> Document different perspectives and social commentaries

Institutional Collecting
> Support collection of electronic records to meet requirements
> Collect publications/documents that are no longer in print form
> Historical record of an institution or individual’s web presence

Content capture
> Collect specific media types (video) or data for analysis
University of Texas at Austin, Latin American Government Documentation

Use Case:
Archive government documents from 18 countries in Latin America

Content includes:
- Full-text archives of official documents
- Original video and audio recordings of key regional leaders
- Thousands of annual and "state of the nation" reports
- Documentation of elections and political parties
Use Case:
Fulfill a mandate to archive state agency websites and online publications

Content includes:
> Documents in a variety of formats
  
  PDF, text, images, audio/video, social media, and more...
Storage and Preservation

**Storage**

> 2 copies (primary & backup) of archived data are stored at San Francisco area data centers.

> A **third copy** is transferred to the General Archive.

> A copy of archived data can be **shipped on a hard drive**.

> Partners can always **download their archived data** from Internet Archive servers.

**Preservation partnerships**

> 2008: LOCKSS

> 2013: DuraCloud
Long-term preservation of WARC

Long-term preservation of WARC

Collaborative Preservation

- A LOCKSS network for preserving AIT data is a well-established use case; the Canadian Gov Info LOCKSS Network has been in place since 2013.

- IA/AIT and LOCKSS were partners on the WASAPI (Web Archiving Systems APIs) project, developing data transfer APIs for web data preservation.

- A AIT/LOCKSS pilot network is operational and proposed with select Ivy Plus Libraries, with IA and LOCKSS as foundational nodes in the network.

- IA is working with (data) publishers, repos, and services on preservation of open science data and with IA hosting a kickoff nodes in LOCKSS networks for distributed data preservation and delivery to local partners.
LOCKSS/IA Workflow

Crawl
Partner archives web data using Archive-It or uploads web or other data to their Archive-It / IA collections

New partners = new nodes
Additional collections/institutions can be added to this network via “flipping a switch” on the AIT side

IA Storage
WARCs go to multiple IA storage environments (multiple copies in multiple repositories in multiple places) per normal processing

New crawls = new ingests
WARCs are continually added to the IA LOCKSS node as they are collected

IA + LOCKSS copy
Institutions/collections earmarked for LOCKSSification have another copy of their WARCs added to both the IA LOCKSS nodes and are mirrored across other PLN nodes
Goals

● Further tools and joint services for a more networked approach to web archiving and the distributed preservation and access of archived web data

● Provide additional solutions to (currently fraught) digpres service landscape

● Leverage scalable, low-cost infrastructure (and self-owned data centers) of IA, the global network of LOCKSS, and technical expertise of both

● Supporting web archive preservation collaboratively is likely to scale better at lower collective cost than our doing so individually

● Offer community-owned, values-aligned infrastructure, independence from for-profit commercial cloud providers
Why not just put it in the cloud?

- They call them FAANGs for a reason - commercial cloud will suck you dry.

- Community-owned infrastructure supports better mission alignment, governance, technical flexibility, and open access.

- As libraries, we maintain the capacity and the commitment to continue to fulfill our fundamental role as stewards of the information that matters to our communities.

- Local custody of the content means that preservation is resilient to temporary funding gaps and interruptions that could prove catastrophic for content preserved under outsourced, pay-as-you-go models.
Future Directions

- IA + LOCKSS networks for preservation of all/any digital content beyond web archives or publications
- IA + LOCKSS can provide fully-hosted, multi-location, even multi-national preservation networks for institutions that can’t or don’t want to host a node in a distributed network.
- The networks include materials beyond that curated by libraries, including data supporting open science, open scholarship, and open government.
THANK YOU!

Maria Praetzelis
maria@archive.org