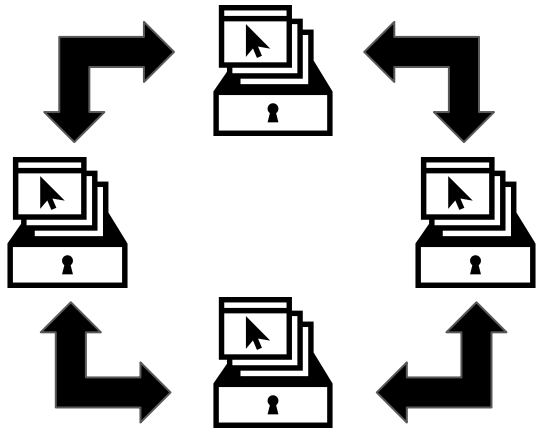
Web Archiving Systems APIs (WASAPI)

for Systems Interoperability and Collaborative Technical Development



Jefferson Bailey (@jefferson_bail), Internet Archive
Nicholas Taylor (@nullhandle), Stanford Libraries
11 December 2017 | CNI Fall Membership Meeting

ARCHIVE-IT

Stanford LIBRARIES

Talk Outline

- Background
- Community Activities
- Outcomes
- What's Next
- Discussion

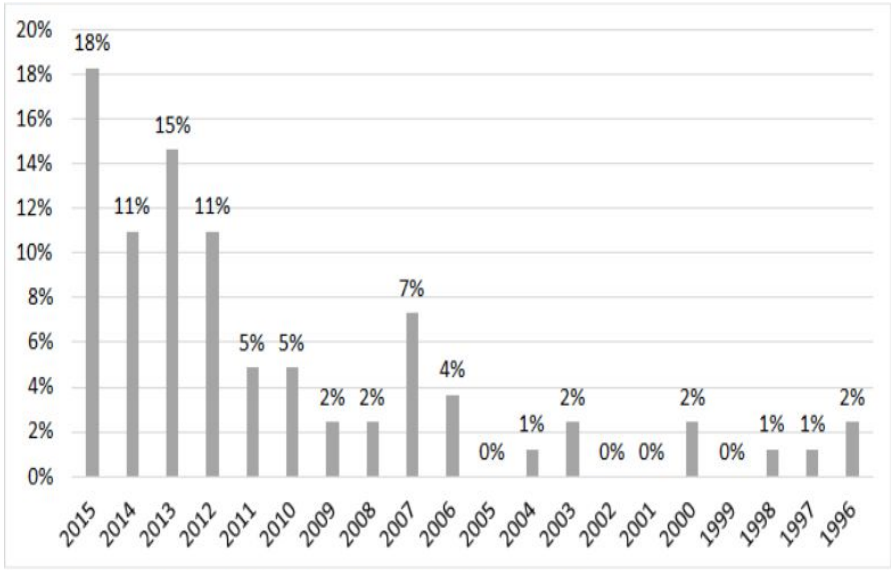
INSTITUTE of Museum and Library SERVICES

RUTGERS

Stanford LIBRARIES

Background

More Orgs Doing Web Archiving (NDSA & AIT)



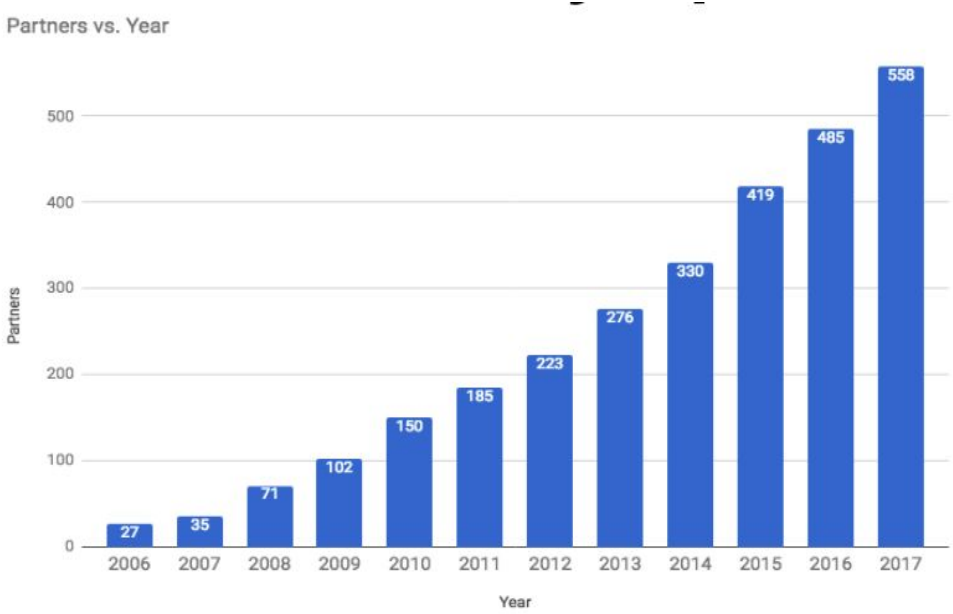


FIGURE 6: YEAR INSTITUTIONS BEGAN ARCHIVING WEB CONTENT

ARCHIVE-IT

Stanford LIBRARIES

Local Web Archive Preservation Still Uncommon

Recent surveys of local preservation

- NDSA: 18 - 20% (2011, 2013, 2016)
- Archive-It: 20% (2016)
- Reasons include:
  - No local preservation plan
  - Trust in service provider
  - Lack of integration
  - Unmanageable data volume

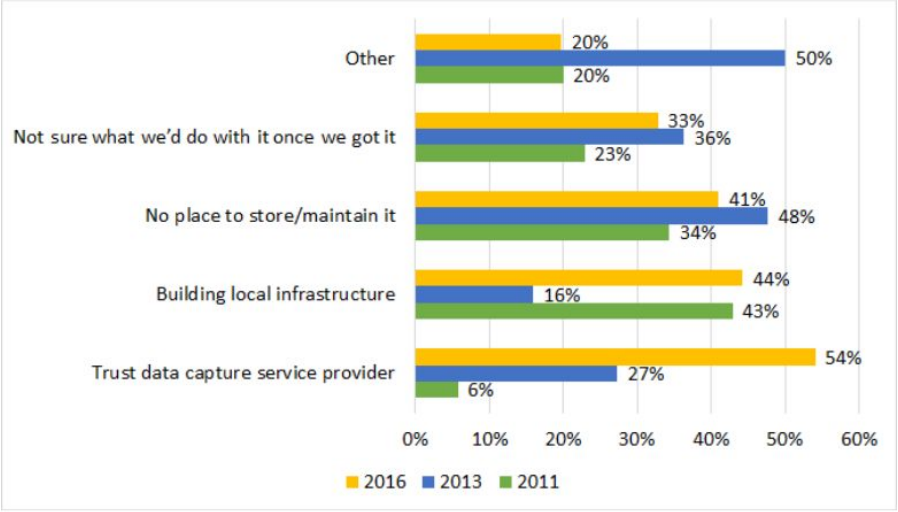


FIGURE 18: REASONS FOR NOT TRANSFERRING DATA FROM AN EXTERNAL SERVICE

ARCHIVE-IT

Stanford LIBRARIES

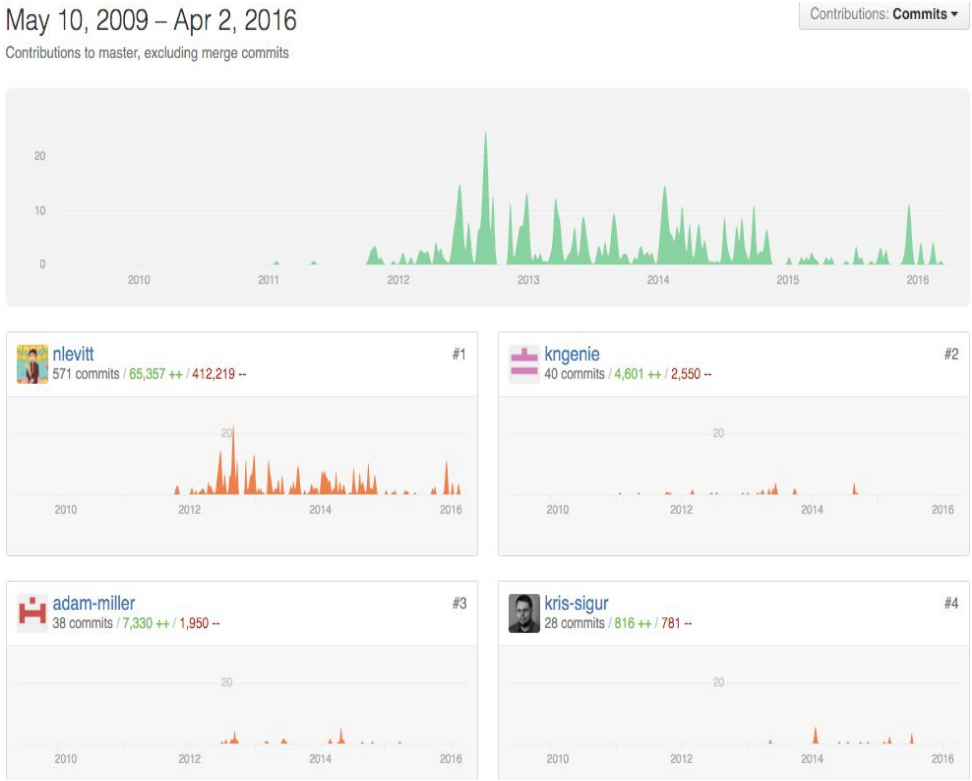
Other Challenges & Motivations

Broader Stewardship

- WA still a niche collecting activity
- Collection use not well measured or aligned with research services

Community Technical Development

- Few coordinated efforts on shared tools
- Marginal distributed dev capacity
- Interoperability not high priority
- Emergence of broader WA CoP
- Nascent community impetus for broad technical development initiatives



ARCHIVE-IT

Stanford LIBRARIES

Grant Project Overview

- "Systems Interoperability and Collaborative Development for Web Archives"
- IMLS National Leadership Grant / National Digital Platform / R&D
- IA/AIT (PI), Stanford, UNT, Rutgers
- 2-year project (Jan 2016 - Dec 2017)
- R&D + National Symposium + APIs

INSTITUTE of Museum and Library SERVICES

RUTGERS

Goals & Outcomes

1. **Build WARC & derivative dataset APIs** (AIT & LOCKSS); test partner data transfer (Stanford, UNT, Rutgers) for better distributed preservation & access
2. **Seed & launch a community** modeled on the characteristics of successful development and w/ participation from communities ID'ed by project
3. **Sketch a blueprint and technical model** for future web archiving APIs informed by project R&D
4. **Seed a technical infrastructure** to facilitate more computational research use of web archives

INSTITUTE of Museum and Library SERVICES

INTERNET ARCHIVE

ARCHIVE-IT

Stanford | LIBRARIES

LOCKSS LOTS OF COPIES KEEP STUFF SAFE

NORTH TEXAS

RUTGERS

ARCHIVE-IT

Stanford LIBRARIES

Community Activities

Work on Research, Education, & Community

- Technical Working Group
- National Symposium
- WARC & Digital Preservation Surveys
- Online Webinars & Demos
- Working with CWAC (Canadians), research tool developers, other harvesting services

INSTITUTE of Museum and Library SERVICES

RUTGERS

Stanford LIBRARIES

Surveys

- WASAPI & AIT “State of the WARC” Preservation Surveys (plus NDSA Web Archiving Survey)
  - 15 - 20% downloading their WARCs for local preservation (33% plan/hope to)
  - Desired API request parameters: institution, collection, crawl, seed, data-range
  - Broad interest in streamlining process, but systems for local preservation remain disparate
  - Transfer of WARCs/datasets for researcher access small but growing

Webinars, Demos, Presos, Working Group

- Video trainings/demos on WA APIs
  - SAA WA Section + SUL demos
- Presentations & Working Groups
  - Archives Unleashed, CNI, IIPC, ReSAW, SAA, TCDL
  - TWG notes

National Symposium on Web Archiving Interoperability

@ Internet Archive, February 2017

- 40+ Institutions from US & Canada
- Orgs included custodial, research, and engineering reps
- Presentations focused on local uses of existing APIs (search, CDX, ASpace, etc) and emerging tools
- Affiliated Archives Unleashed event
- Agenda, docs, presos on GitHub
- Takeaways:
  - Fractured community needs recurring convening forum
  - Facilitate more interaction between practitioners & developers
  - Help for broader institutional understanding & buy-in for WA
  - Marginal distributed capacity necessitates systems interoperability

Outcomes

Research Work

- Preliminary Surveys
- Symposium Summary Report (staff + attendees)
- “Interoperation Among Web Archiving Technologies” white paper (forthcoming)
- Training videos on APIs & WASAPI (AIT + SUL)
- “Community Models for Collaborative Development” white paper (forthcoming)
- Report on iterative development of General Specification & second-level uses (forthcoming)

INSTITUTE of Museum and Library SERVICES

INTERNET ARCHIVE

ARCHIVE-IT

Stanford | LIBRARIES

LOCKSS LOTS OF COPIES KEEP STUFF SAFE

NORTH TEXAS

RUTGERS

ARCHIVE-IT

Stanford LIBRARIES

Technical Work

- General Specification (on Github)
- Archive-It Implementation + docs & videos (on Github/AIT)
- SUL-DLSS downloader + videos (on Github)
- UNT ingest utility (on Github)
- LOCKSS Implementation (on Github)
- Rutgers researcher pipeline (on Github)
- Further testing and utilities (in progress)
- Affiliate APIs (warcprox, WAT APIs)

INSTITUTE of Museum and Library SERVICES

RUTGERS

Archive-It Data Transfer API

Written in python, meets all gen-spec criteria, swagger yaml in the repos

Auth: Uses AIT Django framework (same as web app) -- Auth is not defined in the gen spec

- Browser cookies OR http basic auth (login or pass creds via CLI)

Basic endpoint: https://partner.archive-it.org/wasapi/v1/webdata (in production!)

- Base path returns all WARCs for that account; base/all results are paginated

Query parameters:

- **filename** -- limited use but knowable via AIT CDX/C API
- **filetype** -- currently just WARCs, but others (derivatives) in dev
- **collection** -- ID designating a specific AIT collection [repeatable param]
- **crawl** -- ID designating a specific AIT crawl job
- **crawl-time** -- uses WARC creation date; crawl-time-before / crawl-time-after
- **crawl-start** -- uses crawl job start date; crawl-job-before / crawl-job-after

ARCHIVE-IT

Stanford LIBRARIES

Archive-It Data Transfer API

Sample queries!

Gimme all my WARCs for collection #blacklivesmatter collection (2950)
https://partner.archive-it.org/wasapi/v1/webdata?collection=2950&format=json
Gimme all my WARCs for a specific crawl (300208)
https://partner.archive-it.org/wasapi/v1/webdata?crawl=300208&format=json
Gimme all my WARCs from Q1 of 2017 and collection 1068
https://partner.archive-it.org/wasapi/v1/webdata?collection=1068&crawl-time-after=2016-12-31&crawl-time-before=2017-04-01
WARRRRRRCs:

```
curl --user username:password 'https://partner.archive-it.org/wasapi/v1/webdata?collection=2950&format=json' | jq -r '.files | .[] | .["filename"] | .[]' > WARRRRRRCs.txt
```

ARCHIVE-IT

Stanford LIBRARIES

Archive-It Data Jobs API

GET A JOB!

Supports submitting jobs for generation of derivative datasets re WASAPI goal of expanding researcher / analytic access and use

- Functions;
  - build-wat: build WAT (Web Archive Transformation) files
  - build-wane: build WANE (Web Archive Name Entities) files
  - build-cdx: Build a CDX (Capture Index) files
  - more later!
- Use existing API query syntax to specify content targeted for job
- Receive token for checking job status and use API to poll for status, a la https://partner.archive-it.org/wasapi/v1/jobs/136

ARCHIVE-IT

Stanford LIBRARIES

Archive-It Data Jobs API

GET A JOB! (Done)


```
{
  "account": 1177,
  "function": "build-wat",
  "jobtoken": "136",
  "query": "collection=4783&crawl-time-after=2016-01-01&crawl-time-before=2017-01-01",
  "state": "complete",
  "submit-time": "2017-06-03T22:49:13Z",
  "termination-time": "2017-06-06T01:37:54Z"
}
```

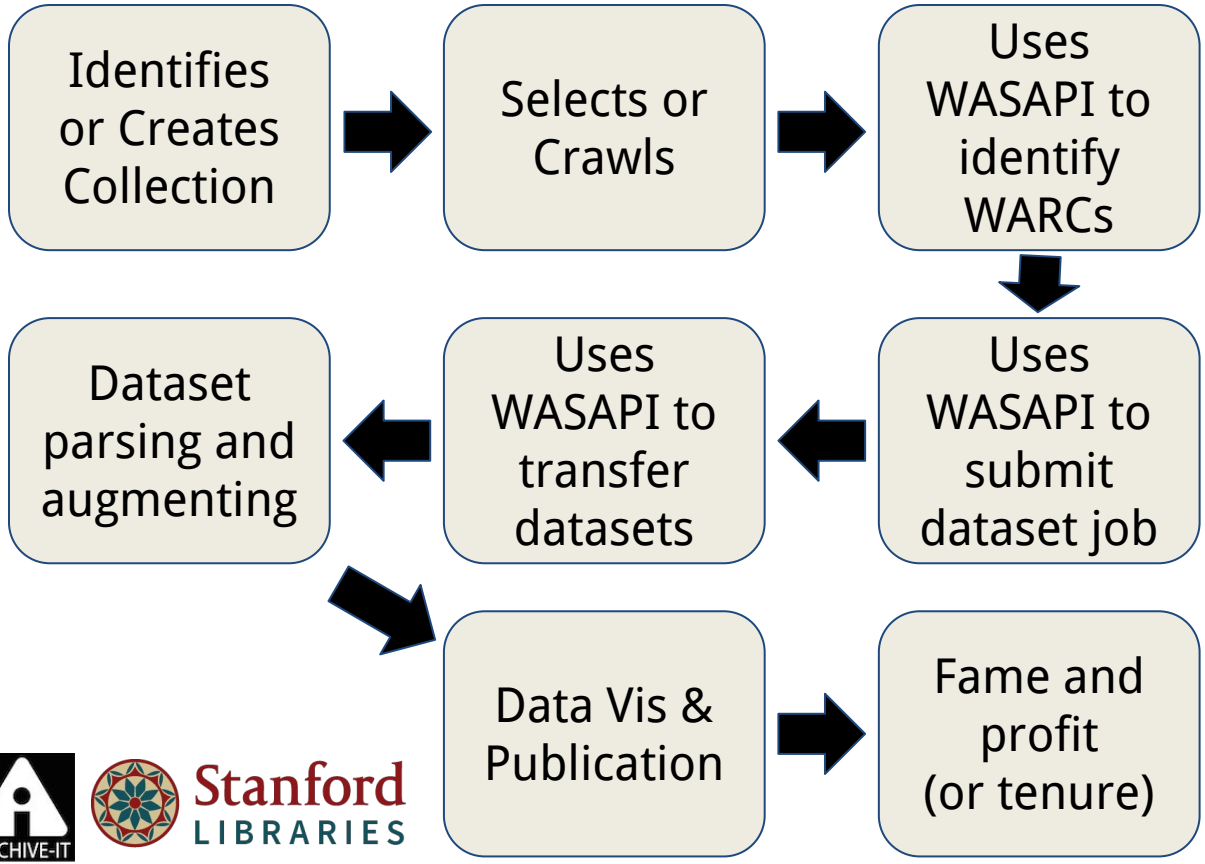

GET A JOB! (Results)

- same as file fields array, with relevant changes to hash, location, size, filetype/name, etc
- query by filetype or job, a la https://partner.archive-it.org/wasapi/v1/jobs/{jobtoken}/result

ARCHIVE-IT

Stanford LIBRARIES

Researcher Workflow



INSTITUTE of Museum and Library SERVICES

INTERNET ARCHIVE

ARCHIVE-IT

Stanford LIBRARIES

LOCKSS LOTS OF COPIES KEEP STUFF SAFE

NORTH TEXAS

RUTGERS

ARCHIVE-IT

Stanford LIBRARIES

Research Services

News Measures Research Project

- 663 local news sites from 100 communities
- 7 crawls for a composite week
- 2.3TB & 17 million URLs captured
- Post-project ongoing monthly crawls
- Access to the collection: https://archive-it.org/collections/7520
- Research datasets publicly available, https://archive.org/details/NMRP_Datasets

INSTITUTE of Museum and Library SERVICES

INTERNET ARCHIVE

ARCHIVE-IT

Stanford | LIBRARIES

LOCKSS LOTS OF COPIES KEEP STUFF SAFE

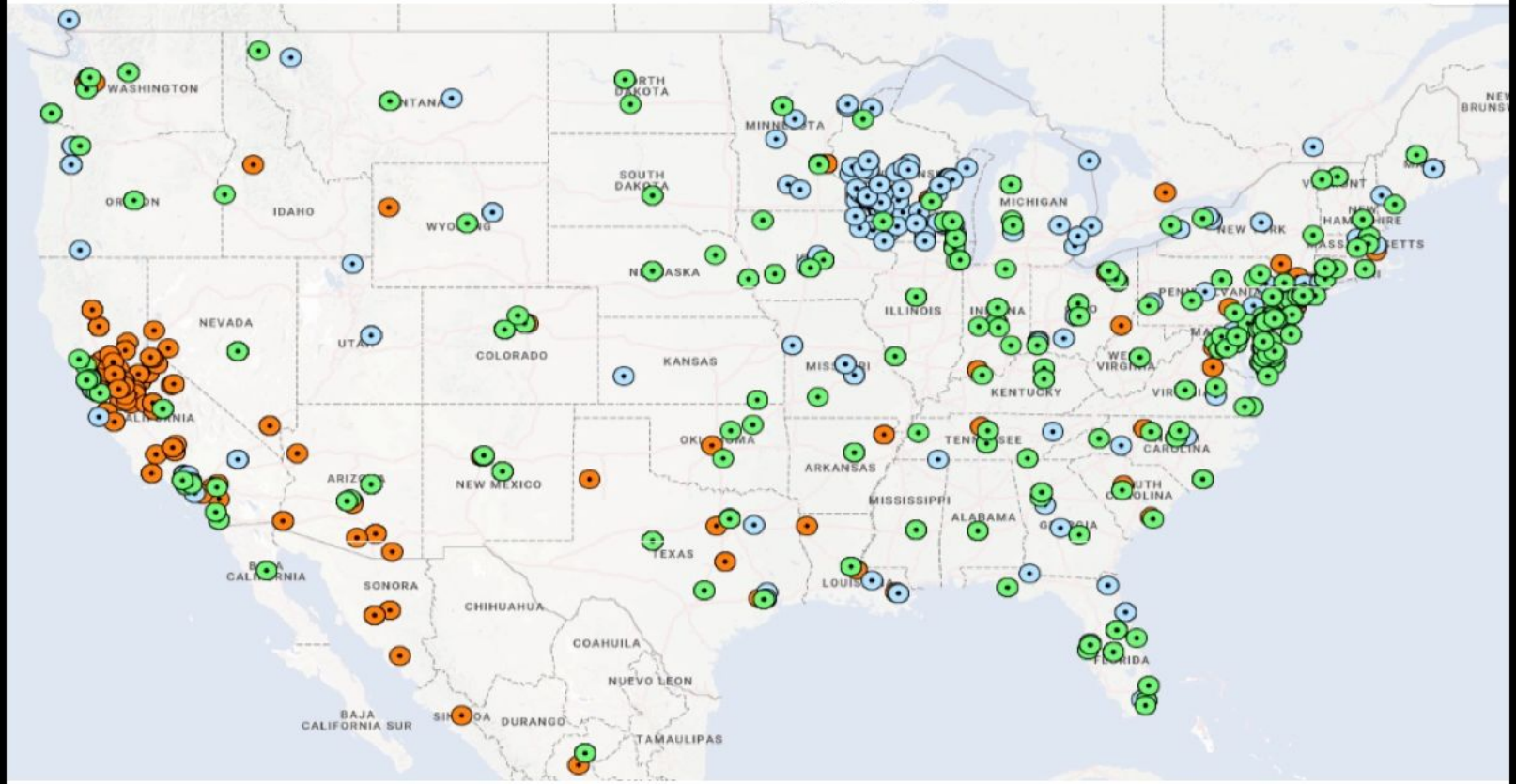
NORTH TEXAS

RUTGERS

ARCHIVE-IT

Stanford LIBRARIES

Location Identification and Mapping



Mapping representation community coverage of local news in Stockton, CA (reddish orange), La Cross, WI (blue) and Newark, DE (greenish yellow). Data were extracted using the Location Identification API, converted to latitute / longitude and mapped using tools available in R.


What’s Next

Ongoing Work

- Expanded production use of APIs
- Continued documentation of recipes and utilities (testers welcome!)
- Ongoing community building & research
- More derivative dataset jobs
- More secondary services
- Integrate other existing APIs
- Identify candidate APIs for WASAPI

INSTITUTE of Museum and Library SERVICES

LOTS OF COPIES KEEP STUFF SAFE

RUTGERS

WASAPI in Action!

- Most AIT partners have transitioned to WASAPI API for local data preservation
- Production datasets job with NMRP
- Second-level preservation service in development by OCUL/COPPUL
- Second-level research service in development by Archives Unleashed
- Integration with other capture tools

THANKS!

WASAPI on the webs

https://github.com/WASAPI-Community
https://archive.org/details/wasapi
https://wasapi.slack.com/ (We can add you)
https://groups.google.com/forum/#!forum/wasapi-community

Jefferson Bailey (@jefferson_bail), Internet Archive
Nicholas Taylor (@nullhandle), Stanford Libraries

Discussion Questions

- Any and all are welcomed. Here are some prompts:
  - What areas or aspects of web archiving do you think can benefit from better technical and social integration?
  - What is your local capacity for partial contribution to technical development?
  - What part of the web archiving lifecycle would most benefit from next-stage API development, post-grant?
  - How might you use the data transfer APIs or utilities?