Systems Interoperability and Collaborative Development for Web Archiving

Filling Gaps in the IMLS National Digital Platform

Mark Phillips, University of North Texas
Courtney Mumma, Internet Archive

Talk outline

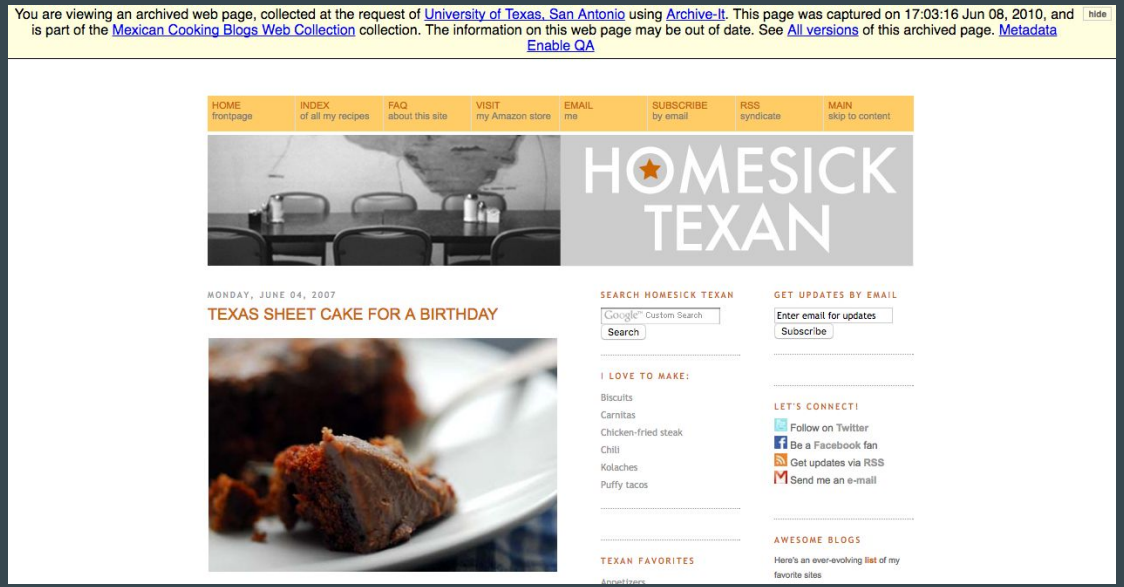
- Growth in Web Archiving
- Problems we face
- How IMLS NLG WASAPI grant will help us improve as a community
- Web archiving APIs and use cases
- What we've done so far and where we're headed
- How you can participate!

INTERNET ARCHIVE

NORTH TEXAS

Web Archiving is just over 20 years old!

- Internet Archive turns 20 this year! Founded in 1996
- International Internet Preservation Consortium (IIPC) was founded in 2003
- UNT CyberCemetery added first site in 1997



UT San Antonio, Texas Sheet Cake for a Birthday, http://homesicktexan.blogspot.com/2007/06/you-say-its-your-birthday.html

Growth in Web Archiving (NDSA & Archive-It)

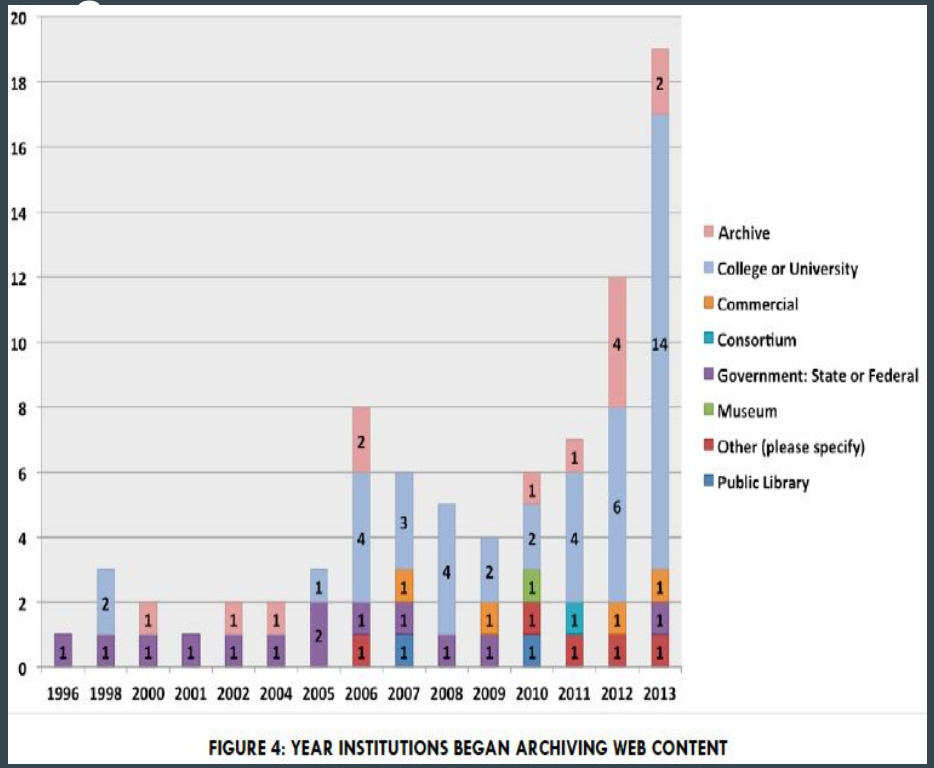
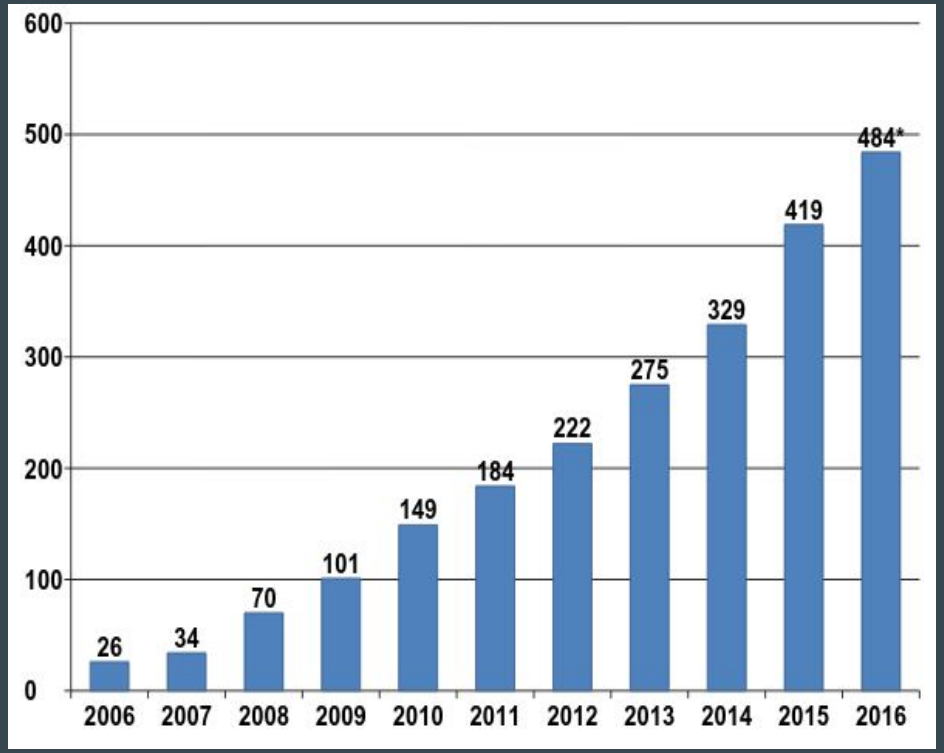


FIGURE 4: YEAR INSTITUTIONS BEGAN ARCHIVING WEB CONTENT



INTERNET ARCHIVE

NORTH TEXAS

Problems we face

Web archiving still a niche collecting activity

Web archives are rarely integrated into Digital Initiatives infrastructure (DL, DA, DR, etc)

Few coordinated efforts on shared tools

Reliance on few providers for entire lifecycle

Convenience of end-to-end services diminishes tech needs

Use largely TBD or not measured

Little familiarity with formats, software, or processes

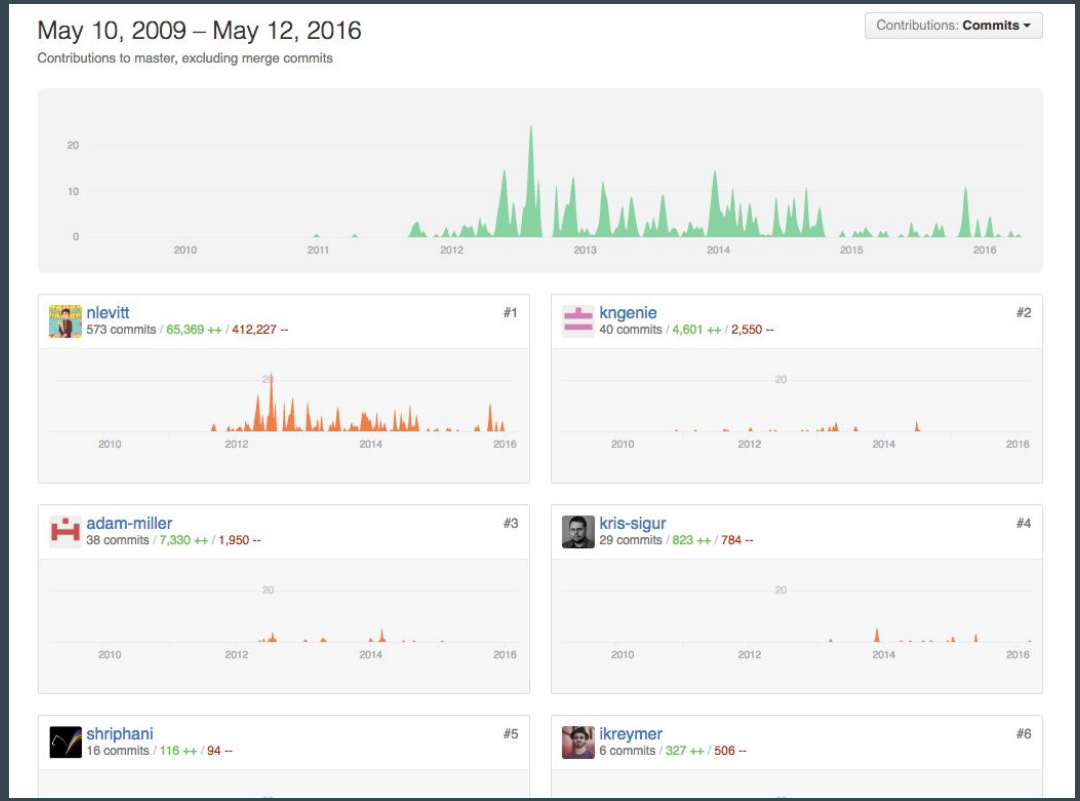
Few on-ramps for non-developer and developer participation

Variance of coordination on emergent efforts & foresight on interoperability

INTERNET ARCHIVE

NORTH TEXAS

Community Involvement in Web Archive Development



INTERNET ARCHIVE

NORTH TEXAS

Local Preservation of Web Archives

Recent Surveys:

NDSA: 18%-20% (2011, 2013, 2016)

AIT: 20% of respondents (2016)

- Data stays with trusted service
- No local preservation plan
- Existing workflows don't consider WARC+ preservation
- Scale is too big and growing

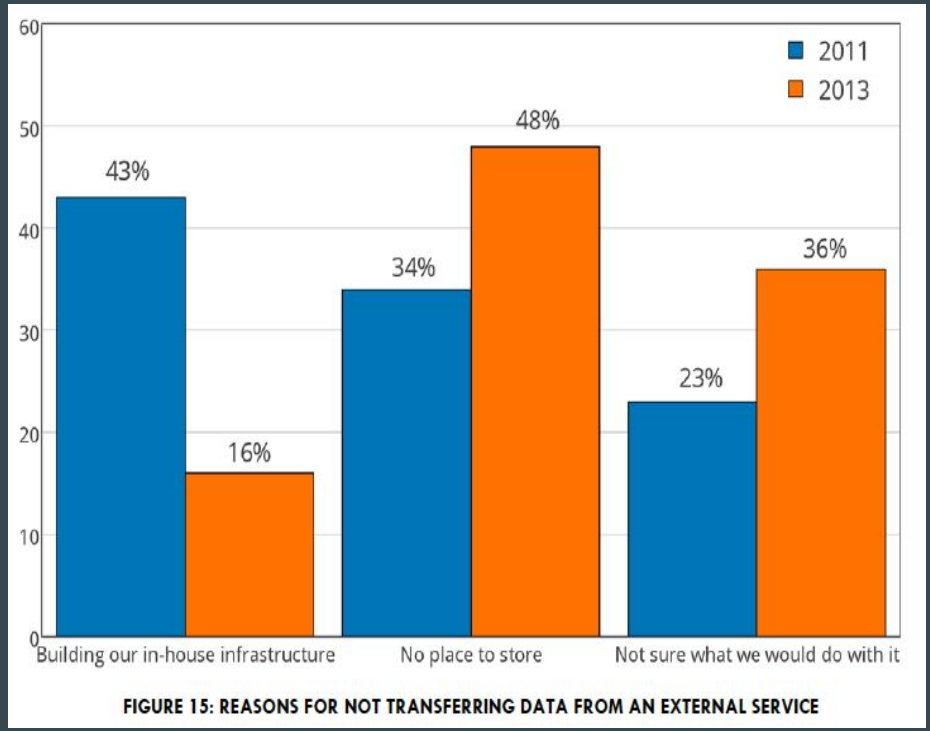


FIGURE 15: REASONS FOR NOT TRANSFERRING DATA FROM AN EXTERNAL SERVICE

Standard “do it yourself” stack of tools

- Heritrix (crawl)
- OpenWayback (replay)
- Nutch/Solr/ElasticSearch (search)

- Spaghetti of scripts to make everything work.

- No current mechanism to integrate with other web archiving services and tools
- Few standardized ways of moving content into preservation environment

Community Involvement in Web Archive Development

Some successful collaborative digital library efforts

Broader web archiving community of practice

There is hope. We can do this!

WASAPI: Web Archiving Systems APIs

- “Systems Interoperability and Collaborative Development for Web Archives”
- National Leadership Grant, National Digital Platform, R&D
- IA/AIT (PI), Stanford, UNT, Rutgers
- 2-year project started January 2016
- National Symposium Early 2017

WASAPI: Web Archiving Systems APIs

Three Key Areas of R&D:

1) What are the attributes of a community model that can support sustainable and broad-based collaborative web archiving technology development?
2) What are the community needs and downstream uses for the planned Export APIs to facilitate transfer of web archive data between distributed systems and what other prospective APIs does it point to?
3) How can better interoperability of web archiving systems support new forms of access and research use?

INSTITUTE of Museum and Library SERVICES

INTERNET ARCHIVE

STANFORD UNIVERSITY

LOCKSS LOTS OF COPIES KEEP STUFF SAFE

RUTGERS THE STATE UNIVERSITY OF NEW JERSEY

NORTH TEXAS

WASAPI: Web Archiving Systems APIs

Outcomes:

1) Seed & launch a community modeled on the characteristics of successful development and participation communities ID'ed by project
2) Build WARC & derivative dataset APIs (AIT & LOCKSS) and test via transfer to partners (SUL, UNT, Rutgers) to enable better distributed preservation and access
3) Sketch a blueprint and technical model for future web archiving APIs informed by R&D
4) Seed a technical infrastructure that will facilitate more computational and distributed research use of web archive collections

INSTITUTE of Museum and Library SERVICES

INTERNET ARCHIVE

STANFORD UNIVERSITY

LOCKSS LOTS OF COPIES KEEP STUFF SAFE

RUTGERS THE STATE UNIVERSITY OF NEW JERSEY

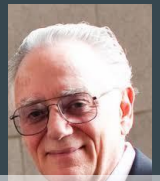
NORTH TEXAS

WASAPI Technical Working Group and Current Progress

Technical Working Group

Stephen Abrams
California Digital Library

Andy Jackson
British Library

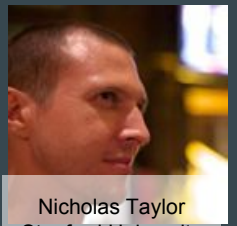


David S.H. Rosenthal
Stanford University

Jefferson Bailey
Internet Archive

Vinay Goel
Internet Archive

Courtney Mumma
Internet Archive



Nicholas Taylor
Stanford University

Tom Cramer
Stanford University

Mark Phillips
University of North Texas

Matt Weber
Rutgers University

INTERNET ARCHIVE

NORTH TEXAS

Why APIs?

(ICYMI: application programming interface)

- Interoperability
- Flexibility and modularity
- Scalability - Bulk data upload and download
- Loose coupling of services (so we can improve pieces as needed)

APIs as Drivers

- Community building and coordinated collection
- Determination of systems of record
- Local ingest and preservation of WARCs
- Assessing archival concepts and practice (ie provenance)
- Broadening options for use (data mining, bulk access, federated browsing)

INTERNET ARCHIVE

NORTH TEXAS

Related API work

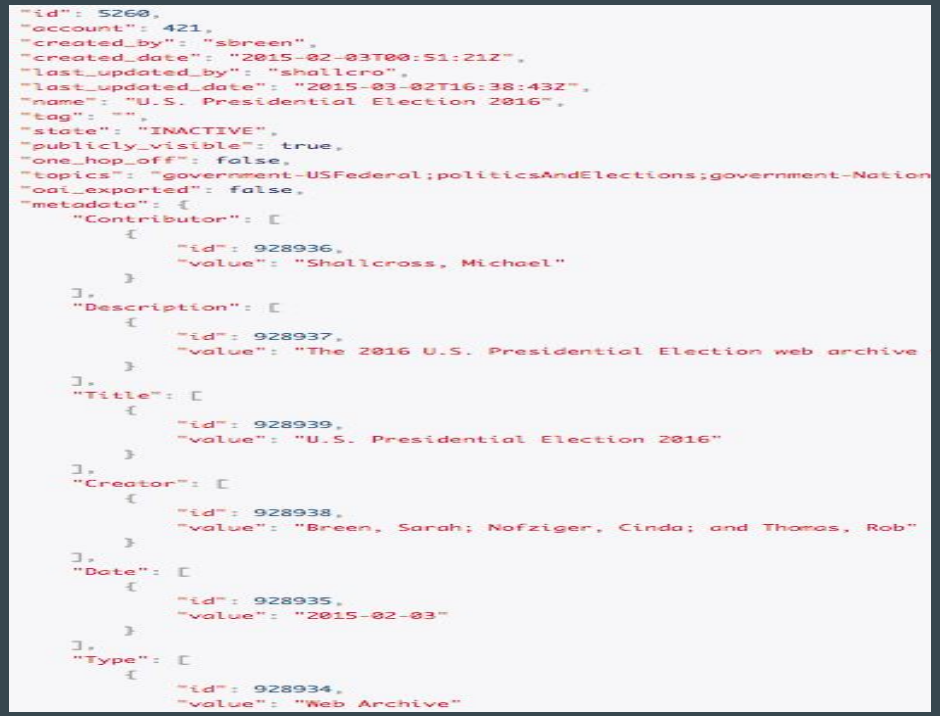
- CDX Server API (IA, IIPC)
- derivative formats (Archive-It, BL)
- crawl logs/partner data (Archive-It)
- Wayback Machine APIs (IA)
- proliferating capture tools (GWU, IA, Rhizome)
- Cobweb (CDL, Harvard, UCLA)

INTERNET ARCHIVE

NORTH TEXAS

Related API work (Internet Archive)

- Wayback APIs
- Archive-It Partner Metadata APIs
- Data Analytics APIs (crawl logs and reports)
- Index (CDX) APIs (AIT and Wayback)
- Upload APIs (non-web)
- Internal APIs



https://github.com/ArchiveLabs/api.archive.org

Use cases

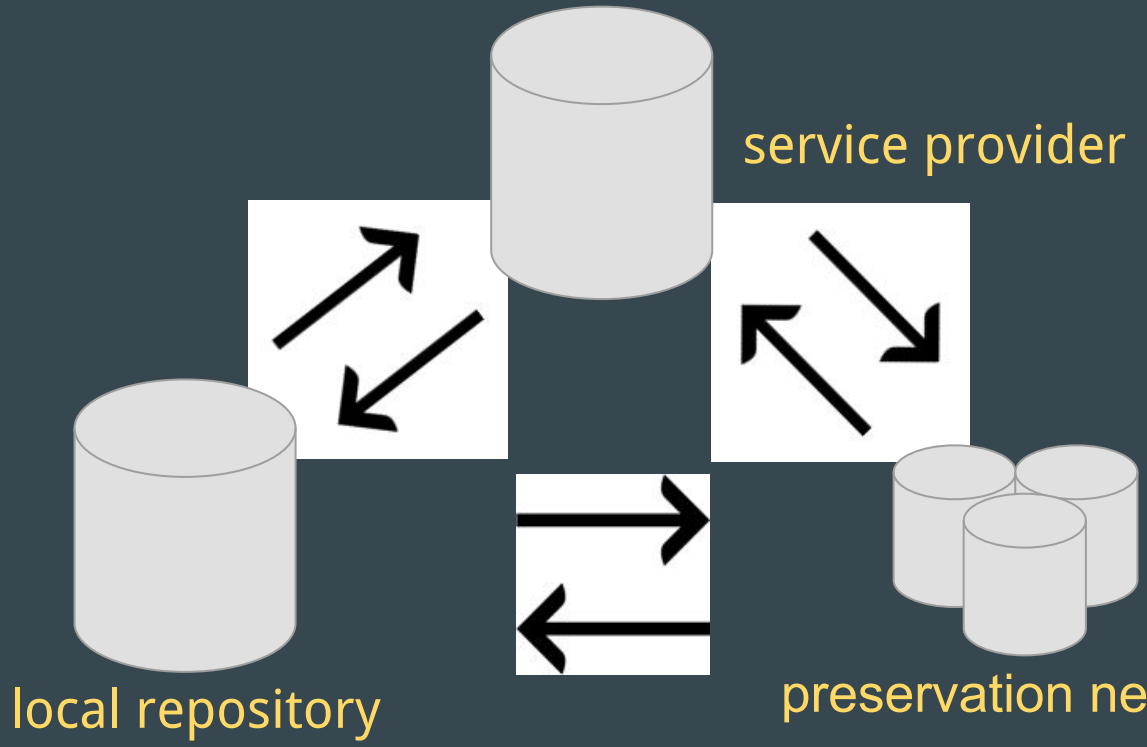
- Archive-It →
  - partner IR/local use
  - DPN
  - LOCKSS (PLN)
- CDL → Archive-It (migration)
- DLSS → IA (WebBase)

- [EoT partners] ← → [EoT partners]
- IA global Wayback→
  - LOCKSS (OA content)
  - national libraries
- LOCKSS (.gov) → IA
- [any web archive] →
  - researcher
  - original publisher

INTERNET ARCHIVE

NORTH TEXAS

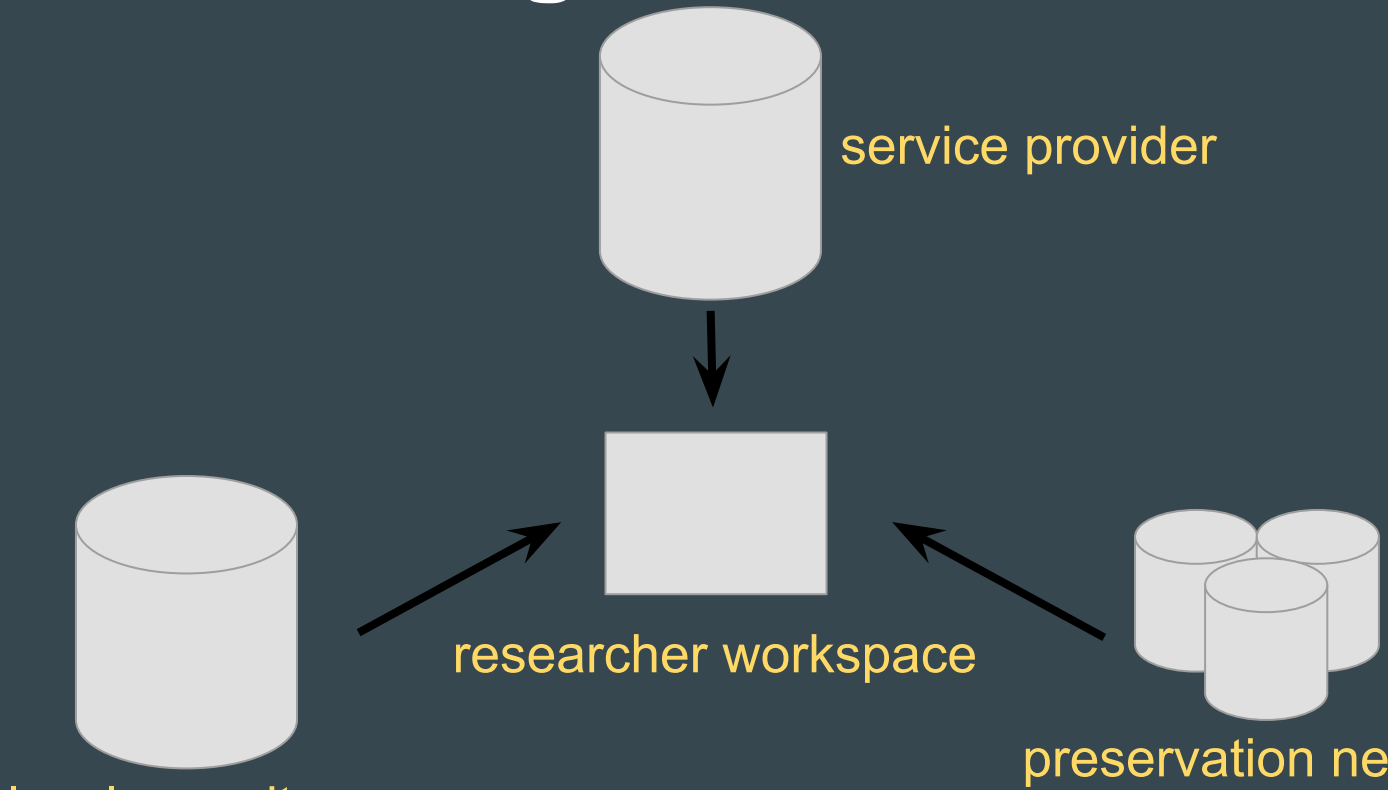
Data exchange between repositories



INTERNET ARCHIVE

NORTH TEXAS

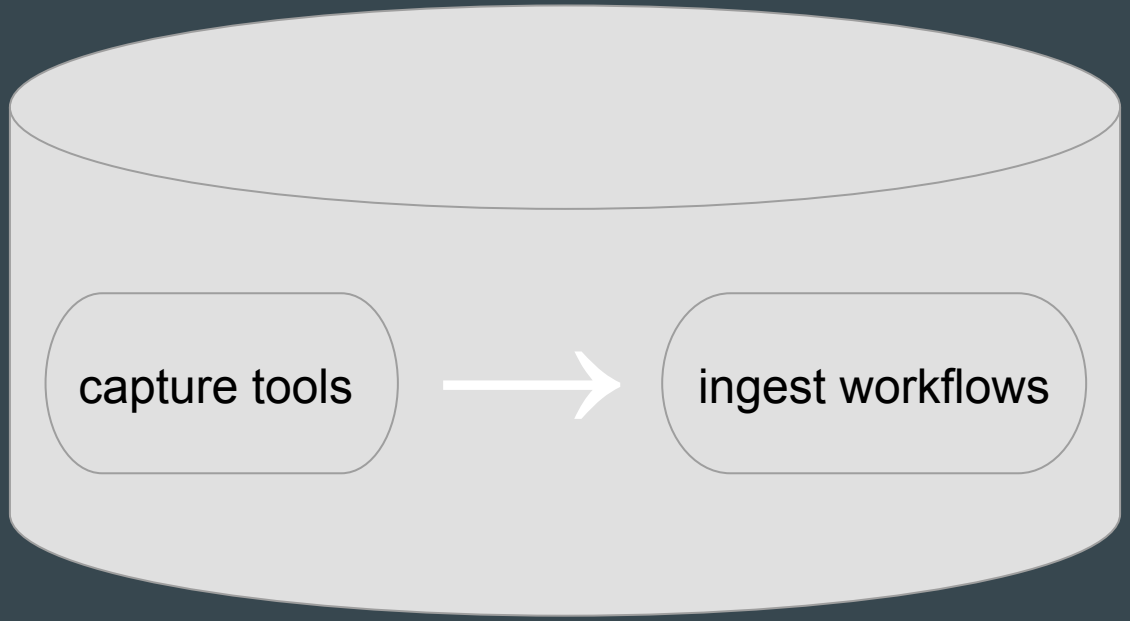
Standardizing researcher data access



INTERNET ARCHIVE

NORTH TEXAS

Data exchange within repositories



Candidate features discussed

- content negotiation for W/ARC or derivatives
- protocol negotiation for transfer handoff
- ability to specify parameters for custom export
- metadata for provenance, crawler configuration, crawl logs, description
- request custom data extraction
- authentication + privileges management

INTERNET ARCHIVE

NORTH TEXAS

How does this contribute to the National Digital Platform?

- Build a foundation for collaborative technology development for web archives
- Improve interoperability of existing and future systems
- Model for identifying and developing APIs
- Engagement of communities of users around the creation, access to, and research us of web archives.

While focused on web archives, many concepts will transfer to other areas

INTERNET ARCHIVE

NORTH TEXAS

Timeline going forward

- API specs
- Test API implementation with LOCKSS and Archive-It
- Rutgers/UNT to build tools to acquire data from test API implementation
- Gather additional use cases
- Build community
- 2017 Summit

INTERNET ARCHIVE

NORTH TEXAS

Your use cases ??? Discussion ?s

- What APIs have attendees built, or are currently using, in their web archiving activities?
- Are these APIs RESTful? If not, why not?
- What frameworks/languages were they built with? What are other notable characteristics of their development and maintenance?
- What part of the web archiving lifecycle would most benefit from next-stage API development, post-WASAPI?
- What has worked and what hasn't worked for your API development (doesn't have to be web archive related)

THANK YOU!!

WASAPI

https://groups.google.com/forum/#!forum/wasapi-community

https://github.com/WASAPI-Community

https://www.imls.gov/sites/default/files/proposal_narritive_lg-71-15-0174_internet_archive.pdf

INTERNET ARCHIVE

NORTH TEXAS