Internet Archive & Archive-It
Digital Preservation Services

Jefferson Bailey, Director
Lori Donovan, Senior Program Manager

May 12, 2020
What & Why

What Digital Preservation Services?

- We provide storage and preservation for our partners’ data already, including multiple copies, locations and fixity checks
- We have had many requests for additional customization and features and provide some of these on an on-demand, contracted basis
- We want to provide these extra features to everyone as a standard service
- For many, existing AIT guarantees are fine, for others, we want to learn more about what they want, need, or think is most important

Why IA Digital Preservation Services?

- The Internet Archive’s efficiencies of scale and mission alignment allow us to provide end-to-end collecting, access, and preservation services
- Our costs are far less than the commercial cloud & others (plus, comm cloud is lame!)
- Services build on significant engineering, technical, and community building capacity
- Discussions with partners around their preservation needs and goals have made clear additional opportunities for collaboration and services
Do you download WARC files or have them shipped to you? (%)
STATE OF THE WARC

If no, why not? (%)

- Trust or prefer for AIT to manage storage: 19
- Currently building own storage infrastructure: 20
- No place to store/maintain the files: 22
- Not sure what to do with the files once we have them: 18
- Other: 20
Background & Context

Internet Archive Digital Preservation
- Non-profit 501c3; mission-aligned, global network of partners
- Self-owned and self-operated data centers in many locations
- Geographic & geopolitical redundancy
- No monetization of ingest/egress, free storage/access for general IA users
- Scale, efficiency, custody, commodities mean low-cost and at-cost operations

Archive-It Standard Storage & Preservation
- Standard with any subscription; forever storage and access
- Minimum 2 copies in multiple places; basic fixity and repair
- ISO-standard formats (WARCs), metadata and search APIs
- Tools and APIs to automate download, mirroring, etc; ingest features
ADDITIONAL DIGITAL PRESERVATION SERVICES

Custom Digital Preservation Services
- Determined / negotiated by contract or individual institution
- Tend to be ad-hoc or on-demand, not generalized
- Include location, fixity, reporting, distribution, replication options
- Some ingest, API, and service integrations

Other Requests
- Include content beyond web archive (WARC) files
- Custom fixity frequencies and reporting
- Additional replication, derivatives,
- Easier IR or local integrations
RELATED WORK

Affiliated and Overlapping Projects

- State of the WARC report
- WASAPI project, API, and downstream tools
- Duracloud, ArchivesSpace, and LOCKSS integrations
- Decentralized Web (dat) pilot with CDL & research data
- Center for Open Science integration mirroring OSF data
  - APTrust, LOCKSS, UND (PresQT), LANL
DIGITAL PRESERVATION SERVICE POSSIBLE FEATURES

- **User-defined Fixity Check Frequency**: Ability for users to specify frequency of fixity checks, including different frequencies for different subsets of archived data.
- **Additional Replication**: Ability for the storage and preservation of a user-defined additional number of copies of archival data.
- **Additional Geolocation**: Ability to have preservation data stored in additional physical locations, including in at least 3 nations on 2 different continents, all IA data centers.
- **Additional Architectures**: Ability to have preservation data stored in multiple technical storage architectures, to reduce technology risk.
- **Inclusion of Non-Web Data**: Ability to upload content into the service beyond just web archive or Archive-It data. The service will be data agnostic.
- **Third-Party Cloud Replication**: Ability to have additional copies of archival data mirrored in third-party cloud systems, commercial or institutional, on demand.
Internet Archive Services Include:

- **Monitoring & Reporting Portal**: An interactive dashboard reporting the status of all archived data in the system, including storage location, fixity jobs (completed and scheduling), resulting manifests, and other data characteristics and reports.

- **API Interoperability**: As with Archive-It, all reports would be available by API and all ingest and export functions would also be possible via API. Basic APIs for syncing to existing popular repository services would also be possible, in line with Archive-It’s current integrations. We expect ingest and export functions to be easy and automated.

- **Cost-Effective**: By owning and operating its own data centers, Internet Archive infrastructure is far lower in cost than the commercial cloud, is a 501(c)(3) non-profit fellow library/archive, can offer cost efficiencies for those utilizing multiple IA services, and can sometimes subsidize costs via coordinate funding.
Q&A/DISCUSSION

Tell us more! Ask us anything!

More updates:
- Archive-It Blog
- Twitter feed

Email us anytime:
- jefferson@archive.org
- lori@archive.org