PROJECT MYCELIUM: NEXT GENERATION DISTRIBUTED DIGITAL PRESERVATION

Nathan Tallman
Penn State University
ntt7@psu.edu
@ntallman@digipres.club

Glenn Heinle
KeeperTechnology
glenn.heinle@keepertech.com

Library of Congress Designing Storage Architectures 2023
Introduction

• MetaArchive started as a project in 2004 and became a distributed digital preservation service provider in 2007.

• Since the beginning, MetaArchive has used Stanford's Lots of Copies Keep Stuff Safe (LOCKSS) to provide bit-level preservation.

• LOCKSS is great for tightly controlled networks and static journal content. But MetaArchive embraces administrative diversity with distributed network management and members deposit a wide variety of cultural heritage materials, in a variety of packaging standards.

• LOCKSS has served us well but is now inhibiting network growth.
Triple Bottom Line Sustainability and UN Sustainable Development Goals

- Triple Bottom Line Sustainability
  - Environmental Sustainability (Planet)
  - Economic Sustainability (Prosperity)
  - Labor Sustainability (People)

- UN Sustainable Development Goals
  - Goal 9: Reduced inequality within and among countries

BUILD RESILIENT INFRASTRUCTURE, PROMOTE INCLUSIVE AND SUSTAINABLE INDUSTRIALIZATION AND FOSTER INNOVATION
Research and Development

- Partnership with Keeper Technology
- Multi-phased Approach
  - 1: Requirements gathering, system design
  - 2: Prototyping, proof-of-concept
  - 3: Deployment of a pilot network
**Discovery**

**Questionnaire**

- Send out the questionnaire to all Metasave members, then schedule interactive sessions with two or more active members. Ideally, we would like to hold one session with a public and one without. The top section, "Questions we want," would be forwarded to Metasave members.
- We can dig into some of the questions in the "Who interacted" section as well as probe items in the "Stuff we need to know" section.

**Questions we want to ask of the Metasave members:**

- What level of granularity is desired for access and sharing? (Examples might help)
- What is the most you have to hand out in a given capacity, and where? (If they don't exist, would you embrace more data into the system?)
- What kind of documentation metadata do you keep track of in your environment?
- What could be done to make this easier?
- How large are the archive objects you want to send to the storage environment?
- How many such objects do you expect to submit?
- As of April 2021, members said they expect Project to ingest its #TB within the next three years (2022-2025).
- How do you define where your data is stored? Which machine(s) or
  - Do you want to have more control over where our data is located?
- How is your data stored? What types of storage are targeted by current staff based on cache usage and geographic distribution? How could these be asked if they had the capability?
- Should the high costs from your organization be utilized for giving access to the system?
- Is there an access denial?
- [Q:] What do you think will be the future of the [name of the project/department]?
- [Q:] What do you think will be the future of the [name of the project/department]?

**Responses:**

- Difficult to Use
- Understaffed
- Automation
- Ease of Ingest
- Physical Shipment

**Wide variety of expertise, infrastructure, workflow**
Architecture

VPN / Secure Network

Site 1
- Storage Service
- Cache Service
- Storage Stack

Site 2
- Storage Service
- Cache Service
- Storage Stack

Site 3
- Storage Service
- Cache Service
- Storage Stack

Site 4
- User Service
- Ingest Stack

Site N
- User Service
- Ingest Stack

Central/Cloud Site
- Administration
- Catalog
- Review
- Admin Stack

Ingest Stack

User Service

Goals:
- Easy
- Automation
- Increase Usage

Distributed Architecture
- Across Member Sites

Modular Design
- Well Defined Interfaces

Scalable
- Secure
- Extensible
Components

**Ingest Stack**
- Primary User Interface
- Local or Cloud
- Archive-Ready Packages and Native Assets

**Storage Stack**
- Data Protection
- Internal Consistency Checks
- External Fixity Verification
- Content Un-Aware

**Admin Stack**
- Limited Access
- Globally Unique Registration
- Additional Metadata
- Policies, Membership, Etc.
THANK YOU

Nathan Tallman  
Penn State University  
ntt7@psu.edu  
@ntallman@digipres.club

Glenn Heinle  
KeeperTechnology  
glenn.heinle@keepertech.com

Library of Congress Designing Storage Architectures 2023