July Data-Facing Call

Date: 2021-07-06 Time: 10am PST / 11am MST / 12pm CST / 1pm EST
Call coordinators: Deb McCaffrey (UMich Med) & Galen Collier (Rutgers Univ) & Amy Koshoffer (UCincinnati)

CaRCC Code of Conduct

Announcements:

- About CaRCC
- Code of Conduct
- Steering committee
    - email df-coordinators@carcc.org if interested
- August plenary on Security

Title: Digital Preservation and Access at Scale: Managing the HathiTrust Digital Library
Presenter: Aaron Elkiss (HathiTrust)
Description:
Founded in 2008, HathiTrust is a not-for-profit collaborative of academic and research libraries preserving 17+ million digitized items comprising over 1PB of digitized images and OCR text. HathiTrust offers reading access to the fullest extent allowable by U.S. copyright law, computational access to the entire corpus for scholarly research, and other emerging services based on the combined collection. This presentation will give an overview of HathiTrust, the data we hold, and the services we provide. It will then describe our storage architecture, our workflows for managing data coming into the repository, our strategies for ensuring long-term preservation of this material, and conclude with some of the unmet challenges of data management that HathiTrust faces.

Biography:
Aaron Elkiss is the Enterprise Architect for HathiTrust. Since 2010, Aaron has had a leading role in a variety of projects and issues around data management and digital preservation for HathiTrust.

Slides: **Digital Preservation and Access at Scale: Managing the HathiTrust Digital Library**

CaRCC YouTube Channel (today's call will be added soon):
https://www.youtube.com/channel/UCMU1PEMM7V4X_KtPDcfB6HA

NOTE: This document is publicly accessible. The link to the Google Drive folder containing this document is posted on the Track Webpage. Please keep that in mind when entering information.

Sign-in (Name / Affiliation / E-mail):

1. Deb McCaffrey / UMich Med School / debmccaf@med.umich.edu
2. Amy Koshoffer / University of Cincinnati / koshofae@ucmail.uc.edu
3. Aaron Elkiss / University of Michigan/HathiTrust / aelkiss@hathitrust.org
4. Jim Leous / Penn State Office of the Associate CIO for Research / leous@psu.edu
5. Rebecca Olson/ University of Cincinnati / rebecca.olson@uc.edu
6. Harrison Dekker / University of Rhode Island / hdekker@uri.edu
7. Plato Smith / University of Florida /plato.smith@ufl.edu
8. Bob Freeman / Harvard Business School / rfreeman@hbs.edu
9. Stephen Tahan / Washington University in St. Louis / tahan@wustl.edu
10. Kirk M. Anne / Rochester Institute of Technology / kirk.m.anne@rit.edu
11. Erik Lundberg / University of Washington / edl@uw.edu
12. Sue Oldenburg / Rutgers / sue.oldenburg@rutgers.edu
13. Mike Hutcheson / Baylor University / mike_hutcheson@baylor.edu
14. Amy Schuler/ Cary Institute of Ecosystem Studies schulera@caryinstitute.org

Max join: 23

Session Notes

- August meeting is a plenary session focused on security
- HathiTrust
    - Define volume as one physical thing, such as one part of a multivolume work or one issue of a journal
- provides full text search on everything! :O
- how to store?
    - two replicated data centers
    - three backup copies
        - 2 on UMich Data Den tape archive
        - 1 on Amazon Glacier Deep Archive
    - separate ingestion process for content and metadata (MARC format)
- Metadata - processed through Zephir Metadata Management System
- only one process has permission to write to the database
- Rights Determination is a major consideration - highlights limit of MARC
    - lots of manual work to track
    - generalization: data set with some restricted data, some open data

Challenges and Unmet needs

Working toward event driven architecture
Improve OCR - newer engines that exist now
Fixing problems now have to go back to digitizer

Metadata challenges
Assumes that this is a representation of a physical thing
Leads to some modern materials not being accepted
Makes changing and updating to modern times extremely difficult

Access is biased towards English

Questions

- ✓ Do you have a specific tool for backing up to Glacier Deep Archive?
  Not really, still working on initial migration
- ✓ During the pandemic, Hathi Trust made many more resources openly available. What were the usage stats with this more open access?
  Using **HathiTrust Emergency Temporary Access Service (ETAS)** - End of 2020 - usage was up by 50%, double
- ✓ How much of the data is usually computed on e.g. through the Data capsule?
  Small amount (1000s of volumes)
  Only computes on the text (~12TB total) and metadata
- ✓ Where is the compute located?
  HPC at IU
- ✓ Does HathiTrust maintain a Globus endpoint?
  No, talk to Research Center - not done yet because it is challenging to manage data for individual volume
- ✓ Latency issues?
  Ingestion process is quite slow, maybe a week from submission to adding to the dataset
  Uses more batch processes
  Keeping Research Center in sync with dataset is difficult