

Kyle R. Rimkus
Librarian for Digital
Programs and
Partnerships
Associate Professor

University of Illinois at
Urbana-Champaign Library
rimkus@illinois.edu
[https://orcid.org/0000-](https://orcid.org/0000-0002-9142-6677)
[0002-9142-6677](https://orcid.org/0000-0002-9142-6677)

File Fixity in the Cloud: Policy, Business, and Technical Considerations



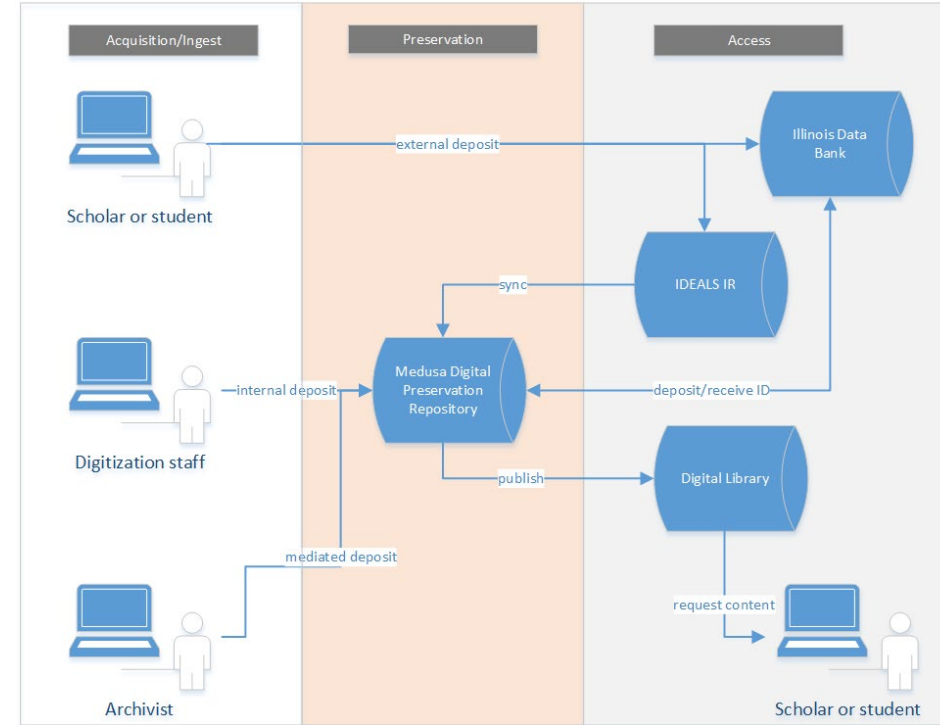
Presented March 24, 2025 for Digital Storage Architectures at
the Library of Congress

Based on a paper written with Genevieve Schmitt for iPres
2024 : <https://ipres2024.pubpub.org/pub/d3akwsah/release/1>

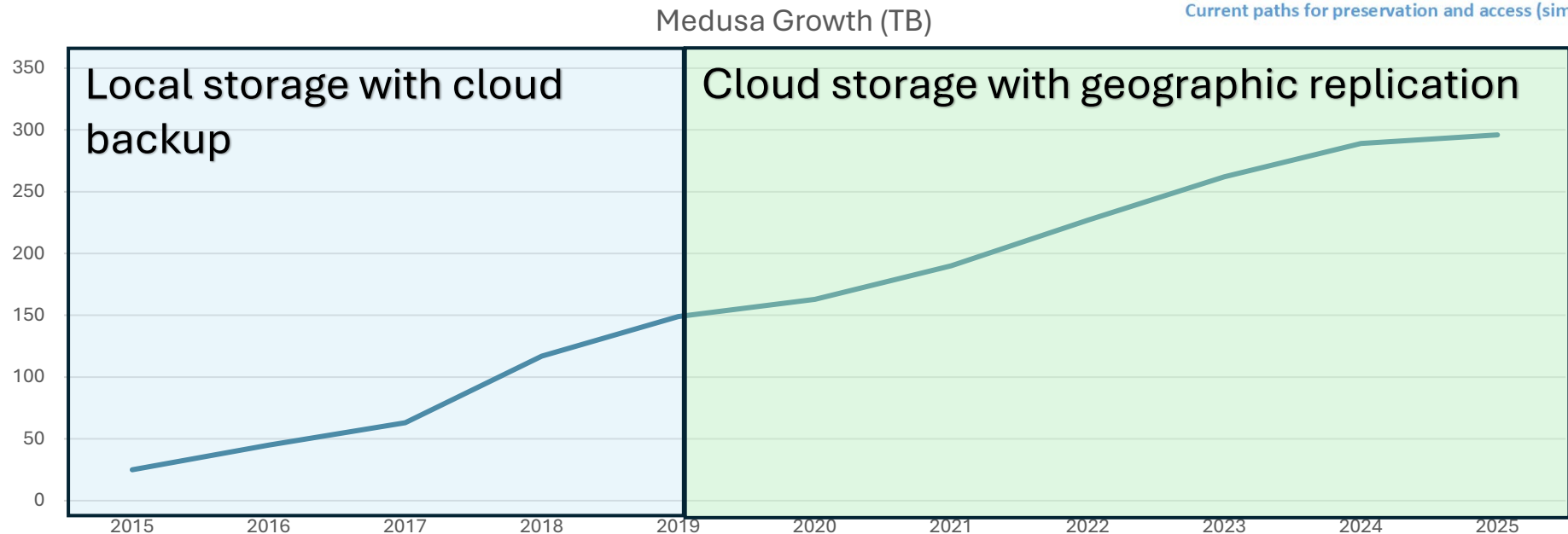
Medusa digital preservation repository

Total size: 300 TB

Total files: 19,000,000



Current paths for preservation and access (simplified view).



The Promise of a “Lift and Shift”



We built Medusa for the file system

1. A file system hierarchy makes sense intuitively and intellectually.
2. The file system is the native environment of electronic records. Storing them in a file system provides the simplest path to accurately representing archival concerns like original order (represented by the folder and file hierarchy) of acquired materials. This applies for files born natively on Windows or Mac operating systems, even when stored on a Linux file system.
3. Most digital preservation tools designed to be run as repository microservices (e.g. FITS or checksum verification) have been written to run on a file system.
4. A file system is portable. If repository managers need to pick things up and move them somewhere else, they can.
5. While important differences exist between operating systems, a file system is largely software-independent, whereas object storage options tend to be vendor-specific with a risk of locking users into proprietary technologies.

Flat Object Storage as an Alternative to the Hierarchical File System

AWS Elastic File System: \$540,000 per year!!!

As an alternative we investigated a more affordable option on offer called the **Simple Storage Service or S3**. S3 is not a file system, but a flat object store of infinitely scalable containers called buckets. In these buckets, bitstreams are identified by unique key values

Technical decisions are financial decisions

- Everything in the cloud is metered.
- We need more research on the economics of digital preservation, and transparent discussion of how much digital preservation costs.
- How being in “the cloud” impacts your technical approach may differ from one cloud vendor to another



UNIVERSITY LIBRARY

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Medusa

Logout (rimkus@illinois.edu)

View lists of:

Institutions

Repositories

Collections

Projects

More

History: Philosophy, and Newspaper Library > Farmer's Wife (St. Paul, Minn.) > Content > 1008/3387 > issues > 1906010101 > 039-FFW-1906-01-01-149-single.tif

File: 039-FFW-1906-01-01-149-single.tif

File information

Storage level: bit-level

MD5: d915981216a14a38d973a0d158ec9545

Last fixity check: 2023-11-26 17:10:27 UTC

Fixity check status: ok

UUID: 5ece0630-35ed-0136-4e01-0050569601ca-e

FITS:

Belongs to: Content

File size: 479 KB

Mimetype: image/tiff

Fixity Check History

Date	Status
2023-11-26	ok
2019-09-18	ok
2019-06-02	ok
2019-03-04	ok
2018-11-05	ok
2018-08-07	ok

Amazon S3 > Buckets > medusa-main > 1008/ > 3387/ > issues/ > 1906010101/ > 039-FFW-1906-01-01-149-single.tif

039-FFW-1906-01-01-149-single.tif

Copy S3 URI

Download

Open

Object actions

Properties

Permissions

Versions

Object overview

Owner
faefaddbb72855c1fa09d54408e0856b48ec87f8f0cbcd4bbdc2b531de3c027

AWS Region
US East (Ohio) us-east-2

Last modified
December 2, 2018, 13:19:14 (UTC-05:00)

Size
479 KB

S3 URI
s3://medusa-main/1008/3387/issues/1906010101/039-FFW-1906-01-01-149-single.tif

Amazon Resource Name (ARN)
arn:aws:s3:::medusa-main/1008/3387/issues/1906010101/039-FFW-1906-01-01-149-single.tif

Entity tag (Etag)
d915981216a14a38d973a0d158ec9545

We store the file path as its S3 key in object storage

Adopting Object Storage and S3 Intelligent Tiering

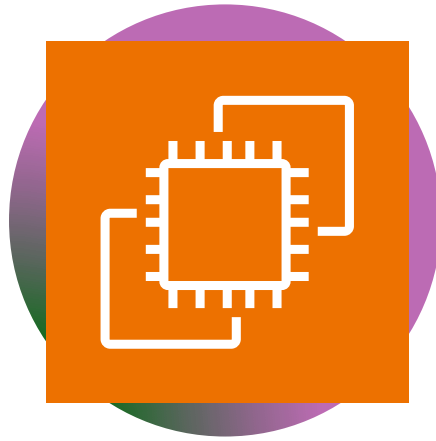
Category	TB	Percentage
Archive Instant Access tier	233.7	91.62%
Infrequent Access tier	14.6	5.76%
Frequent Access tier	6.5	2.55%
Standard Storage	0.2	0.07%

Figuring Out Fixity



Fargate

Reuse Current Ingest
Process in serverless
environment: \$30,000



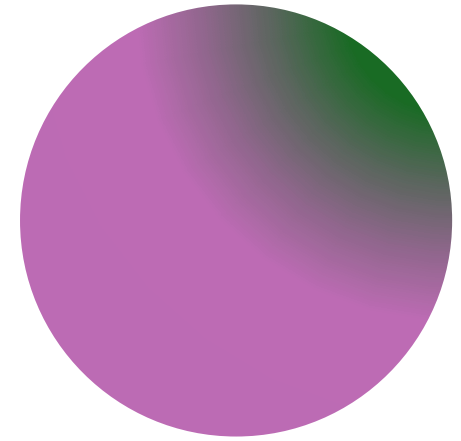
Lambda

AWS-provided
serverless solution:
\$25,000 to complete
processing in a week



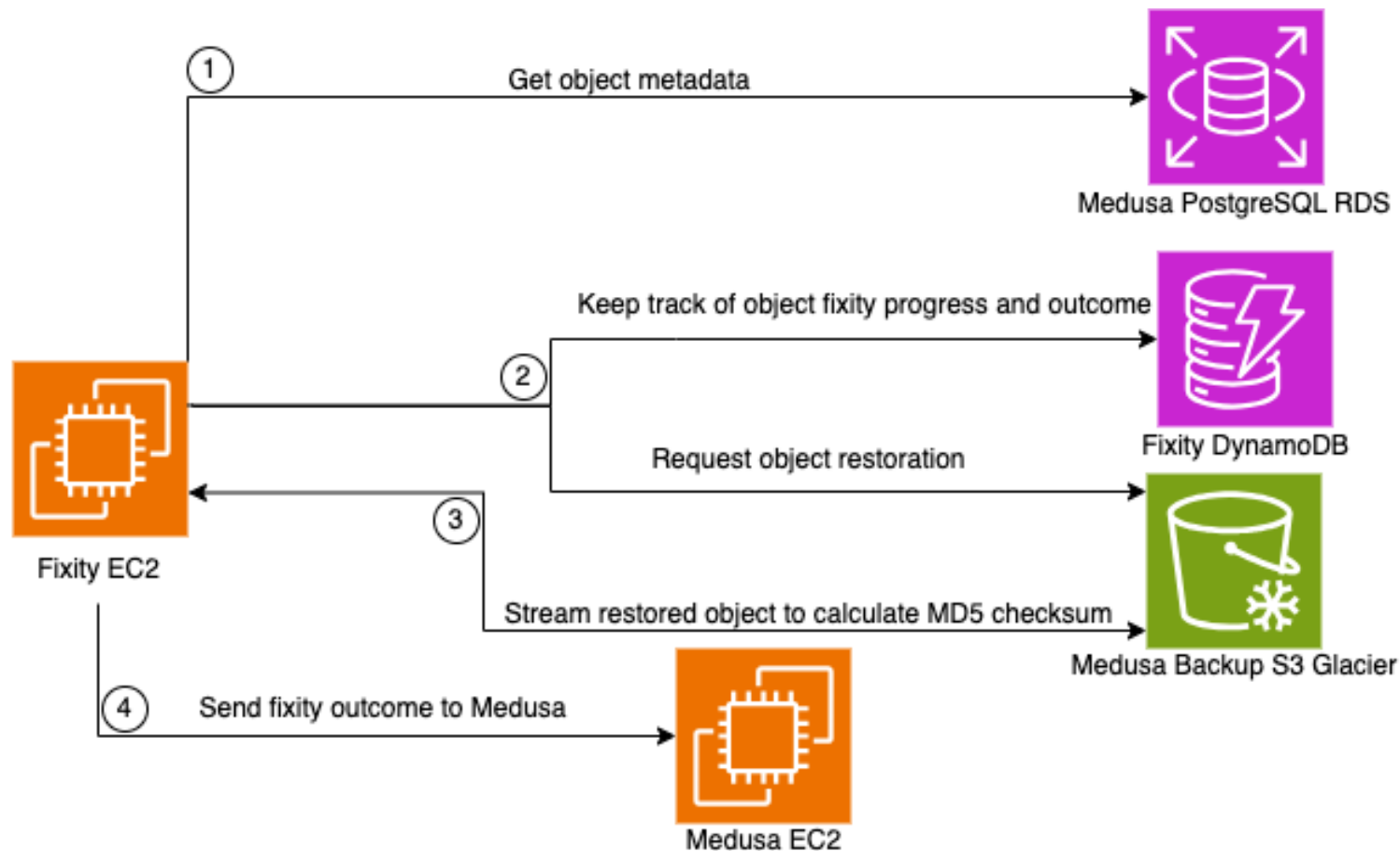
EC2

Custom in-server
application:
estimated \$1700



Fixity Application Simplified

Verifying fixity on 300,000
files per day.



Results

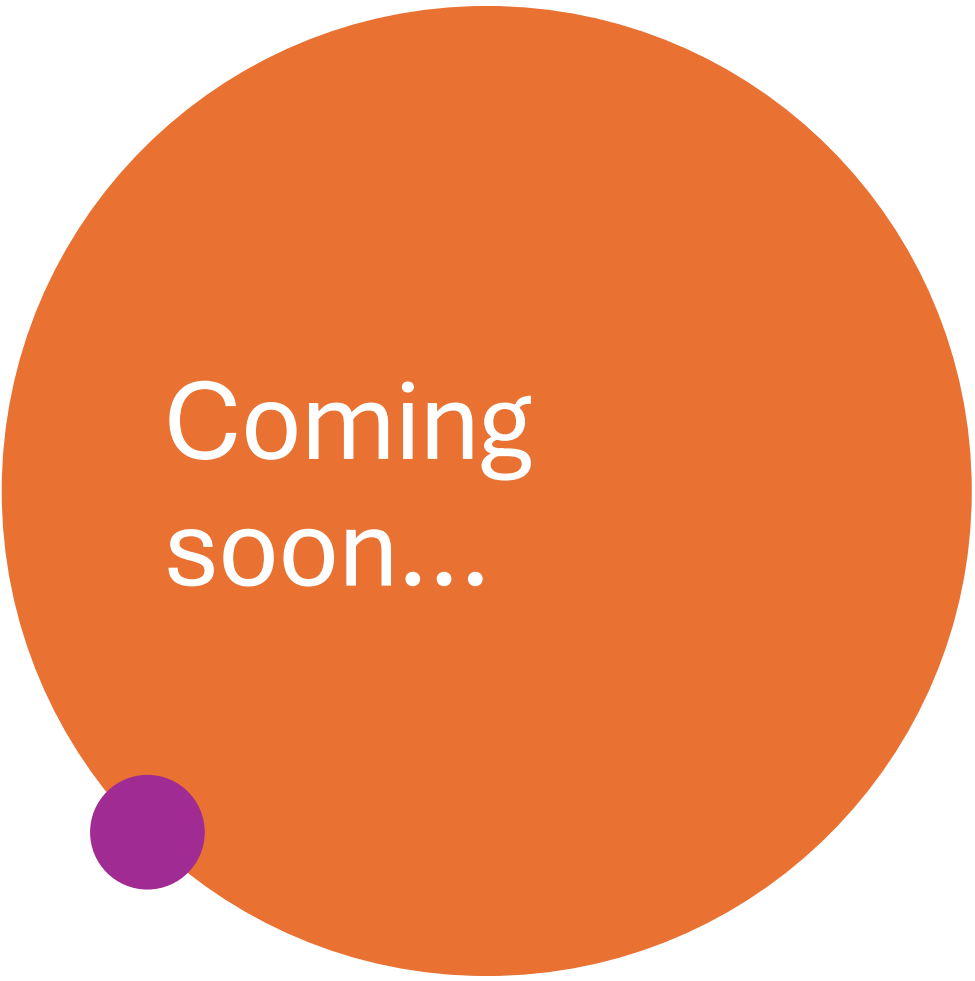
The only files (1600) that failed fixity were due to an internal workflow problem in checksum generation for files that had been replaced on disk. Our analysis verified the reliability of our storage.

“You may set up fixity checking thinking that it’s going to alert you to problems in your hardware, and then you find out it actually alerts you to some problems in your process. And either way, those are things you want to know.”

-- Andrew Diamond, APTrust

Technical Policy

- 1.New content:** Run two fixity checks on all new items in primary storage as defined above within two years of deposit or updating. The purpose: catch flaws in work processes and underlying storage management in a timely manner.
- 2.Stable content:** Run continuous random sampling on all stable content on secondary storage in a financially responsible manner. Trust that the vendor-assured level of file durability will continue, but verify this durability at a modest pace.



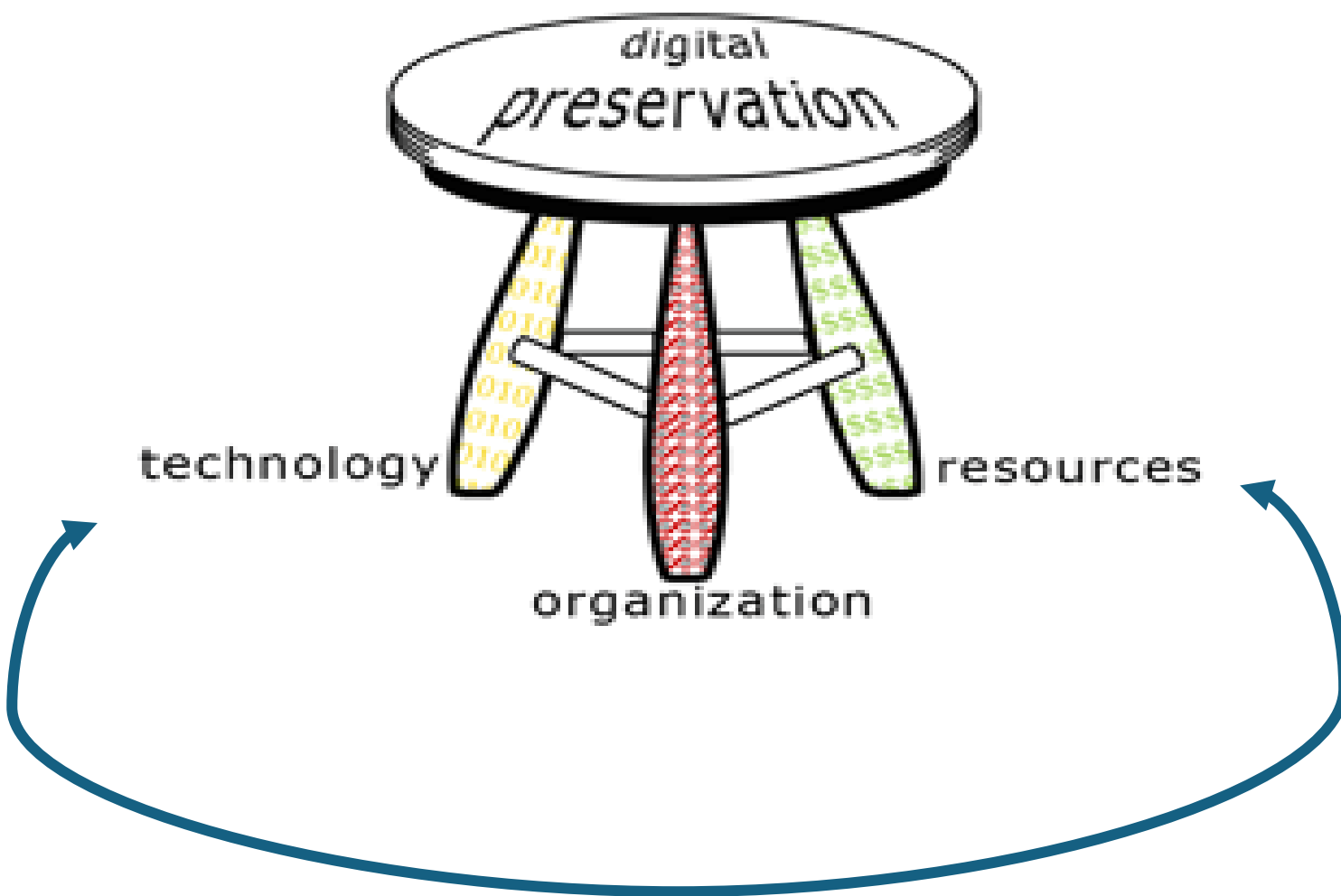
Coming
soon...



AWS S3 now allows for fixity values to be generated and managed in bitstream metadata

- <https://aws.amazon.com/about-aws/whats-new/2024/12/amazon-s3-default-data-integrity-protections/>
- <https://aws.amazon.com/blogs/aws/introducing-default-data-integrity-protections-for-new-objects-in-amazon-s3/>

Analysis: Shifting Costs Between Technology and Resources



AI-generated image of the new breed of software developer / accountant required to flourish in the cloud environment.

Thank you!

Kyle R. Rimkus
Librarian for Digital Programs and
Partnerships
Associate Professor

University of Illinois at Urbana-Champaign
Library

rimkus@illinois.edu

<https://orcid.org/0000-0002-9142-6677>