

# DuraCloud Pilot Program

Experiences, Use Cases, and  
Lessons Learned

# DuraCloud Team

- Michele Kimpton
- Carissa Smith
- Andrew Woods
- Bill Branan

# Presenters

## New York Public Library

Barbara Taranto

## Biodiversity Heritage Library

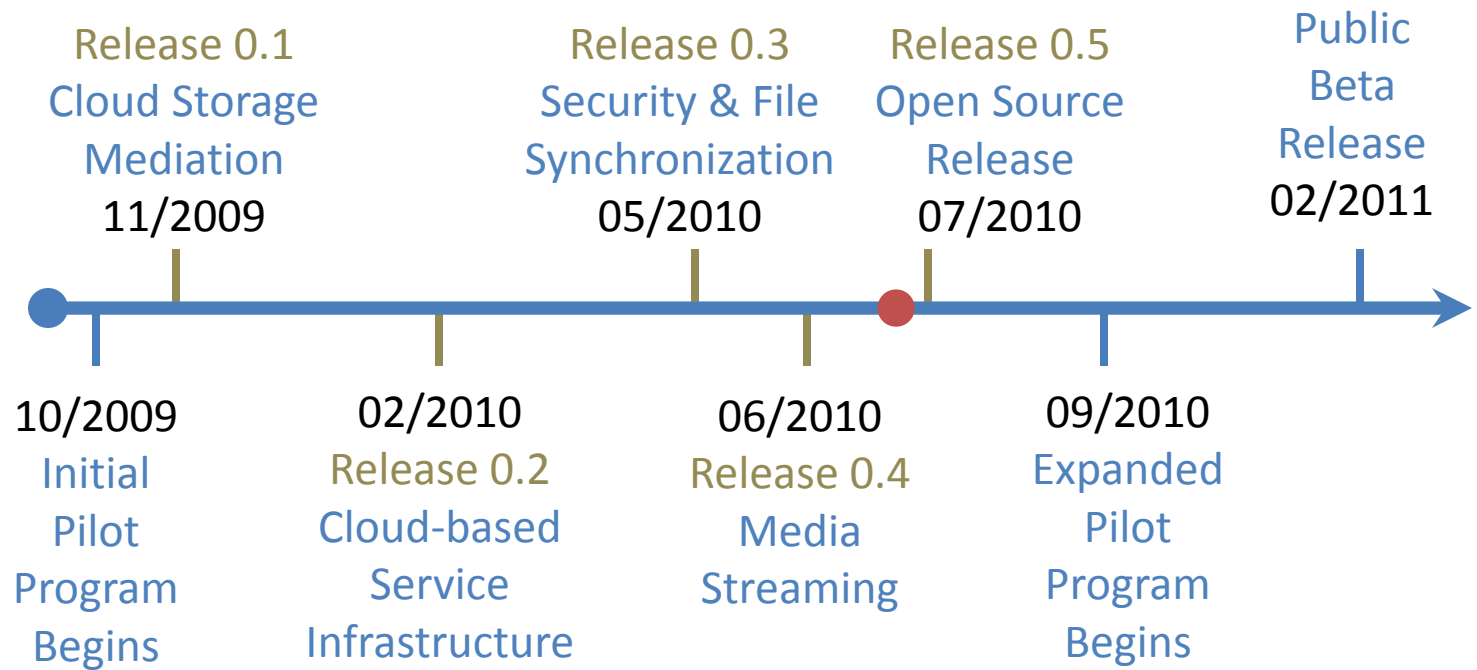
Chris Freeland - Missouri Botanical Garden

Tom Garnett - Smithsonian

## WGBH Educational Foundation

Peter Pinch

# Pilot Timeline



# Reasons for the Pilot Programs

- DuraSpace
  - Real use cases
  - Real data at scale
  - Real users testing the software
  - Help in discovery of opportunities and obstacles
  - Opportunity to engage with potential customers
- Pilot Partners
  - Gain a better understanding of the capabilities and limitations of the cloud
  - Help to shape the DuraCloud offering into something truly useful
  - Discover how to meet real business needs

# **DuraCloud Pilot Program: New York Public Library Pilot Partner Report**

Barbara Taranto  
Managing Director, NYPL Labs  
Office of Strategic Planning  
NDIIP 2010 Partners Meeting  
July 21, 2010

# New York Public Library

- NYPL Repository
- What Prompted Our initial interest in DuraCloud
- NYPL Pilot program – 2009 Use Cases
- Outcome
- Where we are now – 2010 Use Cases

# New York Public Library

- Real data at scale
- 60 TB of mostly image files
- 40 TB of multimedia files waiting to be loaded
- In final stages of implementing Fedora repository
- Migrating from SAN to Isilon storage cluster
- New workflows in development
- Digital Gallery – Primary front end application (700,000 metadata records)



# What Prompted Our initial interest in DuraCloud

- Possibility of buying licensing, services that we didn't need to develop, host, support or upgrade ourselves
- Belonging to a larger community of libraries, museums and cultural organizations working in concert on a common problem
- Need to address some serious issues with the creation, support and web delivery of large zoom-able files
- Streamline workflow
- Tie delivery directly to repository workflow

# NYPL DuraCloud Pilot Goals

- Preservation

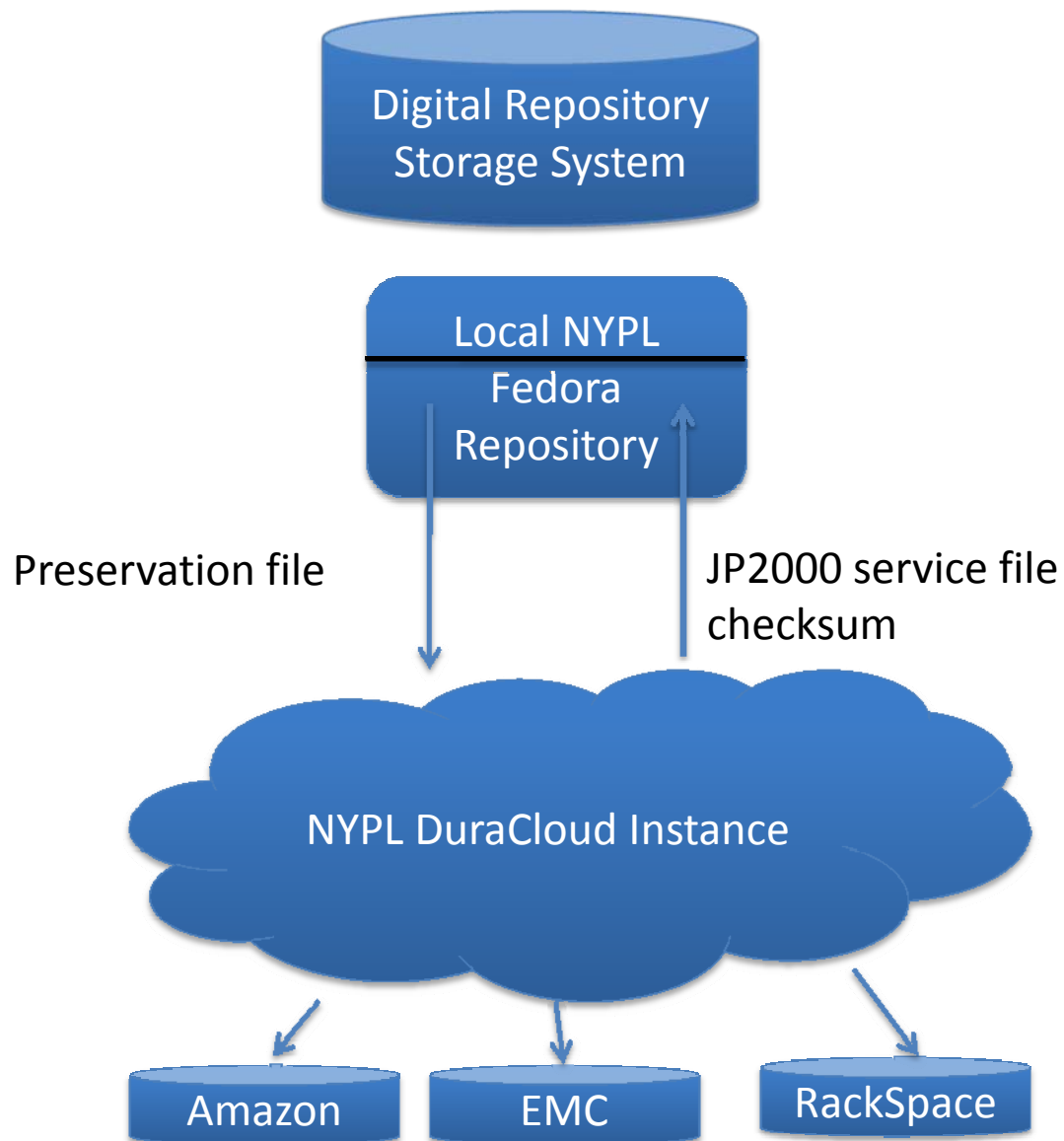
- Migration of service files from an unsupported format to a supported format – Mr. Sid to JP2000
- Data integrity checking of new format
- Ingest of new data streams associated with existing objects in Fedora Repository

## Access

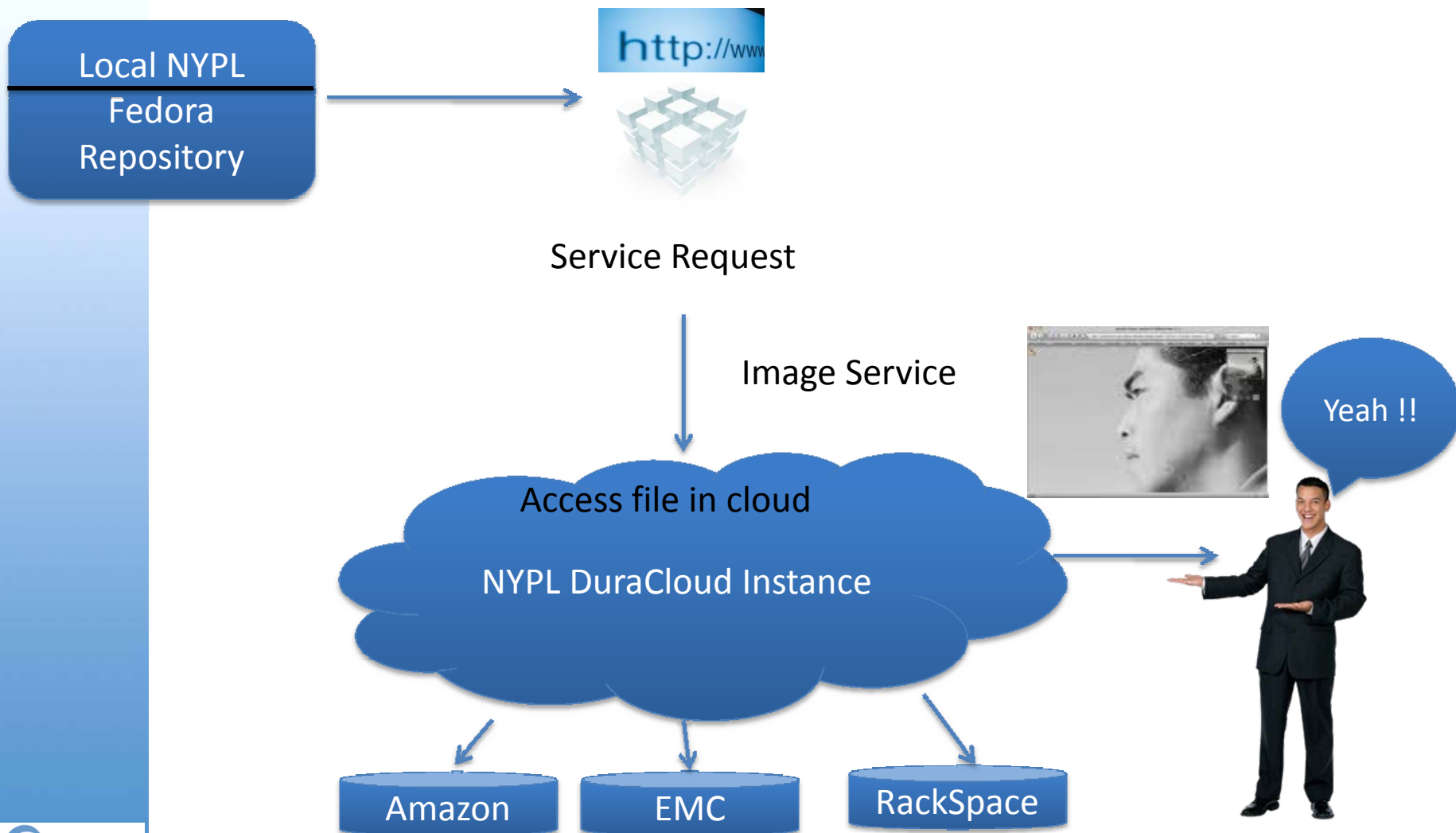
- Reduce number of service files
- Reduce number of services used in delivering service files
- Provide reliable and dependable service and access to those service files

# Key Advantages Cloud provides NYPL

<b>Most Impactful Advantages Electronic Survey</b>	<b>Responses</b>
Scalability	79
Remote, Off Campus Storage of Digital Assets	64
Ease of Implementation	54
Flexibility	53
Don't Have to Staff Locally	39
Cost	33
Elasticity	26
Pay for Use	14
Other	5



## NYPL Access Services utilizing DuraCloud



# Outcomes

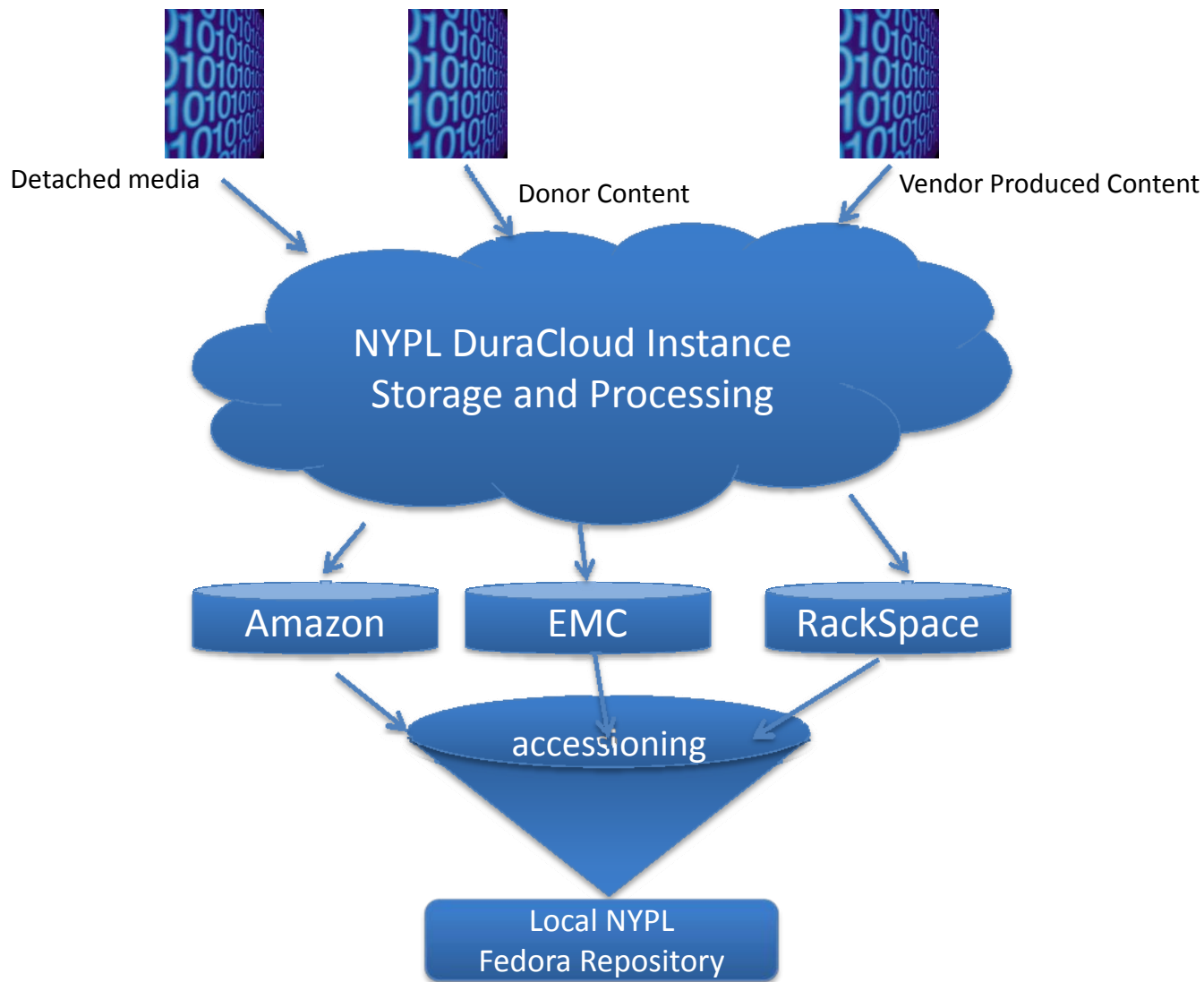
- Loaded 10 TBs of TIFFS to cloud
- Were able to view, convert, and serve a subset of those files via Adore Djatoka image service
- We were able to download and verify end products of conversion process with the DuraCloud API and JHOVE
- We were able to demo the process of chunking and storing multi-gigabyte media files with the DuraCloud sync tool
- Delays in hardware did not allow for provisioning additional services from DuraCloud

# Lessons Learned

- Constraints are at many levels
  - Policy favors public service over preservation needs
  - IT does not support dedicated, segregated bandwidth
  - Restricted funds (restrictions on use of capital funds)
  - Flexible storage is needed at many points in the NYPL workflow – loading content up to the cloud requires local storage as well
- Quickly reached the limitations of single thread processing. Image conversion.
- Local processing is quicker because there is no latency due to the I/O with moving files.
- Flexibility, scalability, Elasticity of the cloud are important to NYPL at the beginning of the process

# Use Case – 2010

## Collaborative Evaluation and Processing Space







# Questions?



# BHL in the cloud



**A Pilot Project**

**Tom Garnett, BHL Executive Director**  
**Chris Freeland, BHL Technical Director**

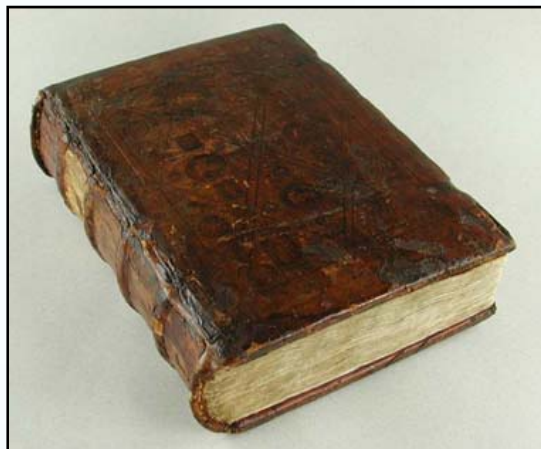
<http://www.biodiversitylibrary.org>

The Biodiversity Heritage Library (BHL) is a global community of natural history libraries and research institutions who have formed a partnership to digitize and make available the world's biodiversity literature.



**BHL Partners**

<http://www.biodiversitylibrary.org>



../  
[albumofabyssinia00fuer.divu](#)  
[albumofabyssinia00fuer.gif](#)  
[albumofabyssinia00fuer.pdf](#)  
[albumofabyssinia00fuer\\_abbyy.gz](#)  
[albumofabyssinia00fuer\\_bw.pdf](#)  
[albumofabyssinia00fuer\\_dc.xml](#)  
[albumofabyssinia00fuer\\_divu.txt](#)  
[albumofabyssinia00fuer\\_divu.xml](#)  
[albumofabyssinia00fuer\\_files.xml](#)  
[albumofabyssinia00fuer\\_flippy.zip](#)  
[albumofabyssinia00fuer\\_jp2.zip](#)  
[albumofabyssinia00fuer\\_marc.xml](#)  
[albumofabyssinia00fuer\\_meta.mrc](#)  
[albumofabyssinia00fuer\\_meta.xml](#)  
[albumofabyssinia00fuer\\_metasource.xml](#)  
[albumofabyssinia00fuer\\_names.xml](#)  
[albumofabyssinia00fuer\\_names.xml\\_meta.txt](#)  
[albumofabyssinia00fuer\\_raw\\_jp2.zip](#)  
[scandata.zip](#)

Biodiversity Heritage Library

Feedback | About | Tools | Tutorials | BHL Members | Copyright | Contact

Search  All Categories

Advanced Search

Browse By: [Titles](#) | [Authors](#) | [Subjects](#) | [Names](#) | [Map](#) | [Year](#) Published In: (Any Language) For: (All Contributors)

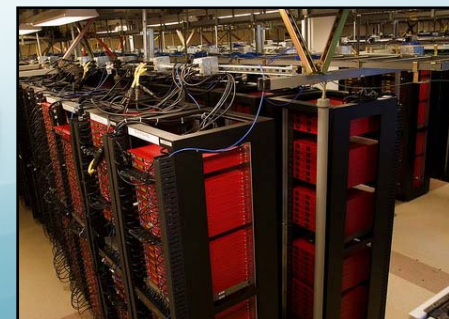
Pages  
 Page 10  
 Page 11  
 Page 12  
 Page 13  
 Page 14  
 Page 15  
 Page 16  
 Page 17  
 Page 18  
 Page 19  
 Page 20  
 Page 21  
 Page 22  
 Page 23  
 Page 24  
 Page 25  
 Page 26  
 Page 27  
[Link to this page](#) [View Text](#)

Names on this page  
[Corythornis cristata](#)  
 -powered by uBio

Album of Abyssinian birds and mammals / Download/About this book

Zoom: Auto

Book contributed by University of Illinois Urbana Champaign







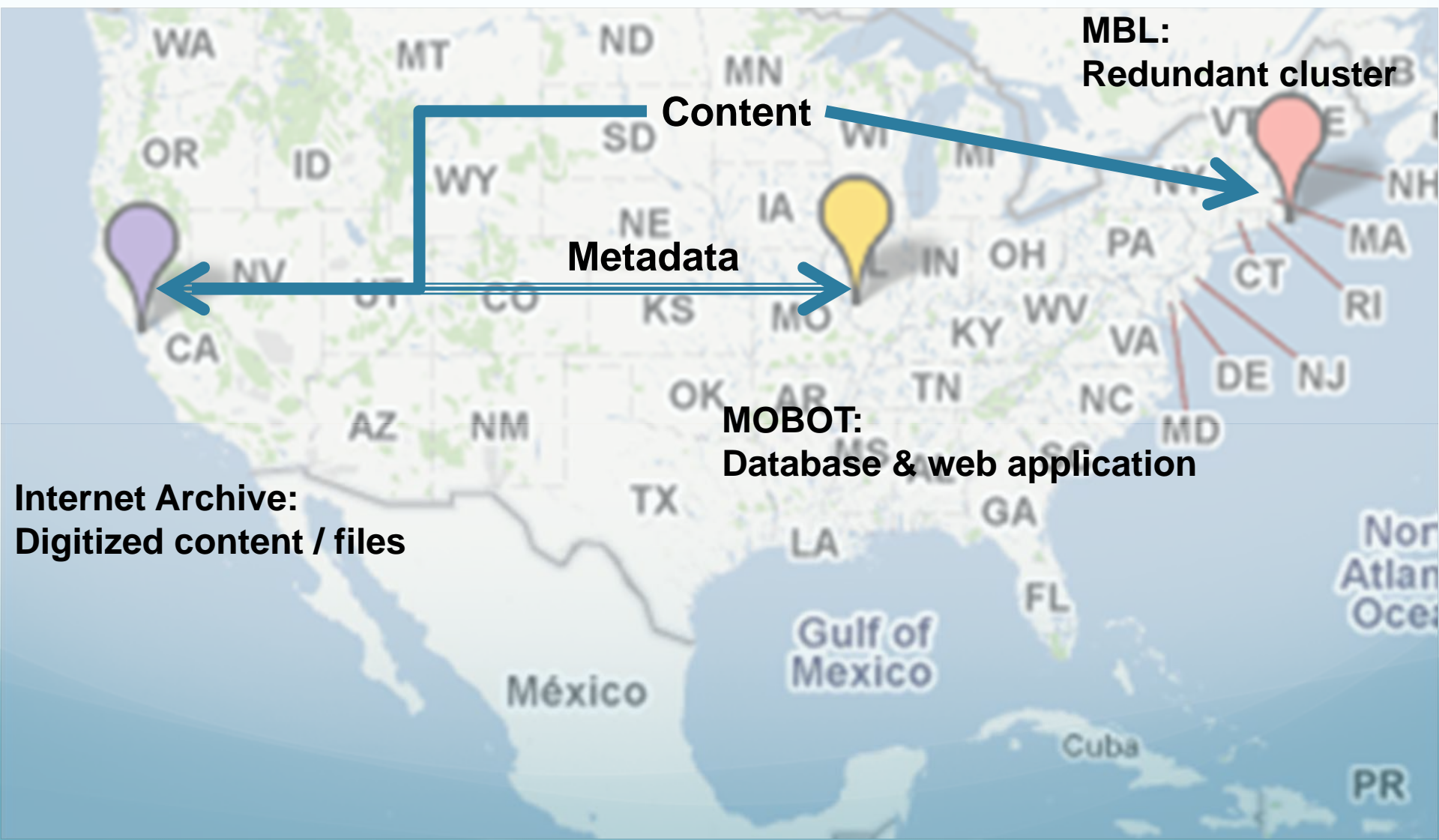
# BHL by the Book

One 380 pg (avg) volume = multiple files, varying sizes, relationships among them



	Name	Last Modified	Size	Type
	Parent Directory/		-	Directory
	<a href="#">mushroomsofameri00palm.djvu</a>	2007-Oct-10 00:13:27	704.9K	image/vnd.djvu
	<a href="#">mushroomsofameri00palm.gif</a>	2007-Oct-09 23:56:53	253.9K	image/gif
PDF	<a href="#">mushroomsofameri00palm.pdf</a>	2007-Oct-10 00:15:56	2.0M	application/pdf
	<a href="#">mushroomsofameri00palm_abbey.gz</a>	2007-Oct-10 00:11:19	367.7K	application/octet-stream
	<a href="#">mushroomsofameri00palm_bw.pdf</a>	2007-Oct-10 00:36:19	1.6M	application/pdf
	<a href="#">mushroomsofameri00palm_dc.xml</a>	2007-Oct-04 13:19:41	0.4K	application/xml
OCR	<a href="#">mushroomsofameri00palm_djvu.txt</a>	2007-Oct-10 00:36:28	24.9K	text/plain
	<a href="#">mushroomsofameri00palm_djvu.xml</a>	2007-Oct-10 00:11:35	225.1K	application/xml
	<a href="#">mushroomsofameri00palm_files.xml</a>	2008-Apr-30 16:10:51	4.4K	application/xml
JP2	<a href="#">mushroomsofameri00palm_flippy.zip</a>	2007-Oct-09 23:57:09	848.8K	application/zip
	<a href="#">mushroomsofameri00palm_jp2.zip</a>	2007-Oct-09 23:56:36	10.5M	application/zip
	<a href="#">mushroomsofameri00palm_marc.xml</a>	2007-Oct-04 13:19:41	2.1K	application/xml
	<a href="#">mushroomsofameri00palm_meta.mrc</a>	2007-Oct-04 13:19:41	0.6K	application/octet-stream
	<a href="#">mushroomsofameri00palm_meta.xml</a>	2008-Feb-05 18:47:48	1.4K	application/xml
XML	<a href="#">mushroomsofameri00palm_metasource.xml</a>	2007-Oct-04 13:19:41	0.4K	application/xml
	<a href="#">mushroomsofameri00palm_orig_jp2.tar</a>	2007-Oct-09 19:27:16	24.2M	application/x-tar
	<a href="#">mushroomsofameri00palm_scandata.xml</a>	2007-Oct-09 19:27:15	24.5K	application/xml

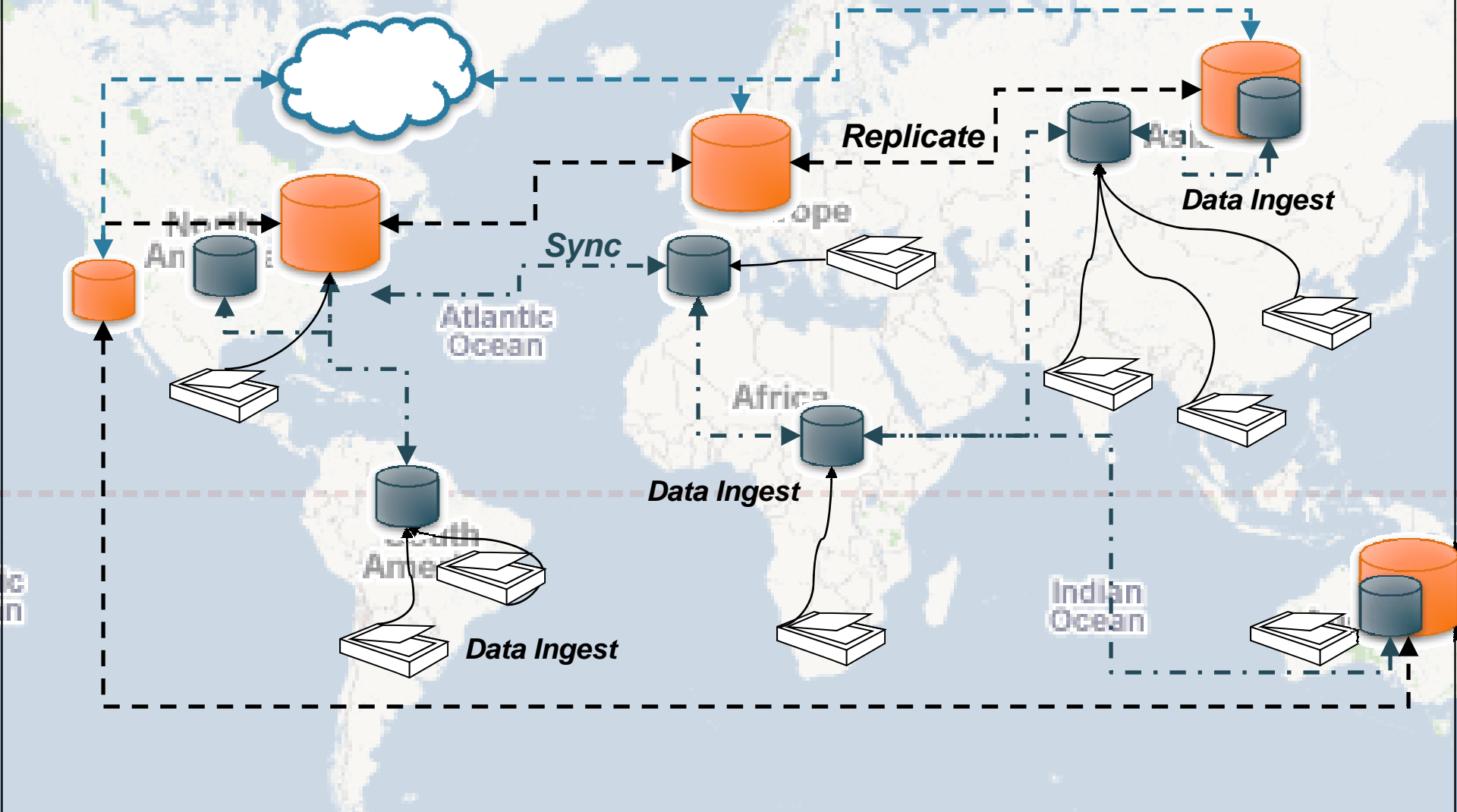
> 70TB, growing every day...

# Current distributed infrastructure



# BHL Vision: Global Infrastructure

-  Preservation System – multiple redundant copies of all digitized content.
-  Access System – files, metadata & services needed to deliver content.



## Motivation for joining pilot

- Community interest in cloud storage
  - (Funding organizations, too!)
- Wanted to evaluate applicability of cloud storage for large-scale digitization activities
  - Solutions for efficient transfer of 10-100s TB data
  - Lower cost alternatives to maintaining large data centers





# BHL as a research space

- BHL nodes are autonomous centers serving the digitized texts under their applications in response to users.
- But the BHL corpus as whole is a data set of biodiversity data in its own right. Embedded in it are:
  - Predator/prey relationships
  - Habitat/distribution data
  - Host/parasite data
  - Pathogen/disease vector data
- Third party researchers and projects are interested in mining the BHL texts for multiple research needs.
- One site for serving/accessing/downloading digital texts AND for data mining is messy. Separate out and put a version of the corpus in a public-like cloud space.

# BHL Policy Challenges

- **Money** - At present in the US, one BHL member library (MBL) is willing to provide essentially free redundant hosting. This is a very attractive financial offer. Since the MBL is BHL member it provides a level of administrative commitment. If this changes, DuraCloud becomes very attractive.
- **Skill level** - Multiple global partners needing all or some of the current holdings - have varying levels of technical skills. For some shipping hard drives might be easier. For some uploading to and downloading from DuraCloud might be preferable.
- **Timing** – at the time of the closing of the pilot our partners, while very close are not quite ready for the initial large data transfer. As they get their marbles lined up, we can evaluate DuraCloud as a transfer mechanism on a node-by-node basis.
- **Control** – in cultural-scientific digital projects no clear models using DuraCloud. Early-adopter paranoia.

# Data Transfer Methods & Limitations



Problems: Hardware failure, data loss, shipping fees

VS



Problems: Available bandwidth, upload/download fees

# Data transfer: Cloud vs. Cluster

- Inventory & audit lists
- Checksums for data integrity
- Heavy lifting at BHL scale, regardless of endpoint
  - weeks->months, not minutes->days
- Differences
  - In cluster environment, have to be intimately involved in hardware decisions, maintenance, troubleshooting
  - In cloud environment, those worries are part of your fee

# Challenges for adopting cloud storage

- BHL is embedded in longstanding institutions with megainfrastructure.
  - Already support data storage & maintenance at BHL scale
- Little funding for alternative infrastructure / storage
  - Current storage is (really, truly) free through Internet Archive
- Costs associated with download / use of content
  - BHL is a global resource for a broad community
  - User community wants to “do things” with data

## Lessons Learned

- Cloud infrastructure & applicability to BHL are no longer a mystery
- Nothing is free
  - Except when it is
- Cloud storage provides ability to quickly scale infrastructure
  - No lost time procuring & configuring hardware
- Useful for the right kinds of datasets
  - It's not the size of the corpus, it's the size of the files
  - Huge files are problematic

# Outcomes

- 10-13TB transferred from Internet Archive to DuraCloud over wire
  - Simple, without bandwidth limitations
- Became intimately familiar with our data
  - Larger files in corpus than expected (GB+ files)
  - Issues with “checksums”
  - Need to know your data to efficiently manage it
- Spent less time moving data than checking / verifying data

# Perceptions about cloud infrastructure after pilot participation

- More possibilities than expected:
  - Features
  - Movement
  - Support available from commercial providers.
  - Increasing menus of choices
- There is no silver bullet
  - Cloud is just a different endpoint for file storage
  - It doesn't solve all problems related to repository management



# Future opportunities for cloud infrastructure & BHL

- Depending on BHL partner needs/abilities use DuraCloud to transfer/synch files
- Seek research grants for data mining and include line items for DuraCloud hosting of BHL “research space” for multiple informatics projects.
- If “free” turns into “not so free” use DuraCloud as ongoing redundant preservation storage.
- As we explore synchronization across projects, is DuraCloud an alternative?



# BHL in the cloud



**A Pilot Project**

**Tom Garnett, BHL Executive Director**  
**Chris Freeland, BHL Technical Director**

<http://www.biodiversitylibrary.org>

# WGBH DuraCloud Pilot

---

Audio and video in the cloud

Digital Preservation Partners Meeting  
Wednesday, July 21, 2010



Peter Pinch  
Director of Technology for Interactive



# WGBH Media Library and Archives

36



# DuraCloud Pilot Goals

37

- Access
  - Streaming video
  - Integration with <http://openvault.wgbh.org>
  - Cost savings?
  - Improved sustainability?
- Preservation
  - Uncompressed audio and video storage
  - Cost Savings?
  - Improved reliability?
- Future Services

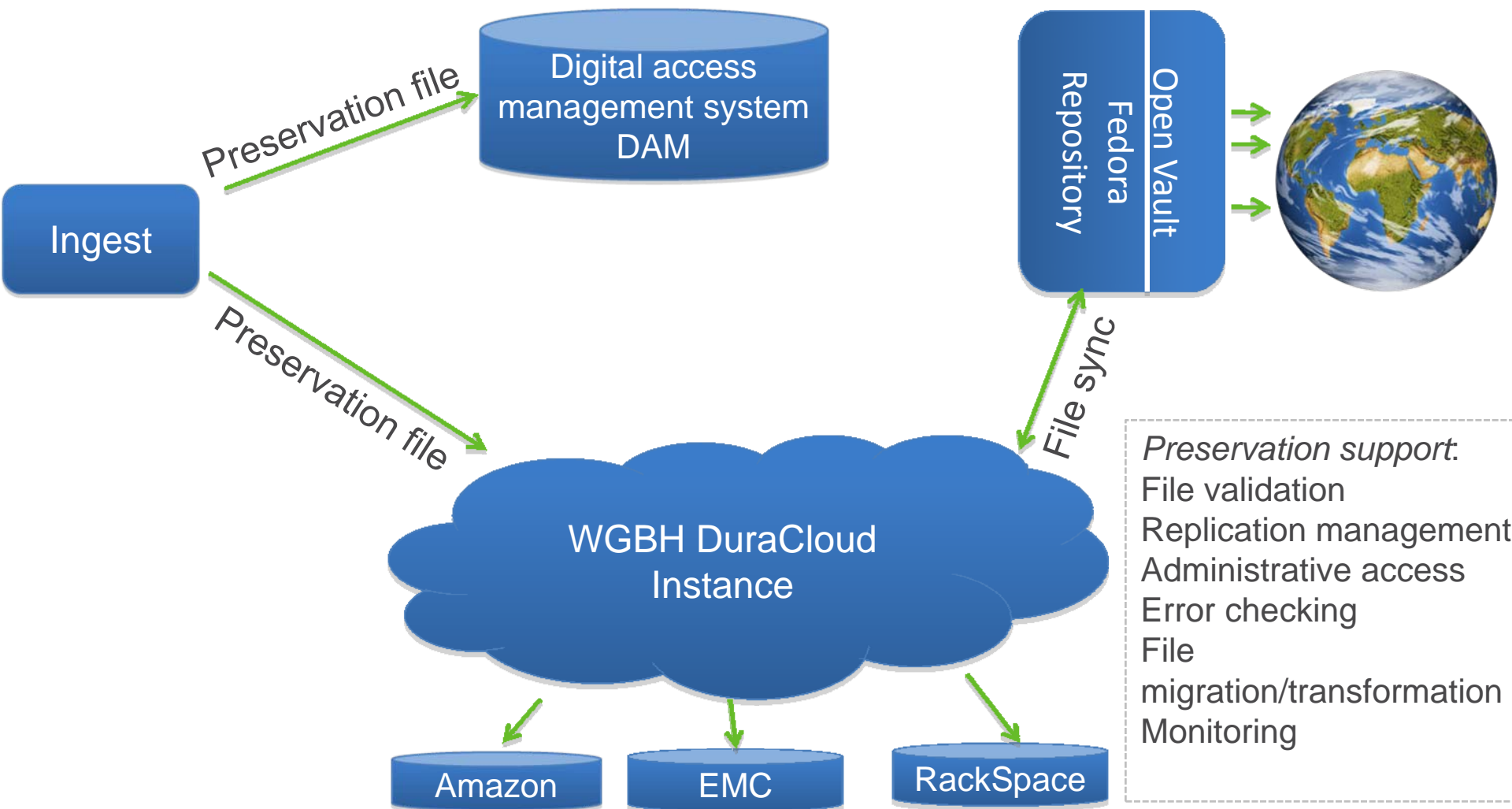
# DuraCloud Use Case: American Archive Pilot

38

- CPB pilot project, 20 stations including WGBH
  - civil rights era and World War II
  - Stations responsible for preservation & hosting
- Preservation
  - 110 hrs of video, 8.5 TB
  - 120 hrs of audio, 150 GB
- Access (streaming)
  - 12GB of H.264 video
  - 4GB of mp3 audio

# Preservation using DuraCloud

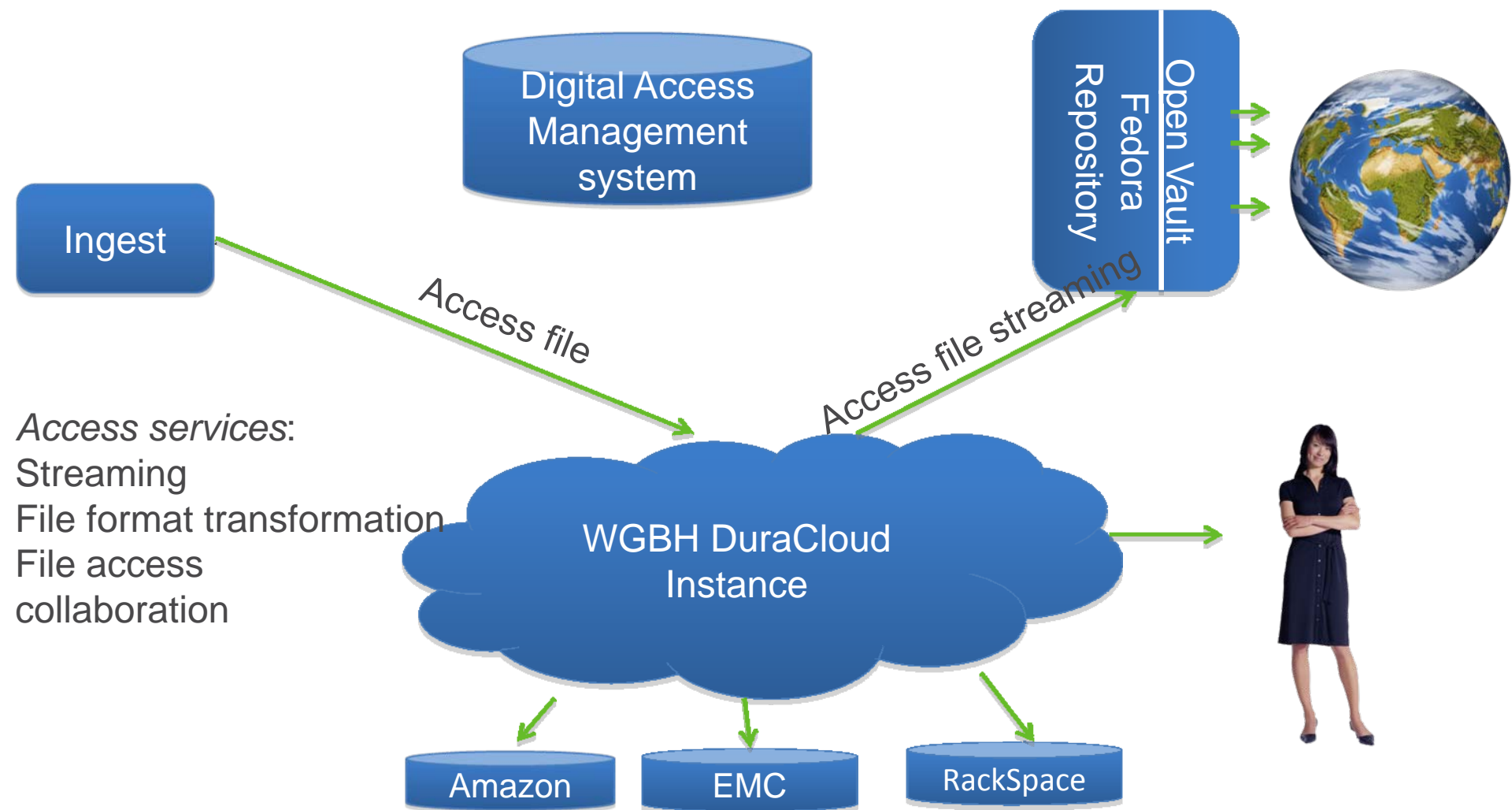
39





# Access using DuraCloud

40



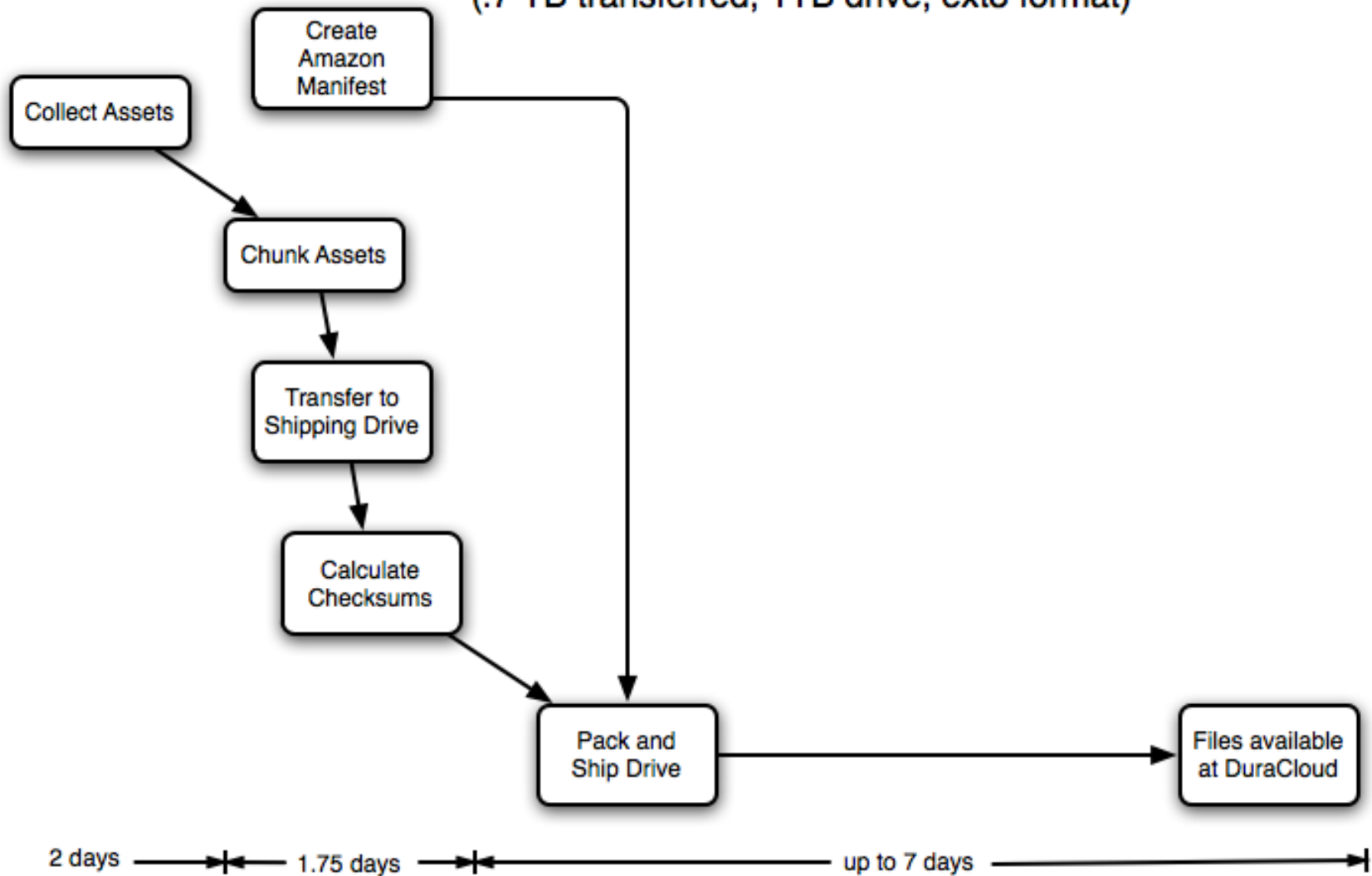


# Lessons Learned

# Sending disks to the cloud

42

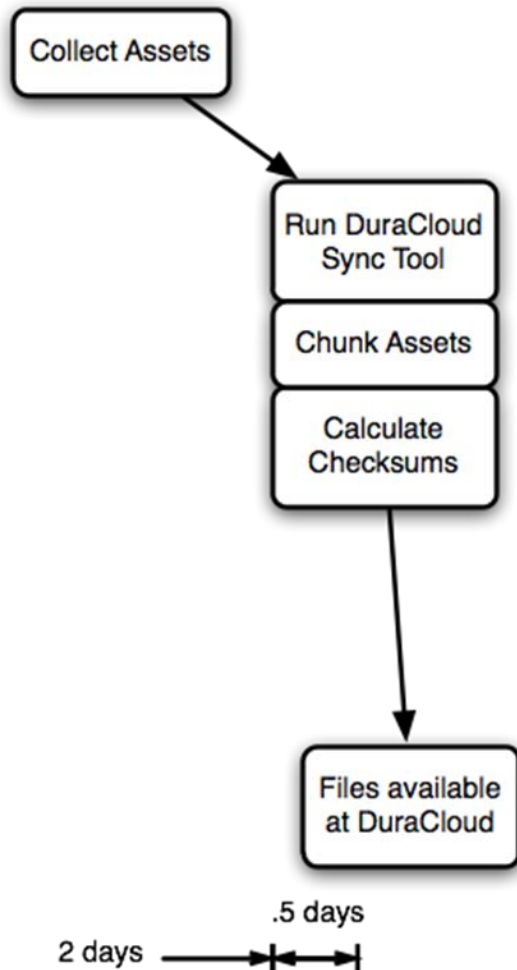
## WGBH Delivery to DuraCloud via Hard Drive (.7 TB) (.7 TB transferred, 1TB drive, ext3 format)



# Using tubes and wires

43

## WGBH Delivery to DuraCloud via Internet (.7 TB transferred, 300Mbps pipe)



- Gathering data
- Sunk costs
  - DAM (including hierarchical storage)
  - Bandwidth (to the cloud)
- Incremental costs
  - Off-line storage
  - Cloud storage
  - Streaming bandwidth (+1 for cloud)

# Cost Comparisons

45

	In-house	Cloud
Bandwidth to storage	n/a	“free”
Bandwidth for access	\$1 per GB transferred	17¢ per GB transferred
Storage	8.8¢ per GB	15¢ per GB per month

- 5.5 TB of audio & video uploaded
  - Preservation and access files
- Still working with sync tool
- Streaming service works
  - But still need to integrate with Open Vault web site (for access)

# The Future

47

- Complete integration with Open Vault site
- Dealing with file size limits
  - Editing (clipping)
- Transcode services?
  - Proposal with NCSA
- Speech to text?
  - Transcript alignment
- Recommend for American Archive when it moves to preservation phase of project

Questions?

<http://openvault.wgbh.org>

[peter\\_pinch@wgbh.org](mailto:peter_pinch@wgbh.org)



## Achievements during Initial Pilot

- Demonstrated large scale data transfer
  - 30 TB moved into the cloud
- Demonstrated feasibility of large scale data processing in the cloud
  - Image format conversion
- Demonstrated cloud capabilities for content access
  - Media streaming
  - Image display

# Lessons Learned

- Initial content load requires time and effort
  - Preparing content for transfer is often non-trivial
  - Transfer over http: simpler, faster, and cheaper than disks
  - Time to load content: determined by bandwidth available at the source location
  - Client-side utilities can help ease burden
- Tool development is required to overcome or mitigate cloud provider limitations
- Latency due to transfer over the web can be an issue for applications
  - Minimize transactions across the wire
  - Keep data close to compute
- Must allow for “eventual consistency”
  - Adds to latency if existence guarantees are required

# Lessons Learned

- Cloud market is still developing
  - New capabilities becoming available frequently
  - Sun and EMC have both exited the market in the past year
- Storage capabilities more mature than Compute
  - Cloud storage provides robust performance
  - Storage APIs beginning to converge
  - Compute services and capabilities vary widely
- Each vendor is seeking ways to differentiate offerings
  - Amazon way out in front
  - Building only to lowest-common-denominator equals missed opportunities to leverage provider offerings

# Expanded Pilot Partners

University	Use Case	Repository
Rice U	Preservation	DSpace, meta archive
Hamilton College	Access/international collaboration	Fedora
Northwestern U	Preservation books, audio, image	Fedora
U of PEI	Image viewing/hosting	Fedora/Islandora
Cornell U	Data stream access and preservation	Fedora
ICPSR	Access and Preservation	Fedora
SUNY Buffalo	Preservation	DSpace
IUPUI	Preservation	DSpace
Rhodes College	Image Access	DSpace
North Carolina State U	Preservation	DSpace
CARL	Preservation and Services	Fedora
Orbis Cascade Alliance	Preservation and Services	DSpace
MIT	Preservation, OAIS compliance	Dspace
NYPL	Preservation and Services	Fedora
WGBH	Access and Preservation	DAM

# Thank You!

<http://www.duracloud.org>

<https://wiki.duraspace.org/display/duracloud>