Getting Creative: Moving Digital Content to the Cloud

Presenter: Jim Coleman
Brief Overview of the Center

- The Center for Creative Photography was founded by then President John Shaefer and Ansel Adams
- It opened in 1975 with the Archives of five living master photographers – Ansel Adams, Wynn Bullock, Harry Callahan, Fredrick Sommer, and Aaron Siskind
- Its collection includes more than 130,000 fine prints by more than 2,200 photographers
- The archives have grown to include more than 239 collections, comprising more than 5,000,000 individual items
John and Ansel
The Center
What's and Wherefores

- Center’s IT legacy
- Transition from support by the Library to UITS
- Evolving strategies to manage the Center’s image collection and image data
- Using local software, custom workflows, AWS S3, and Glacier as foundation for cost containment and on-going object management
Reverse Q & A

- How much data do you have?
- Number of files? Number of file paths?
- Data naming conventions?
- Data duplication? How many files? How much data in the aggregate? How do you know they are duplicates?
- How easy is it to find material for you? For your users?
- Vulnerability to change/alteration? How would you know?
Center’s Reverse Q & A

```
-- How Much Data Do You Have?

```sql
select sum(file_size)/(1024 * 1024 * 1024) 'How Much Data' from assets where FILE_TYPE in ('jpg', 'tif')
```

-- How Many Files?

```sql
select count(*) 'How Many Files' from assets where FILE_TYPE in ('jpg', 'tif')
```

-- How Many File Paths?

```sql
select count(distinct file_path) 'How Many File Paths' from assets where FILE_TYPE in ('jpg', 'tif')
```

-- How Much Data Duplication?

```sql
select sum(file_size)/(1024 * 1024 * 1024) 'How Much Data Duplication' from assets a,
hash_count_vw b
where a.MD160HASH = b.MD160HASH
and a.FILE_TYPE in ('jpg', 'tif')
and b.count > 1
```

-- How Many Duplicate Files?

```sql
select count(*) 'How Many Duplicate Files' from assets a,
hash_count_vw b
where a.MD160HASH = b.MD160HASH
and a.FILE_TYPE in ('jpg', 'tif')
and b.count > 1
```
Maroon Bells, near Aspen, Colorado

Ansel Adams
Center’s IT Infrastructure

- Supported by Library IT until the Center moved under Vice President for Research in 2013
- Three month project to move desktops to UITS and 20TB of image and document data from the Library to Central IT
- Initiated thinking about data archiving / backup / recovery with initial solutions provided by Central IT
Cost of Doing Business

- While under the Library, costs were largely “hidden” and not part of the Center’s budget
- Move from the Library IT to UITS made the costs of doing business visible
- Costs included
  - local storage
  - local storage replication
  - back-up of both documents and objects
  - Servers for the Center’s applications, databases, local workflows
  - Time Involved to accomplish all of the data/file manipulation
## Costs of Doing Business, Year 1

<table>
<thead>
<tr>
<th>Storage</th>
<th>Quantity</th>
<th>On-going Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Backup</td>
<td>18TB</td>
<td>37,000</td>
</tr>
<tr>
<td>Image Local</td>
<td>18TB</td>
<td>6,200</td>
</tr>
<tr>
<td>Image Local Rep</td>
<td>18TB</td>
<td>6,200</td>
</tr>
<tr>
<td>User Document</td>
<td>750GB</td>
<td>255</td>
</tr>
<tr>
<td>User Document Backup</td>
<td>750GB</td>
<td>5,700</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>55,355</strong></td>
</tr>
</tbody>
</table>
Quest of Continual Becoming

Jerry Uelsmann
Outsourcing Storage

- CCP conducted a selection process among five Cloud providers
- At the time of the initial vendor selection, Amazon was not an option
- CCP chose Zetta.net for Year 2 because they met all of the requirements and were the fastest by an order of magnitude in backup and recovery
- Zetta, although meeting all requirements and decidedly better than the local option, was as expensive as local backup solutions
Vendor Selection Considerations

- It must be easy for a technically savvy user to configure synchronizations and data restores without needing to write scripts or code
- Solution must be able to sync with CCP shared directories
- Up- and downloads must be performant, i.e., all of the CCP mission-critical document data (approximately 300GB) must be capable of a restore within a business day or less (8 hours)
- Solution should be able to do restores in place (i.e., restore to the original site) and maintain file attributes and file path information
- Solution should be able to maintain versions of files (i.e., sync as of a specific day) for at least <n> months
- Solution should report at least daily on its health (e.g., email indicating the status of restores)
- Solution should have built-in reporting
- Solution should have 24/7 support
AWS, S3, and Glacier

- With the agreement with Amazon, the Center was able to look at other solutions that would provide much of the same benefits as Zetta, but at a lower cost.
- Local, managed storage or cloud storage like S3 has predictable pricing. Here at UA, Isilon and S3 are very similar, but S3 requires home-grown or third-party interfaces and active maintenance.
- Glacier is cheap, but requires home-grown or third-party interfaces, and active maintenance.
- Zetta was a product, Amazon is largely “roll you own” in the space at the current time.
Toadstool
Edward Weston
Content Discovery

- Initially managed by file path / naming conventions
- Need for something to capture and manage file characteristics/metadata
- Absent a DAM (which would automate much of this), write software that captures and posts this information to a database
Content Management

- We capture file name, path, size, date added/updated, MD5 hash, Glacier/S3 key as basic metadata
- The process included an initial metadata data load and MD5 signature creation
- Now nightly scheduled tasks to keep the data store refreshed
- Logic can determine other actions as part of the load process, e.g., TIFF to JP2 conversions, loads to S3 or Glacier
Livia
Fredrick Sommer
In addition to the gathering of basic object metadata, the addition of content metadata at this point is useful/essential.

Absent specific mechanisms to gather this additional metadata, path information can likewise be harvested at the point of ingestion.

Absent a DAMS, metadata ingestion can occur by embedding object metadata at the point of imaging using standard or custom profiles.

When implementing a DAMS, all of the discussion around object and content management still pertains.

The amount of time required to accomplish this should not be underestimated.
Current Architecture

- Master TIFF’s are created by Imaging and placed into local archival space
- TIFF’s are loaded into Glacier (unique only)
- JPEG2000’s are then processed, put into local archival space, and loaded into S3 (unique only)
- Master TIFF’s are then removed from local archival space
- Two copies of JP2’s, one archival copy of TIFF, local use copies created as necessary
Current Archive Architecture Metrics

- Move Master TIFF files to Glacier (6 weeks initially)
- Convert Master TIFF files to JP2 (3 weeks initially)
- Verify JP2 / deal with outliers that do not convert
- Move JP2 to S3 (2-3 weeks initially)
- Remove, as necessary, master TIFF files locally
- Use JP2 to regenerate TIFF as necessary
- Retrieve Master TIFF files from Glacier as necessary
- Develop optimal workflow to support this architecture
- Migrate to a DAMS when the University adopts a suitable platform.
Current Technologies Used

- AWS S3 Web Services API and Workflow
- Kakadu as JPEG2000 rendering engine
- JPYLYZER for JPEG2000 verification
- .NET Windows Forms App for Loading/Management
- MSSQL Server as back-end database
- User-facing .NET Web Pages
- Scheduled Tasks calling C# jobs for On-going Maintenance
Cost Containment and Management

- Cost considerations have a major impact on how the Center has chosen to implement its local and archival storage solutions.
- Deciding to use JPEG2000 as the Center’s main image format, using 6:1 compression, results in a large savings in storage size and costs.
- Retaining archival TIFF in Glacier gives the Center some confidence in maintaining stewardship of its digital objects while enjoying significant savings.
- JPEG2000 has been chosen as the archival format of choice by many cultural heritage institutions.
- That said, its adoption by the community and by software vendors to date has been limited.
Untitled
Aaron Siskind
Now vs Then vs Future
Costs of Doing Business, Then

<table>
<thead>
<tr>
<th>Storage</th>
<th>Quantity</th>
<th>On-going Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Backup</td>
<td>18TB</td>
<td>37,000</td>
</tr>
<tr>
<td>Image Local</td>
<td>18TB</td>
<td>6,200</td>
</tr>
<tr>
<td>Image Local Rep</td>
<td>18TB</td>
<td>6,200</td>
</tr>
<tr>
<td>User Document</td>
<td>750GB</td>
<td>255</td>
</tr>
<tr>
<td>User Document Backup</td>
<td>750GB</td>
<td>5,700</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>55,355</strong></td>
</tr>
</tbody>
</table>
Cost of Doing Business, Now

<table>
<thead>
<tr>
<th>Storage</th>
<th>Quantity</th>
<th>On-going</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Storage (JP2. etc.)</td>
<td>4TB</td>
<td>1390</td>
</tr>
<tr>
<td>Image Storage Backup (S3)</td>
<td>4TB</td>
<td>1500</td>
</tr>
<tr>
<td>Image Storage Backup (Glacier)</td>
<td>17TB</td>
<td>2090</td>
</tr>
<tr>
<td>Document Storage</td>
<td>750GB</td>
<td>255</td>
</tr>
<tr>
<td>Document Backup (Zetta)</td>
<td>1TB</td>
<td>4000</td>
</tr>
<tr>
<td>Document Backup (Local)</td>
<td>500GB</td>
<td>4060</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>13,235</strong></td>
</tr>
</tbody>
</table>

**Net Savings over Year 1**

42,120

Note that this savings is to the base, and thus realized every year.
What Keeps Us Awake

- Long-term viability of JPEG2000, support by community and vendors
- Catastrophic event requiring the retrieval from Glacier of a significant part of the Center’s content
- Continuing to leverage local infrastructure for a local solution
- Continuing to use a solution that is home-grown and uses technologies not widely adopted by the Central IT (.NET, .NET Web Services, Windows Forms)
Panorama of Paris, Suzy Parker in Jacques Fath Gown

Louise Dahl-Wolfe
Q & A