DIGITAL PRESERVATION FOR INSTITUTIONAL REPOSITORIES

ELIZABETH LA BEAUD
DIGITAL LAB MANAGER, UNIVERSITY OF SOUTHERN MISSISSIPPI
ACCESS VS. PRESERVATION

ACCESS

- relies on cutting edge technologies to provide best and fastest access at a point in time
- selects metadata needed to use and understand content
- access systems deliver objects with user-oriented services to make the objects
- purpose: provide content to users
- focus: current users

PRESERVATION

- relies upon proven technologies to preserve digital objects across generations of technology
- accumulates metadata over the life cycle to trace and preserve content
- preservation systems create new versions of digital objects for access to deliver as needs change over time
- purpose: ensure long-term access
- focus: future users
RISKS TO DIGITAL CONTENT

- Obsolescence – hardware & software
- Loss of power
- Hardware failure
- Change and loss – accidental and intentional
- Bit rot
- User error
- Man-made and natural disasters

- Cyber attacks
- Format obsolescence
- Media failure
- Inappropriate access (i.e. confidential data)
- Non-compliance – standards and requirements
- Dust
PRESERVATION MAKES LONG-TERM ACCESS POSSIBLE...
# NDSA’s Levels of Digital Preservation

<table>
<thead>
<tr>
<th>Storage and Geographic Location</th>
<th>Level 1 (Protect your data)</th>
<th>Level 2 (Know your data)</th>
<th>Level 3 (Monitor your data)</th>
<th>Level 4 (Repair your data)</th>
</tr>
</thead>
</table>
| - Two complete copies that are not collocated  
- For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system | - At least three complete copies  
- At least one copy in a different geographic location  
- Document your storage system(s) and storage media and what you need to use them | - At least one copy in a geographic location with a different disaster threat  
- Obsolescence monitoring process for your storage system(s) and media | - At least three copies in geographic locations with different disaster threats  
- Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems |

<table>
<thead>
<tr>
<th>File Fixity and Data Integrity</th>
<th>Level 1 (Protect your data)</th>
<th>Level 2 (Know your data)</th>
<th>Level 3 (Monitor your data)</th>
<th>Level 4 (Repair your data)</th>
</tr>
</thead>
</table>
| - Check file fixity on ingest if it has been provided with the content  
- Create fixity info if it wasn’t provided with the content | - Check fixity on all ingests  
- Use write-blockers when working with original media  
- Virus-check high risk content | - Check fixity of content at fixed intervals  
- Maintain logs of fixity info; supply audit on demand  
- Ability to detect corrupt data  
- Virus-check all content | - Check fixity of all content in response to specific events or activities  
- Ability to replace/repair corrupted data  
- Ensure no one person has write access to all copies |

<table>
<thead>
<tr>
<th>Information Security</th>
<th>Level 1 (Protect your data)</th>
<th>Level 2 (Know your data)</th>
<th>Level 3 (Monitor your data)</th>
<th>Level 4 (Repair your data)</th>
</tr>
</thead>
</table>
| - Identify who has read, write, move and delete authorization to individual files  
- Restrict who has those authorizations to individual files | - Document access restrictions for content  
- Maintain logs of who performed what actions on files, including deletions and preservation actions | - Perform audit of logs | |

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Level 1 (Protect your data)</th>
<th>Level 2 (Know your data)</th>
<th>Level 3 (Monitor your data)</th>
<th>Level 4 (Repair your data)</th>
</tr>
</thead>
</table>
| - Inventory of content and its storage location  
- Ensure backup and non-collocation of inventory | - Store administrative metadata  
- Store transformative metadata and log events | - Store standard technical and descriptive metadata | - Store standard preservation metadata |

<table>
<thead>
<tr>
<th>File Formats</th>
<th>Level 1 (Protect your data)</th>
<th>Level 2 (Know your data)</th>
<th>Level 3 (Monitor your data)</th>
<th>Level 4 (Repair your data)</th>
</tr>
</thead>
</table>
| - When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs | - Inventory of file formats in use  
- Monitor file format obsolescence issues | |
| | | | |
You have options:

- Buy, build, or join
- Local installation

- Open source
- Hosted solutions
- Proprietary
CONSIDER…

– Cost (available resources for preservation)
– Quantity (size and number of files)
– Expertise (skills required to manage)
– Partnerships (geographic distribution)
– Services (outsourcing)
WHERE IS YOUR DATA?
WHERE IS THE CLOUD?
Global Network of Regions and Edge Locations

The AWS Cloud spans 52 Availability Zones within 18 geographic Regions around the world, with announced plans for 12 more Availability Zones and four more Regions in Bahrain, Hong Kong SAR, Sweden, and a second AWS GovCloud Region in the US.
AT LEAST TWO COPIES IN TWO LOCATIONS…

| Storage and Geographic Location | - Two complete copies that are not collocated  
| - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system | - At least three complete copies  
| - At least one copy in a different geographic location  
| - Document your storage system(s) and storage media and what you need to use them | - At least one copy in a geographic location with a different disaster threat  
| - Obsolescence monitoring process for your storage system(s) and media | - At least three copies in geographic locations with different disaster threats  
| - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems |
PLAN FOR STORAGE SYSTEM REPLACEMENT
WHO HAS ACCESS TO YOUR DATA?
WHAT HAPPENS WHEN THEY LEAVE?
<table>
<thead>
<tr>
<th>Information Security</th>
<th>Document access restrictions for content</th>
<th>Maintain logs of who performed what actions on files, including deletions and preservation actions</th>
<th>Perform audit of logs</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Identify who has read, write, move and delete</td>
<td></td>
<td>- Maintain logs of who performed what actions on files, including deletions and preservation actions</td>
<td></td>
</tr>
<tr>
<td>authorization to individual files</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Restrict who has those authorizations to individual</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>files</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Make sure no one person has write privileges to all copies.
<table>
<thead>
<tr>
<th>File Fixity and Data Integrity</th>
<th>Check file fixity on ingest if it has been provided with the content</th>
<th>Check fixity on all ingests</th>
<th>Check fixity of content at fixed intervals</th>
<th>Check fixity of all content in response to specific events or activities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Create fixity info if it wasn’t provided with the content</td>
<td>- Use write-blockers when working with original media</td>
<td>- Maintain logs of fixity info; supply audit on demand</td>
<td>- Ability to replace/repair corrupted data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Virus-check high risk content</td>
<td>- Ability to detect corrupt data</td>
<td>- Ensure no one person has write access to all copies</td>
</tr>
</tbody>
</table>
ALERTS YOU TO A PROBLEM

- Checksums
  - SHA256
  - MD5
- MD5 Summer
- AVPreserve’s Fixity
- File Verifier ++
  - (these are windows examples)
CAN YOUR VENDOR PROVE INTEGRITY ON DEMAND?

HOW OFTEN DO THEY CHECK FILE INTEGRITY?
| File Formats | - When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs | - Inventory of file formats in use | - Monitor file format obsolescence issues | - Perform format migrations, emulation and similar activities as needed |
WHAT ARE YOU ACCEPTING?

HOW ARE YOU GOING TO MIGRATE IT OVER TIME?
TIERED OPTIONS

- **Tier 1**
  - Open Sustainable File Formats (e.g., .txt)
  - We will preserve your data to the best of our ability, and migrate the content over time to ensure accessibility. We’ll keep a copy of your original file as well, just in case you want to revisit the original bitstream (1’s and 0’s).

- **Tier 2**
  - Some well-known, but proprietary files (e.g., Microsoft Word)
  - We will preserve your original file, but are limited in how we can maintain it. We’ll always keep the 1’s and 0’s in the right order, but some formatting or other features may change.

- **Tier 3**
  - Proprietary file formats with little to no open documentation (e.g., .spss)
  - There are times when a specialized format is required for your data. In these cases, we will work with you. We will commit to preserving the original bitstream (1’s and 0’s) and the submitted metadata; however, while the files may or may not be renderable in the future, we can make no guarantees on the form or presentation it will take.
SUSTAINABLE FILE FORMAT CHARACTERISTICS

- Widely used by appropriate research community
- Non-proprietary
- Well documented, open standard
- Uncompressed (size permitting)
- Unencrypted
- Encoded in standard representation (e.g., ASCII, UTF-8)
## Metadata

| Metadata | - Inventory of content and its storage location  
| - Ensure backup and non-collocation of inventory | - Store administrative metadata  
| - Store transformative metadata and log events | - Store standard technical and descriptive metadata  
| - Store standard preservation metadata |
DESCRIBE IT SO YOU CAN FIND IT!
METADATA FIELDS

- Identifier
- Title
- Description
- Creator
- Department
- Legacy department
- Date
- coverage
- Subjects
- Geographic location
- Language
- Type
- File format
- Contributors
- Permissions
- Disciplines
- Related docs
RECORDS CHANGES OVER TIME

- Document metadata changes
- File format migrations
- IP problems & resolutions
- Deaccessions
TRUCK BOOK...
IT’S A COMMITMENT…
IT’S NOT A ONE AND DONE…
BUT...YOU DON’T HAVE TO DO IT ALONE...

RESOURCES
THANK YOU !!

ELIZABETH LA BEAUD
ELIZABETH.LABEAUD@USM.EDU