The Cloudy Outlook for Digital Preservation

Nicholas Taylor (@nullhandle)
Program Manager Emeritus, LOCKSS and Web Archiving
Stanford Libraries

International Conference on Digital Preservation
Cloud Atlas: Navigating the Cloud for Digital Preservation
17 September 2019
“The Cloud” is playing a growing role in digital preservation.
Which “The Cloud” we use, and how we use it, matters both for our missions and the likely success of our efforts.
overview

• threat modeling
• commercial cloud
• community cloud
• wrap up
Threat Modeling
threats to digital information

- **failures**: media, hardware, software, network services, economic, organizational
- **obsolescence**: media, hardware, software
- **errors**: communication, operator
- **attacks**: external, internal
- **natural disaster**
### community best practices

<table>
<thead>
<tr>
<th>Table 1: Version 1 of the Levels of Digital Preservation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>
| 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 |
| File Fixity and Data Integrity | - 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 |

NDSA Levels of Preservation Working Group: “NDSA Levels of Digital Preservation”
what threats are addressed?

- **failures**: media, hardware, software, network services, economic, organizational
- **obsolescence**: media, hardware, software
- **errors**: communication, operator
- **attacks**: external, internal
- **natural disaster**

<table>
<thead>
<tr>
<th>Table 1: Version 1 of the Levels of Digital Preservation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 4 (Repair your data)</strong></td>
</tr>
<tr>
<td><strong>Storage and Geographic Location</strong></td>
</tr>
<tr>
<td>- At least three copies in geographic locations with different disaster threats</td>
</tr>
<tr>
<td>- Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems</td>
</tr>
<tr>
<td><strong>File Fixity and Data Integrity</strong></td>
</tr>
<tr>
<td>- Check fixity of all content in response to specific events or activities</td>
</tr>
<tr>
<td>- Ability to replace/repair corrupted data</td>
</tr>
<tr>
<td>- Ensure no one person has write access to all copies</td>
</tr>
</tbody>
</table>

_NDSA Levels of Preservation Working Group: “NDSA Levels of Digital Preservation”_
what threats are discounted?

- **failures**: media, hardware, software, network services, economic, organizational
- **obsolescence**: media, hardware, software
- **errors**: communication, operator
- **attacks**: external, internal
- **natural disaster**

**Table 1: Version 1 of the Levels of Digital Preservation**

<table>
<thead>
<tr>
<th>Threat Category</th>
<th>Level 4 (Repair your data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage and Geographic Location</td>
<td>- At least three copies in geographic locations with different disaster threats</td>
</tr>
<tr>
<td></td>
<td>- Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems</td>
</tr>
<tr>
<td>File Fixity and Data Integrity</td>
<td>- Check fixity of all content in response to specific events or activities</td>
</tr>
<tr>
<td></td>
<td>- Ability to replace/repair corrupted data</td>
</tr>
<tr>
<td></td>
<td>- Ensure no one person has write access to all copies</td>
</tr>
</tbody>
</table>

*NDSA Levels of Preservation Working Group: “NDSA Levels of Digital Preservation”*
Amazon cloud crash wipes out customer data; Will users be compensated?

By Molly McHugh
Posted on April 28, 2011 10:57 am

Yahoo Quietly Pulls The Plug On Geocities

Posted Apr 23, 2009 by Leena Rao (@leenarao)

“Climate change” and “global warming” are disappearing from government websites

The deletions follow a pattern of policy changes on climate change under the Trump administration.

By Umair Irfan | Updated Jan 11, 2018, 12:30pm EST

Moon landing tapes got erased, NASA admits

Maggie Fox, Health, Science Editor

WASHINGTON (Reuters) - The original recordings of the first humans landing on the moon 40 years ago were erased and re-used,
understand + mitigate threats

- long-term **data integrity** is hard
- needs architecture informed by **actual leading threats** to data
- don’t underestimate:
  - people making **mistakes**
  - **attacks** on information
  - organizational **failure**
Commercial Cloud
commercial cloud considerations

- on-demand access
- economic lock-in
- reliability caveats
- opaque data integrity
- security configuration
- non-operational externalities
on-demand access

• minimize *idle built-out infrastructure* needed for:
  • long-tail access
  • data integrity checks

• shift costs for *corpus-level use cases*

• but *metered access*:
  • complicates cost modeling
  • works less well for popular or oft-accessed content
economic lock-in

- **CapEx → OpEx yields financial flexibility**
- **financial flexibility less valuable for inflexible commitment** (i.e., long-term preservation)
- prices decline but maybe not as quickly as in competitive market for local hardware refresh
- mitigating strategy: maintain a local copy

<table>
<thead>
<tr>
<th>service</th>
<th>ingest</th>
<th>store</th>
<th>export</th>
<th>lock-in factor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AMZN Glacier</strong></td>
<td>$2,250</td>
<td>$4,000</td>
<td>$55,240</td>
<td>13.8x</td>
</tr>
<tr>
<td><strong>GOOG Coldline</strong></td>
<td>$3,600</td>
<td>$7,000</td>
<td>$83,860</td>
<td>12x</td>
</tr>
<tr>
<td><strong>MSFT Archive</strong></td>
<td>$6,350</td>
<td>$2,000</td>
<td>$16,260</td>
<td>8.1x</td>
</tr>
</tbody>
</table>

David S.H. Rosenthal: “Cloud for Preservation”

“Dam II” by Craig Bennett under CC BY-NC 2.0
reliability caveats

• "11 nines" of reliability?
  • modeled on hardware failure
  • accounts for ⅓ of data losses

• ⅔ of data losses due to less rationalizable factors: attacks, errors, software failures
  • highly centralized infrastructure more vulnerable

• chance of billing error interrupting service non-trivially more significant than risk of loss suggested by reliability estimate
opaque data integrity

• **feature**, not a bug?
• hashing data in situ **requires trusting** that the service has performed computation rather than reporting cached value
• may be **prohibitively expensive** to retrieve content to a trusted environment to perform hashing
security configuration

• **monoculture** vulnerabilities

• greater affordances, better defaults for **on-premise security**

• consistent leaks from **misconfigured cloud services** suggest security is a challenge

“Insecure,” by nicholas j. nawroth under CC BY-NC-ND 2.0
not all clouds the same

• “The Cloud is just somebody else’s computer”
• values-aligned partnerships to **build private clouds** e.g.,
  • consortial/community
  • focused on particular content types (e.g., software, web archives)
  • for computational research

---

“Three Layer Storm Clouds” by Razvan Orvendovici under CC BY 2.0
community cloud considerations

- sustain community capacity
- flexibility + interoperability
- diversity + risk mitigation
- pilot models
sustain community capacity

• can we still claim to have **custody + intellectual control** over content stored in commercial cloud?

• can we afford to outsource functions **core to mission** to commercial cloud?

• **scholarly publishing** is an example of a service ceded to commercial providers
flexibility + interoperability

“[T]he needs of today’s diverse scholarly communities are not being met by the existing largely uncoordinated scholarly infrastructure, which is dominated by vendor products that take ownership of the scholarly process and data. We intend to create a new open infrastructure system that will enable us to work in a more integrated, collaborative and strategic way. It will support global connections and consistency where it is appropriate, and local and contextual requirements where that is needed.”

Invest in Open Leadership: “Preamble, The Why”
diversity + risk mitigation

• lots of copies is necessary but not sufficient
• central points of failure can undermine all copies at once
• multi-organizational preservation storage provides:
  • resilience against organizational failure
  • diversity in technical infrastructure
pilot models

• in original, Global LOCKSS Network, **all nodes stored copies**

• private LOCKSS networks moving towards **hosted service models**

• **subset of institutions** host infrastructure w/ governance by + funding from broader community

• **Stanford + trusted partners** may serve as anchor storage hosts
deliberate cloud strategy

- let’s be cautious about reducing future flexibility
- let’s understand the meaningful differences between use cases
- let’s be mindful of trade-offs
- let’s consider what else we can do on open infrastructure, together