Lots of Checksums
Keep Stuff Safer

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

Flexibility and Pragmatism: Thinking Differently about “Better” for Digital Preservation Services
CNI Spring Membership Meeting
8 April 2019
community best practices

<table>
<thead>
<tr>
<th>Table 1: Version 1 of the Levels of Digital Preservation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1 (Protect your data)</strong></td>
</tr>
<tr>
<td>Storage and Geographic Location</td>
</tr>
<tr>
<td>File Fixity and Data Integrity</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**NDSA Levels of Preservation Working Group:** “NDSA Levels of Digital Preservation”
lots more (purposeful) copies

• 3 copies too few for reliable long-term consensus on data integrity
• which is why LOCKSS prefers “lots of copies”
• LOCKSS copies don’t just provide idle redundancy
• LOCKSS copies also employed to provide:
  • data integrity attestations + consensus
  • high-confidence repairs
  • risk diversification
what are we protecting against?

- **familiar** threats:
  - hardware, media, software failures
  - natural disaster

- **more typical** threats:
  - economic failure
  - organizational failure
  - operator error
  - internal/external attack
who’s checking the checksums?

• fixity data **subject to same risks** as data whose integrity it assures

• it’s actually **more vulnerable** because:
  • more valuable
  • more centralized

• LOCKSS copies **tolerate multiple failures**, unlike canonical fixity store
canonical fixity store model
canonical fixity store model
canonical fixity store model
canonical fixity store model
canonical fixity store model
canonical fixity store model
canonical fixity store model
canonical fixity store model
canonical fixity store model
LOCKSS model
LOCKSS model
LOCKSS model
LOCKSS model
is this “better” digital preservation?
for sure

- 4+ LOCKSS copies technically better than 3x copies /w canonical fixity store
- increased confidence in data integrity consensus
- increased probability of good copy to repair from
- increased confidence in executing repairs
- better risk mitigation through decentralization
...or maybe not

- more copies means more cost
- greatest digital preservation threat is economic
- trade-off b/t level of preservation vs. amount of content preserved at given level
  - preserve less at higher preservation level
  - preserve more at lower preservation level
so, how to get the data integrity assurance provided by lots of copies without the cost of lots of copies?
blockchain!
?
not blockchain

• gap b/t hypothetical benefits vs. capabilities of real-world implementations

• recommended reading/viewing: DSHR, *Blockchain: What’s Not To Like?* from 2018 Fall CNI Meeting
LOCKSS fixity service

• make lots of copies…of checksums
• subject to LOCKSS polling + repair
• use consensus as indication of correct content checksum
• provide endpoint
• essentially, a canonical fixity store-like service, powered by LOCKSS

“Measure twice, cut once…” by GretaMichelle Joachim under CC BY 2.0
LOCKSS fixity service
LOCKSS fixity service
LOCKSS fixity service
LOCKSS fixity service
considerations

advantages

• high degree of data integrity assurance w/ few (content) copies
  • lower storage costs for high volume content
• provide data integrity assurance for content stored outside of LOCKSS system
• confidence in repair direction (or feasibility)

disadvantages

• fewer content copies to repair from
  • data integrity assurance useless if no remaining good copies
• low safety margin to detect + fix bad content copies
pilot use case

• **CLOCKSS Archive**
  - 12x-replicated LOCKSS network
  - long-term dark archive for the scholarly record

• supports:
  - scaling capacity
  - ensuring minimum level of preservation for more content
  - tiered appraisal

"Time #time #clock #johnlewis #poole" by Mark Hillary under CC BY 2.0
Question

Any Questions? by Matthias Ripp under CC BY 2.0