LOCKSS

The Cloudy Outlook for Digital Preservation

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

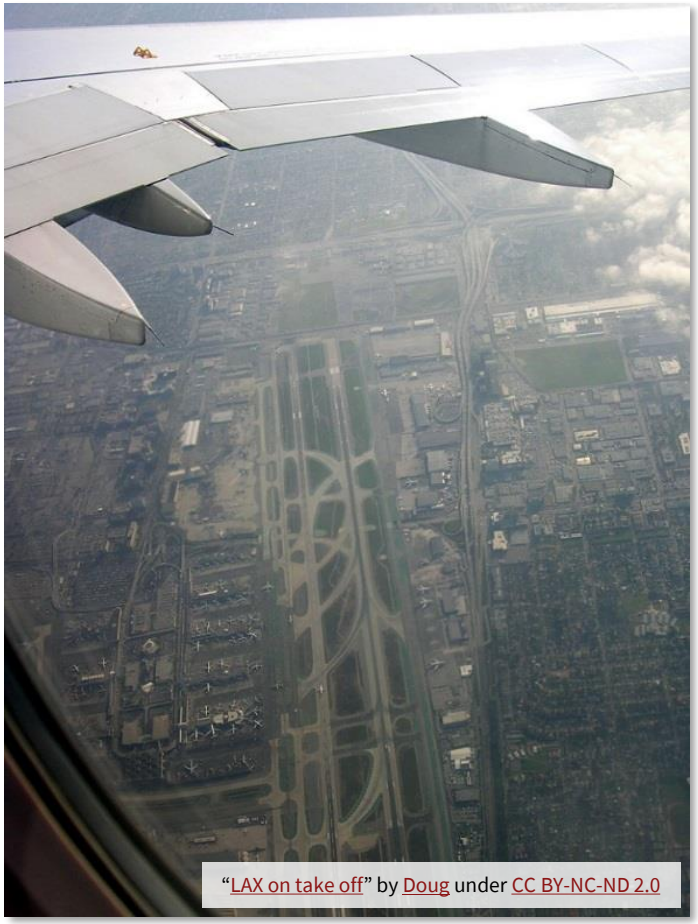
International Conference on Digital Preservation
Cloud Atlas: Navigating the Cloud for Digital Preservation
17 September 2019

“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

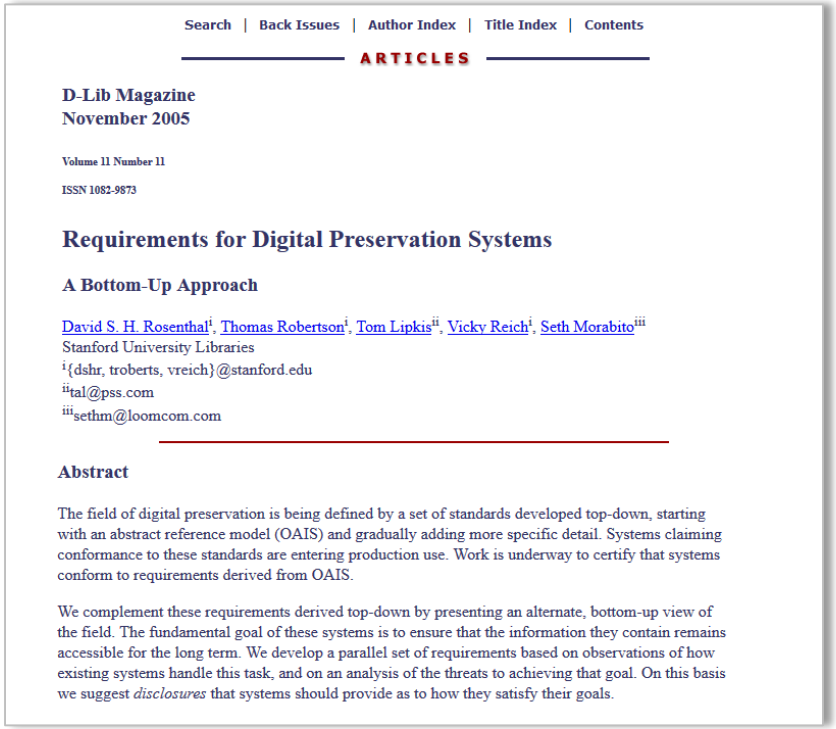


LOCKSS

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**



Search | Back Issues | Author Index | Title Index | Contents

ARTICLES

D-Lib Magazine
November 2005

Volume 11 Number 11

ISSN 1082-9873

Requirements for Digital Preservation Systems

A Bottom-Up Approach

David S. H. Rosenthal[i], Thomas Robertson[i], Tom Lipkis[ii], Vicky Reich[i], Seth Morabito[iii]
Stanford University Libraries
[i]{dshr, troberts, vreich}@stanford.edu
[ii]tal@pss.com
[iii]sethm@loomcom.com

Abstract

The field of digital preservation is being defined by a set of standards developed top-down, starting with an abstract reference model (OAIS) and gradually adding more specific detail. Systems claiming conformance to these standards are entering production use. Work is underway to certify that systems conform to requirements derived from OAIS.

We complement these requirements derived top-down by presenting an alternate, bottom-up view of the field. The fundamental goal of these systems is to ensure that the information they contain remains accessible for the long term. We develop a parallel set of requirements based on observations of how existing systems handle this task, and on an analysis of the threats to achieving that goal. On this basis we suggest *disclosures* that systems should provide as to how they satisfy their goals.

David S.H. Rosenthal et al: "Requirements for Digital Preservation Systems: A Bottom-Up Approach"

LOCKSS

community best practices

Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
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"

LOCKSS

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 1: Version 1 of the Levels of Digital Preservation

	Level 4 (Repair your data)
Storage and Geographic Location	- 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 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"

LOCKSS

what threats are discounted?

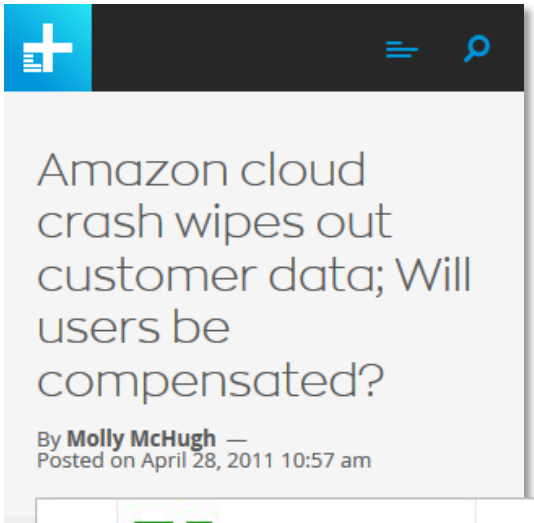
- **failures**: media, hardware, software, network services, economic, organizational
- **obsolescence**: media, hardware, software
- **errors**: communication, operator
- **attacks**: external, internal
- **natural disaster**

Table 1: Version 1 of the Levels of Digital Preservation

	Level 4 (Repair your data)
Storage and Geographic Location	- 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 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

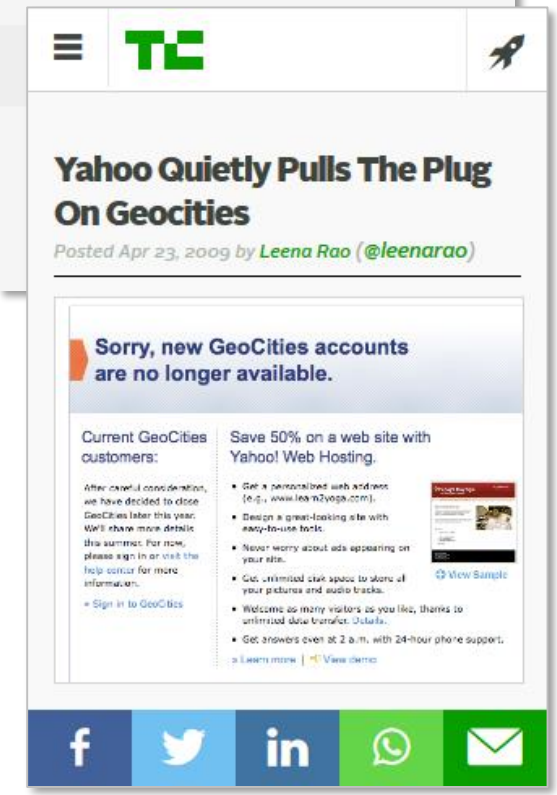
NDSA Levels of Preservation Working Group : "NDSA Levels of Digital Preservation"

LOCKSS



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)

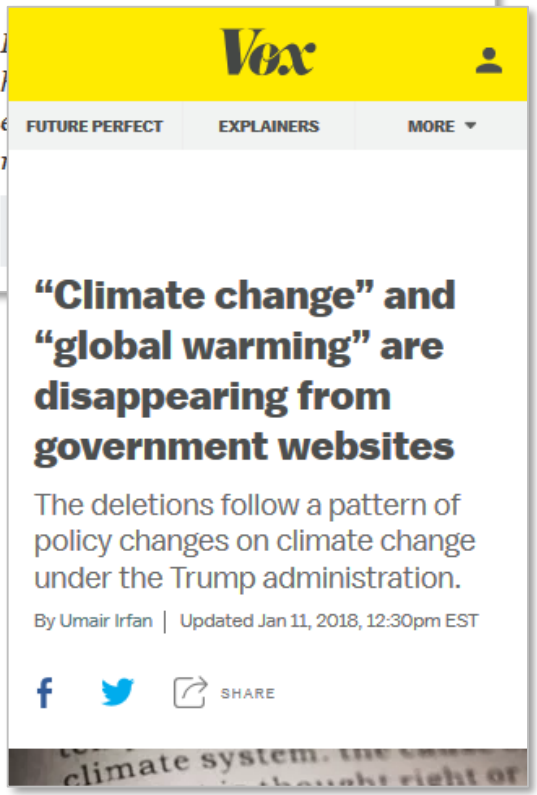
Sorry, new GeoCities accounts are no longer available.

Current GeoCities customers:

Save 50% on a web site with Yahoo! Web Hosting.

SECURITY

Ohio's Recent Spate of Cyberattacks Is Indicative of the National Trend



FUTURE PERFECT | EXPLAINERS | MORE

"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.

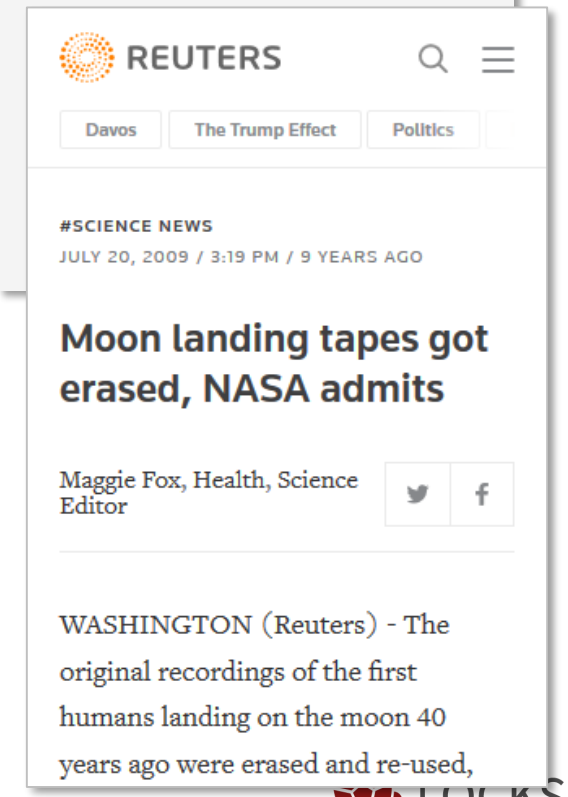
By Umair Irfan | Updated Jan 11, 2018, 12:30pm EST

SHARE

Angry Employee Deletes All of Company's Data

Published January 24, 2008
Fox News

Call it a tale of revenge gone wrong.



Davos | The Trump Effect | Politics

#SCIENCE NEWS
JULY 20, 2009 / 3:19 PM / 9 YEARS AGO

Moon landing tapes got erased, NASA admits

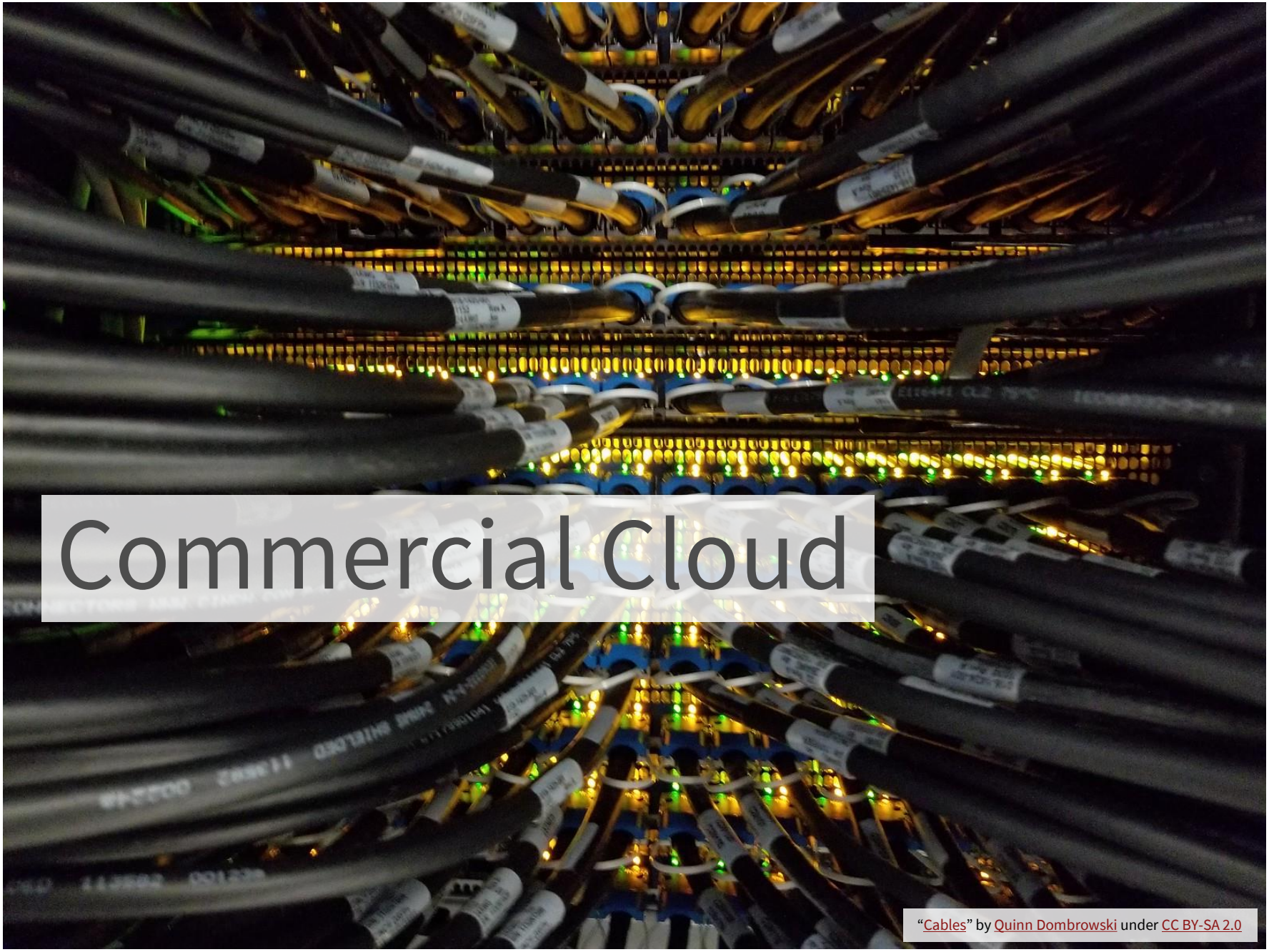
Maggie Fox, Health, Science Editor

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**

LOCKSS



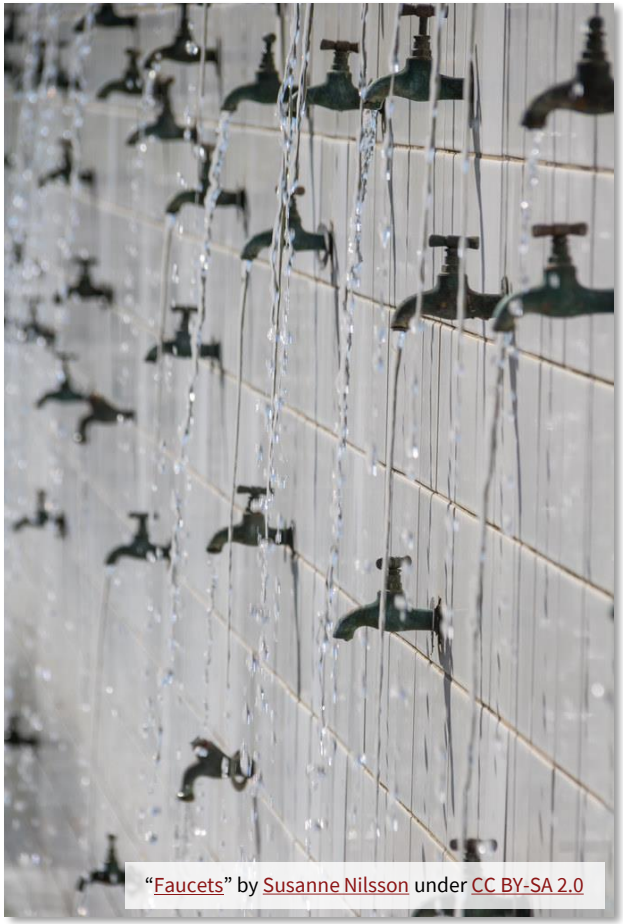
Commercial Cloud

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

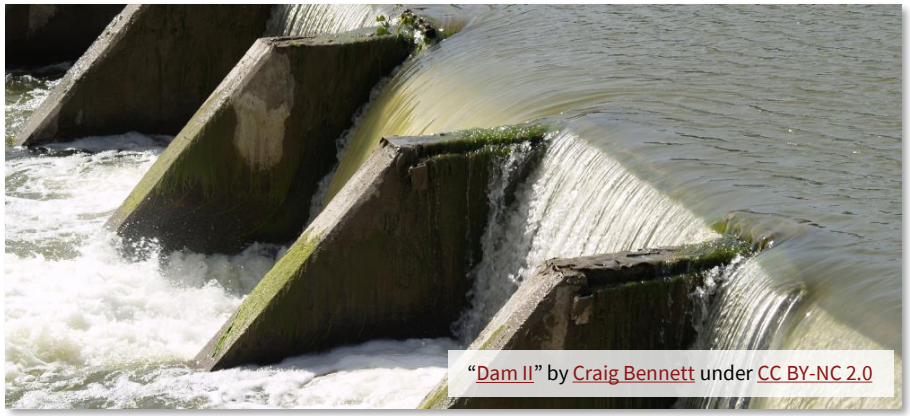


economic lock-in

- 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**

long-term storage services (1 PB, 1 month)				
service	**ingest**	**store**	**export**	**lock-in factor**
AMZN Glacier	$2,250	$4,000	$55,240	13.8x
GOOG Coldline	$3,600	$7,000	$83,860	12x
MSFT Archive	$6,350	$2,000	$16,260	8.1x

David S.H. Rosenthal : "Cloud for Preservation"



LOCKSS

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

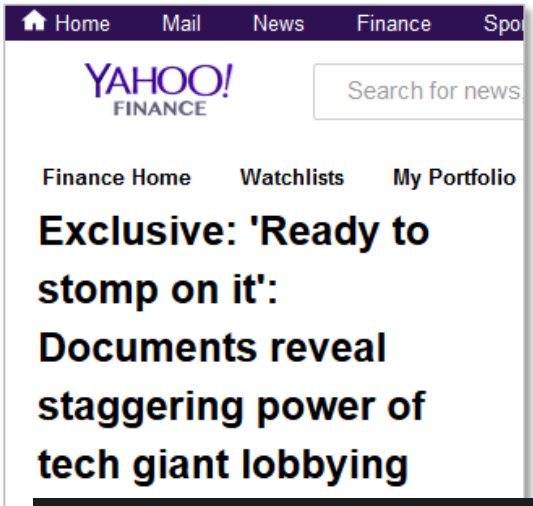
- **feature**, not a bug?
- hashing data in situ **requires trusting** that the service has performed computation rather than reporting cached value
- may be **prohibitively expensive** to retrieve content to a trusted environment to perform hashing

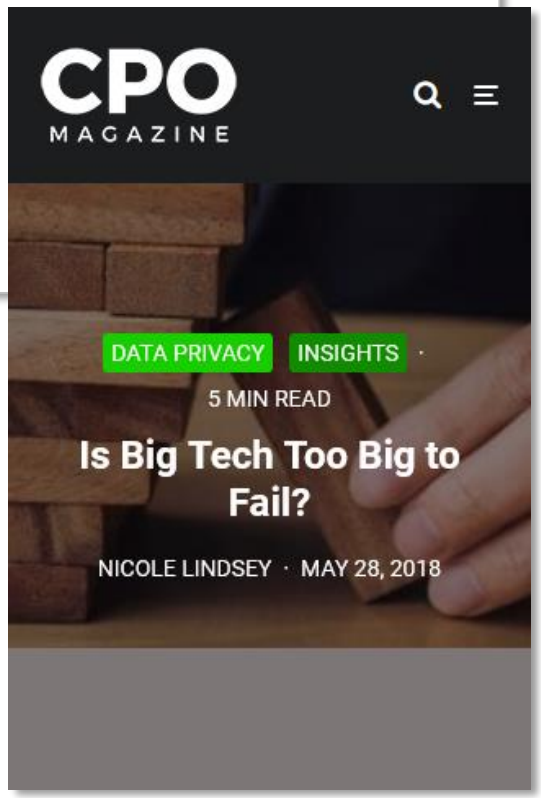
security configuration

- **monoculture** vulnerabilities
- greater affordances, better defaults for **on-premise security**
- consistent leaks from **misconfigured cloud services** suggest security is a challenge

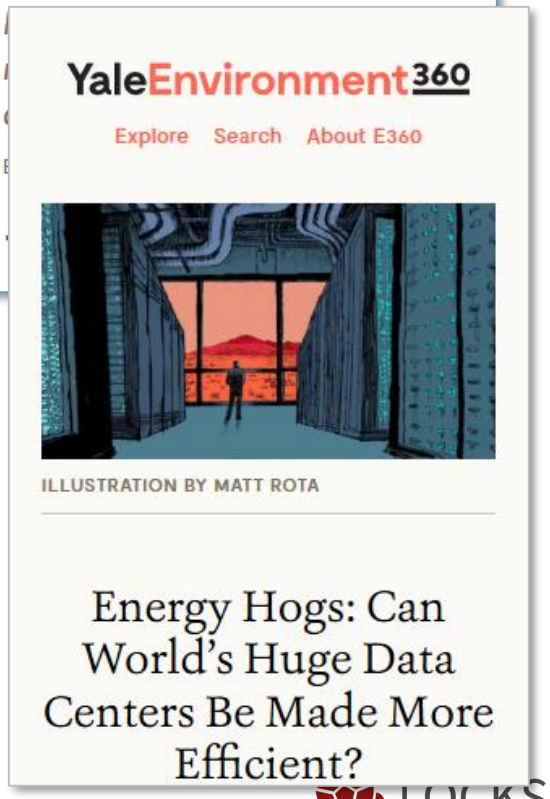

The Telegraph
Log in
News Politics Sport Business Mor
See all Tech
Technology Intelligence
Tech giants face tax avoidance crackdown
share
Save 3


GIZMODO
LATEST REVIEWS SCIENCE IO9 FIELD GU
GOODBYE BIG FIVE
I Cut the 'Big Five' Tech Giants From My Life. It Was Hell
Share
Tweet


Home Mail News Finance Spo
YAHOO! FINANCE
Search for news
Finance Home Watchlists My Portfolio
Exclusive: 'Ready to stomp on it': Documents reveal staggering power of tech giant lobbying


CPO MAGAZINE
DATA PRIVACY INSIGHTS
5 MIN READ
Is Big Tech Too Big to Fail?
NICOLE LINDSEY · MAY 28, 2018

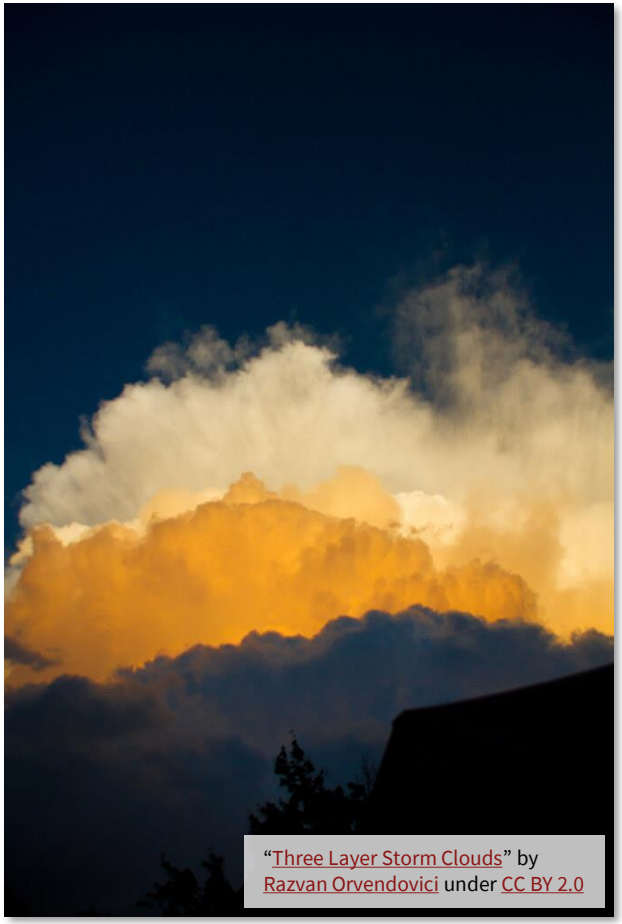

THE VERGE
TECH SCIENCE ENTERTAINMENT MORE
PODCASTS THE VERGECAST TECH
21
Big Tech's problem is its lack of competition


YaleEnvironment360
Explore Search About E360
ILLUSTRATION BY MATT ROTA
Energy Hogs: Can World's Huge Data Centers Be Made More Efficient?

Community Cloud

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



LOCKSS

community cloud considerations

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

LOCKSS

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

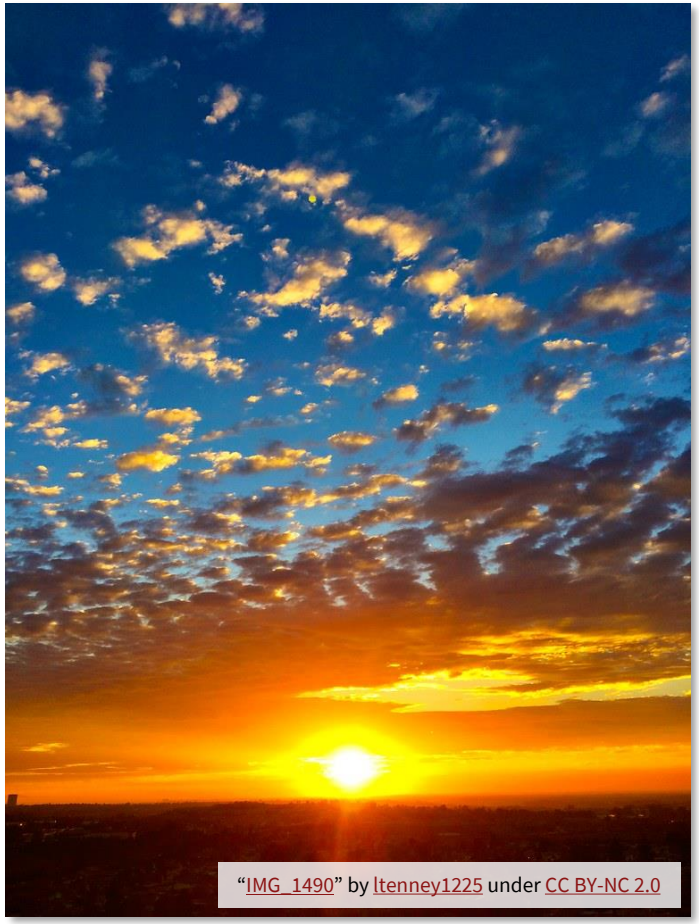
- in original, Global LOCKSS Network, **all nodes stored copies**
- private LOCKSS networks moving towards **hosted service models**
- **subset of institutions** host infrastructure w/ governance by + funding from broader community
- **Stanford + trusted partners** may serve as anchor storage hosts

Ivy Plus Libraries
Confederation

Wrap Up

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**



LOCKSS