NATIONAL AGENDA FOR DIGITAL STEWARDSHIP

2014

NDSA

Prepared for

CNI Fall Membership Meeting

December 2013

A National Agenda for Digital Stewardship

Presenters:

Micah Altman, <escience@mit.edu>, MIT
Michelle Gallinger, Library of Congress
Abigail Grotke, Library of Congress
Trevor Owens, Library of Congress

This Talk

Who are the NDSA?

Why develop an agenda for digital stewardship?

What should national priorities be?
… digital content
… technical infrastructure
… organizational roles
… research areas

What’s next?

NDSA

Collaborators & Co-Conspirators

- The 150+ institutional members of NDSA, and the 10000+ hours contributed by their representatives to NDSA working groups, meetings and reports

- National Agenda Authors:

  Micah Altman, Jefferson Bailey, Karen Cariani, Jim Corridan, Jonathan Crabtree, Blaine Dessy, Michelle Gallinger, Andrea Goethals, Abigail Grotke, Cathy Hartman, Butch Lazorchak, Jane Mandelbaum, Carol Minton Morris, Trevor Owens, Meg Phillips, John Spencer, Helen Tibbo, Tyler Walters, Kate Wittenberg, Kate Zwaard

Who are the NDSA?

About the NDSA

- **Founded** in 2010, the National Digital Stewardship Alliance (NDSA) is a consortium of institutions that are committed to the long-term preservation of digital information.
- **Our mission** is to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations.
- **NDSA member institutions** represent all sectors, and include universities, consortia, professional associations, commercial enterprises, and government agencies at the federal, state, and local levels. The Library of Congress provides organizational support and substantive collaboration as *Secretariat.*
- **Based on collaborative community effort** -- there are no fees for NDSA membership. Each member institution commits to to NDSA principles, and contributes efforts to working groups, reports, surveys, meetings and other NDSA initiatives.

NDSA Initiatives

Working Groups

Content

selection and aquisition of digital collections

Standards

digital formats and best practices

Infrastructure

development and maintenance of tools

Innovation

cutting edge research and development

Outreach

relationships and messaging

Recent Outputs

Extending Knowledge

- Preservation Storage Survey
- Web Harvesting Survey
- Preservation Staffing Survey
- Geospatial Selection & Appraisal report
- Content case studies
- NDSA Interview Series

Tools for Practice

- Levels of Preservation
- Digital Preservation in a Box
- Digital Preservation on Wikipedia

Dissemination

- National agenda for digital stewardship
- NDSA Innovation Awards
- NDSA Social Media

Why develop an agenda for digital stewardship?

Why a national agenda for digital stewardship?

- Effective digital stewardship is vital for:
  - maintaining authentic public records
  - growing a reliable scientific evidence base
  - providing durable access to our cultural heritage
- Knowledge of ongoing research, practice, and organizational collaborations is distributed widely across disciplines, sectors, and communities of practice

Why now?

Climate

Strong trends towards:

- More production of digital content
- More publishing, filtering and access
- More learners and collaborators
- More attention to public information

Weather

The New York Times

The Year of the MOOC

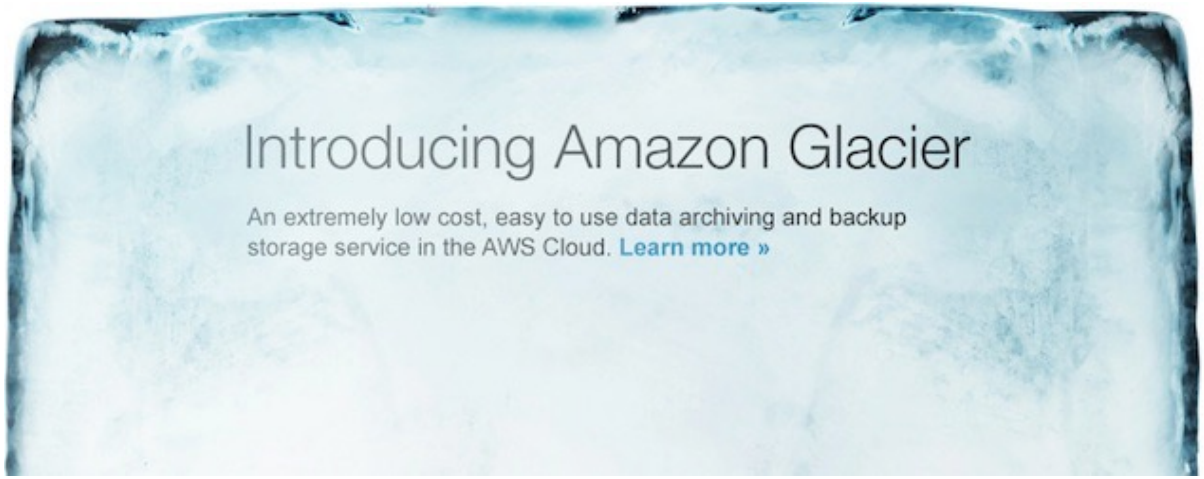
The Washington Post

White House moves to make federally funded research open to the public

THE CHRONICLE of Higher Education

Long-Awaited Ruling in Copyright Case Mostly Favors Georgia State U.

Isn’t digital preservation a solved problem?



- Why not put everything in Amazon?
- Amazon claims reliability of 99.999999999% (Longer odds than winning powerball, being struck by lightning, and finding alien life, combined)

What Was Accomplished?

The
National Agenda for Digital Stewardship
identifies high-impact opportunities to advance:

- *the state of the art*
- *the state of practice*
- *the state of collaboration*

How was this accomplished it?

- **Contributed community effort**

- *Development: contributions from the (now 150+) institutional members through working group participation, workshop discussion, commentary*
- *Writing: LC Staff, chairs of NDSA working groups, coordination committee*
- *Reviewing: expert reviewers in the preservation community*

- **Integrating diverse perspectives from multiple disciplines & sectors**

- **The persistence, organization, and commitment of the Library of Congress in its role as Secretariat**

National priorities for…

Digital Content

Digital Content Areas

- Web and Social Media
- Electronic Records
- Moving Image and Recorded Sound
- Research Data

Digital Content Areas

- Across all areas, content size, value and selection represent a core challenge

- Important to develop theoretically grounded and empirically tested models of information valuation.

Raising Awareness and Articulating Value

- Content area webinars
- Follow up blog posts, interviews
- *Content Case Studies*
- *Content Matters* blog posts

National priorities for…

Technical Infrastructure

2014 Technical Infrastructure Priorities

- File Format Action Plan Development
- Interoperability and Portability in Storage Architectures
- Integration of Digital Forensics Tools
- Ensuring Content Integrity

Technical Infrastructure

File Format Action Plan Development

–Stewardship organizations are amassing large collections of digital materials suggests a need to monitor the heterogeneous digital files the organizations are managing.

–Need for tools and services for creating file-format action plans is needed to make timely execution of file format plans a reality for data stewards.

–The digital preservation community would further benefit from organizations sharing their assessments of institutional risk and their plans for mitigating that risk and addressing file format problems with specific plans.

Interoperability and Portability in Storage Architectures

- As stewardship organizations manage increasingly large and complex data sets, the need for interoperability at various levels within the technical hardware and software stacks that make digital preservation becomes increasingly important.
- Interoperability of storage devices, hardware, data tape, and file systems software and would help alleviate bottlenecks in the interrelationship between distinct functions in workflows.
- Need for establishing and promoting technical means by which lower levels of the technology stack can directly integrate without requiring extensive computation and processing at higher levels.

Integration of Digital Forensics Tools

- Digital Forensics tools are essential for working across the range of heterogeneous kinds of digital materials coming under stewardship
- Projects like BitCurator are pulling together the suite of tools to do this work and developing processes and workflows.
- We are now at the point of implementation, it's time for organizations to start implementing and sharing information about their work
- The result of this work, will be large sets of heterogeneous digital files which will then push for the development of tools to work with these kinds of data at scale.

Ensuring Content Integrity

- Digital preservation is possible through a chain of migration of current hardware and software systems to yet-to-be-established future infrastructures. Essential to develop guidance on how to plan for and manage these changes
- Abstract requirements for fixity are useful as principals, but when applied universally can actually be detrimental in some digital preservation system architectures. Need for best practices for fixity in particular system designs and configurations.
- Need for the development of standards, practices and strategies that directly address migration, in particular, around end-to-end fixity checking

National priorities for…

Organizational Development

Organizational Roles, Policies, and Practices

Identifies need to increase cross-organizational cooperation to increase the impact and leverage investments made by individual institutions.

"People who work together will win, whether it be against complex football defenses, or the problems of modern society."

- NFL coach, Vince Lombardi

2014 Priorities for Cross-organizational Cooperation

1) *Provision networked preservation services*
   – network of preservation service providers with specialized services rather than every organization performing all aspects of digital preservation

2) *Collaborate on shepherding and promotion of standards*
   – digital preservation community representation on the relevant standards bodies rather than each organization needing to participate in every body

3) *Share digital preservation training and staffing resources*

National priorities for…

Research

Research Priorities

- Applied Research for Cost Modeling and Audit Modeling
- Understanding Information Equivalence & Significance
- Policy Research on Trust Frameworks
- Preservation at Scale
- The Evidence Base for Digital Preservation

What does the discipline believe?

- Our digital evidence base erodes
- There are multiple threats to information – diversifying against them is crucial
- Lifecycle analysis is critical for better long-term management of information
- Better practices are needed

How have we learned…
The Limits of Case Studies

Most current evidence for digital preservation practices and outcomes are based on local case studies and convenience samples

- Case studies are useful for:
  - existence proofs
  - raising awareness of problems
  - process tracing
  - hypothesis generation,
- Case studies are not enough to
  - advance our scientific knowledge
  - create robust predictive models
  - test causal hypotheses
  - strongly guide decision making.
- Systematic Evidence is needed both to support
  - general selection of digital preservation practices and method
  - applications of selected digital preservation methods in a specific operational context.

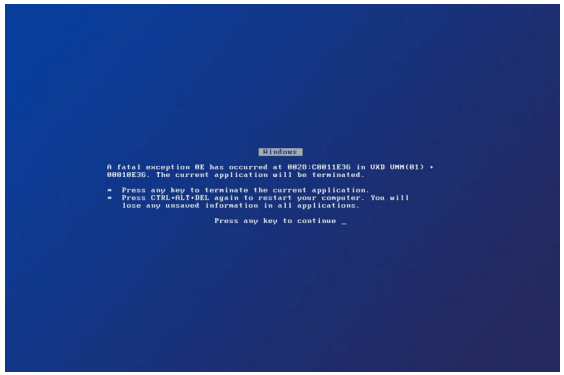
Simple question?

- If you have 1000 files (bitstreams), and you'd like to have 99.99% chance of accessing them in 20 years. How do you store them?

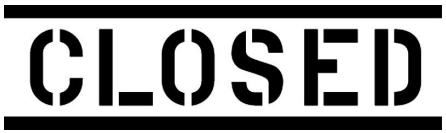
What are some threats?

Physical & Hardware

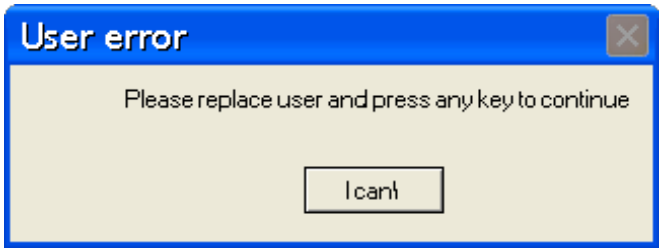
Insider & External Attacks



Software



Organizational Failure



Curatorial Error

Amazon's Unrealistic Nine Nines

- What are the units? - Collection? Object? Bit?
- How many of these do you have?
- Seems to be entirely theoretical
  - MBTF + Independence * enough replicas
  - No details for estimate provided
  - No historical reliability statistics provided
  - No service reliability auditing provided
- Reasons to Doubt Theoretical Calculations
  - Storage manufacture hardware MTBF (mean time between failures) is inaccurate…
  - Failures across hardware replicas are not independent
  - Many potential correlated failure modes not addressed:
    - software failure
      (e.g. a bug in the AWS software for its control backplane)
    - legal threats (leading to account lock-out — such as this, deletion, or content removal);
    - institutional threats (such as a change in Amazon's business model)
    - Process threats (someone hits the delete button by mistake; forgets to pay the bill; or AWS rejects the payment)
- Amazon SLA's do not incorporate or reflect "design" reliability claims even slighltly:
  - No claim to reliability in SLA's (or uptime, availability, response time…)
  - Sole recovery for breach is limited to refund of fees for periods the service was unavailable

The Problem Restated

Keeping risk of object loss fixed
-- what choices minimize $?

"Dual problem"

Keeping $ fixed, what choices minimize risk?

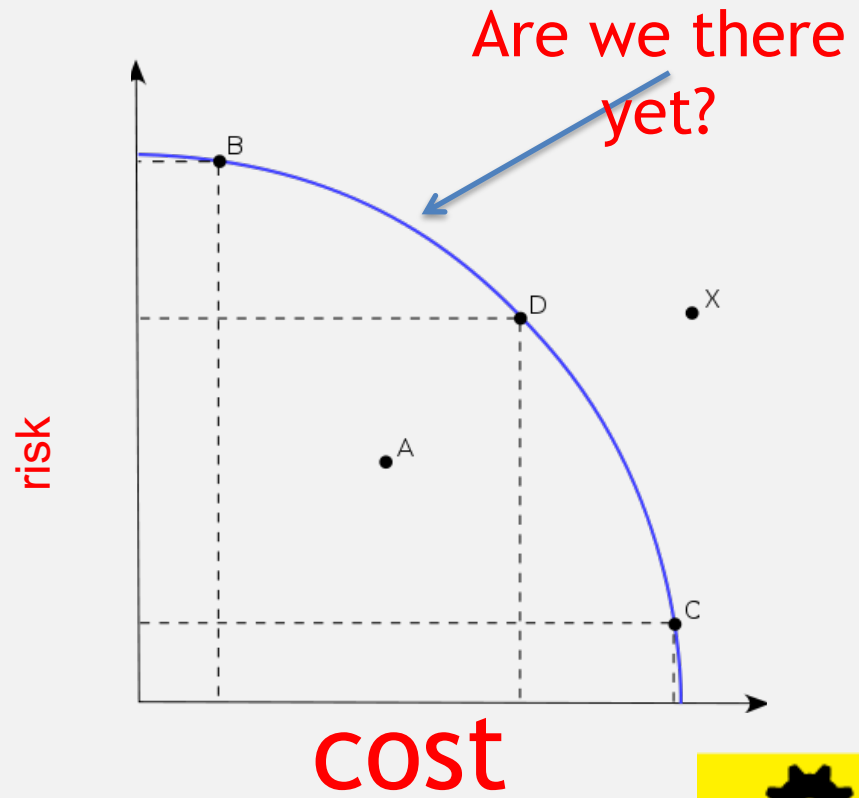
Extension

For specific cost functions for loss of object:

Loss(object_i), of all lost objects

What choices minimize:

Total cost= preservation cost+ sum(E(Loss))



NDSA

Methods for Mitigating Bit-LevelRisk

Local Storage

Physical: Media, Hardware, Environment

Formats

File Transforms: compression, encoding, encryption

File Systems: transforms, deduplication, redundancy

Replication

Diversification of copies

Number of copies

Verification

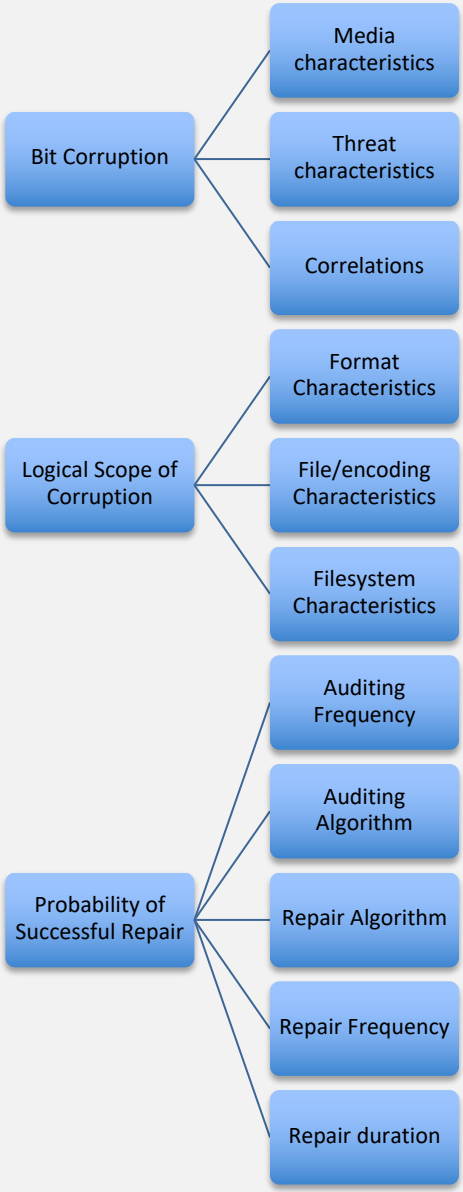
Fixity

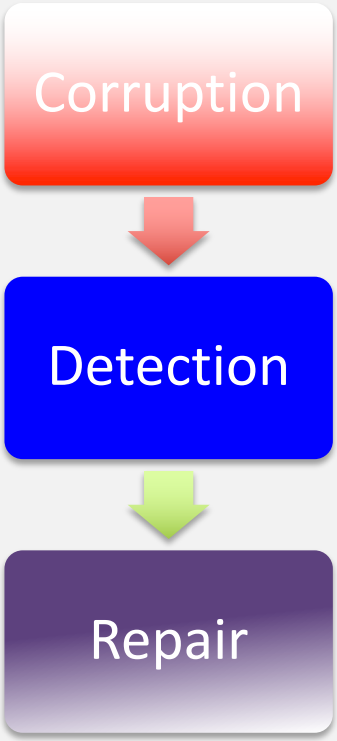
Audit

Repair

NDSA

Modeling


Bit Corruption
Media characteristics
Threat characteristics
Correlations
Logical Scope of Corruption
Format Characteristics
File/encoding Characteristics
Filesystem Characteristics
Probability of Successful Repair
Auditing Frequency
Auditing Algorithm
Repair Algorithm
Repair Frequency
Repair duration


Corruption
Detection
Repair

What Else do We Need To Know?

- What is the expected future value of a specified collection of digital content?
- What content is already being effectively stewarded by other organizations?
- How much is the expected future cost of preserving that content?
- How often do different threats to information manifest
  - storage hardware or media failures
  - software errors cause information loss
  - stored information becomes inaccessible because of obsolete formats, or loss of other contextual knowledge
  - that human error or maliciousness causes loss content in an information system
- What is the reliability of current digital preservation networks and services?
- How successful are other proposed strategies for replication, monitoring, certification, and auditing at preventing loss due to these threats?

How do we learn?

- Apply existing research methodologies from other fields -- especially fields involving observation research on humans and human systems
- Some useful methodologies:
  - probability-based surveys (e.g. of information management practice *and outcomes*)
  - replicable simulation experiments tied to theoretically grounded models of information management and risk;
  - creation of testbeds and test-corpuses which can be used to systematically compare new practices, tools, and methods;
  - field experiments, in which *randomized* interventions are applied and evaluated in real operational environments.

What’s next?

A National Preservation Agenda for 2015 and Beyond

- Drafts and update process starts this winter
- Community review process late spring
- An update will be presented in July at

Digital Preservation 2014

Moving Digital Preservation Forward

NDSA has a commitment to:

- Facilitating broad collaboration
- Promoting dissemination and engagement
- Regular updates and revisions of the National Agenda and core NDSA surveys

Want more information?

Contact NDSA for…

- Briefings, webinars, and consultations on the Agenda or other NDSA work
- Assistance in gathering comments on National policies and programs
- Assistance in recruiting experts for review and discussion panels; grant review
- Referrals to content stewards in specific areas

More Information

digitalpreservation.gov/ndsa/nationalagenda

ndsa@loc.gov