Prepared for
Digital Preservation 2012
Washington, DC
July 2012

Assessing and Mitigating Bit-Level Preservation Risks

NDSA Infrastructure Working Group

INTRODUCTION:

A Framework for Addressing Bit-Level Preservation Risk

Mark Evans
<mark.evans@tessella.com>
Digital Archiving Practice Manager,
Tessela Inc.

Micah Altman
<Micah_Altman@alumni.brown.edu>
Director of Research, MIT Libraries

NDSA

Threats to Bits

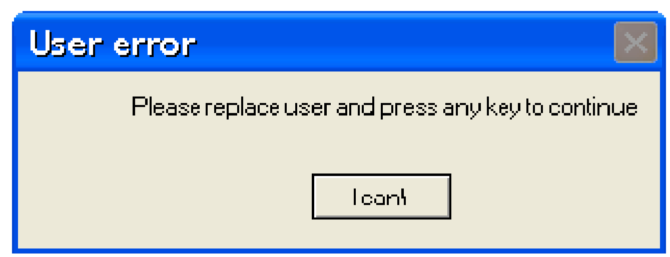
Physical & Hardware

Insider & External Attacks

Software

Media

Organizational Failure



Curatorial Error

Do you know where your data are?

How is content stored?

How is content Replicated?

How is content audited?

NDSA

Encoding

Priscilla Caplan
<pcaplan@ufl.edu>
Assistant Director for Digital Library Services,
Florida Virtual Campus

Compression

- Many types of compression:
  - Format based file compression, e.g. JPEG2000
  - Tape hardware compression at the drive
  - NAS compression via appliance or storage device
  - Data deduplication
- Is it lossless?
- Is it transparent?
- Is it proprietary?
- What is effect on error recovery?

Compression Tradeoffs

- Tradeoffs
  - Space savings allows more copies at same cost
  - But makes files more sensitive to data corruption
- Erasure coding in cloud storage
  - Massively more reliable
  - But dependent on proprietary index

Encryption

- Two contexts:
  – Archiving encrypted content
  – Archive encrypting content
- Reasons to encrypt:
  – Prevent unauthorized access
    - Especially in Cloud and on tape
  – To enforce DRM
  – Legal requirements (HIPAA, state law)
    - Though only required for transmission, not “at rest”

Encryption Concerns

- Increased file size
- Performance penalty
- Additional expense
- But makes files more sensitive to data corruption
- May complicate format migration
- May complicate legitimate access
- Risk of loss of encryption keys
- Difficulty of enterprise level key management
- Obsolescence of encryption formats
- Obsolescence of PKI infrastructure

Redundancy & Diversity

Andrea Goethals
<andrea_goethals@harvard.edu>
Manager of Digital Preservation and
Repository Services,
Harvard University

Failures WILL happen

- Real problem:
  failures you can't recover from!
- A few mitigating concepts:
  redundancy & diversity

Redundancy (multiple duplicates)

- Ecology
  - Redundancy hypothesis = species redundancy enhances ecosystem resiliency
- Digital preservation
  - Example: Multiple copies of content

Diversity (variations)

- Finance
  - Portfolio effect = diversification of assets stabilizes financial portfolios
- Ecology
  - Response diversity = diversification stabilizes ecosystem processes
- Digital preservation
  - Examples: different storage media, storage locations with different geographic threats

What can “fail”? What can’t?

- Likely candidates
  - Storage component faults
    - Latent sector errors (physical problems)
    - Silent data corruption (higher-level, usually SW problems)
    - Whole disks
  - Organizational disruptions (changes in finances, priorities, staffing)

1

Data loss risks **(impact & likelihood?)**	**Redundancy & diversity controls** **(costs?)**
Environmental factors e.g. temperature, vibrations affecting multiple devices in same data center	Replication to different data centers
Shared component faults e.g. power connections, cooling, SCSI controllers, software bugs	Replication to different data centers or redundant components, replication software systems
Large-scale disasters e.g. earthquakes	Replication to different geographic areas
Malicious attacks e.g. worms	Distinct security zones
Human error e.g. accidental deletions	Different administrative control
Organizational faults e.g. budget cuts	Different organizational control

NDSA

Slide 15

1 Andrea

Added software to the list of components... bugs in the software can also cause correlated failure.

- Micah

micah, 7/16/2012

BIT-LEVEL FIXITY

Karen Cariani
<karen_cariani@wgbh.org>
Director WGBH Media Library and Archives,
WGBH Educational Foundation

John Spencer
<jspencer@bmschase.com>
President, BMS/Chase LLC

Bit-Level Fixity

- Fixity is a “property” and a “process” (as defined from the 2008 PREMIS data dictionary)
- It is a “property”, where a message digest (usually referred to as a checksum) is created as a validation tool to ensure bit-level accuracy when migrating a digital file from one carrier to another

Bit-Level Fixity

- It is also a "process", in that fixity <u>must</u> be integrated into *every* digital preservation workflow
- Fixity is common in digital repositories, as it is easily put in the ingest and refresh migration cycles
- Fixity of digital files is a cornerstone of archival best practices

NDSA

So what’s the problem?

- While bit-level fixity solutions are readily available, there remains a large constituency of content creators that place minimal (or zero) value on this procedure
- Legacy IT environments, focused on business processes, are not “standards-driven”, more so by vendors, budgets, and poorly defined archival workflow strategies

NDSA

So what's the problem?

- A vast majority of commercial digital assets are stored "dark" (i.e. data tape or even worse, random HDDs), with <u>no</u> fixity strategy in place
- For private companies, individuals, and content creators with digital assets, bit-level fixity remains a mystery – a necessary outreach effort remains

So what’s the problem?

- Major labels, DIY artists, indie labels, amateur and semi-professional archivists, photographers, oral histories, and born-digital films usually ignore the concept of fixity
- All of the these constituencies need guidance to engage fixity into their daily workflow or suffer the consequences when the asset is needed NOW to monetize…

NDSA

Overview:

Auditing & Repair

Micah Altman
<Micah_Altman@alumni.brown.edu>
Director of Research, MIT Libraries

Audit [aw-dit]:

An independent evaluation of records and activities to assess a system of controls

Fixity mitigates risk *only if used for auditing*.

Functions of Storage Auditing

- **Detect**
  corruption/deletion of content
- **Verify**
  compliance with storage/replication policies
- **Prompt**
  repair actions

Bit-Level Audit Design Choices

- Audit regularity and coverage:
  on-demand (manually); on object access; on event; randomized sample; scheduled/comprehensive
- Fixity check & comparison algorithms
- Auditing scope:
  integrity of object; integrity of collection; integrity of network; policy compliance; public/transparent auditing
- Trust model
- Threat model

Repair

Auditing mitigates risk *only if used for repair*.

Design Elements

- Repair frequency
- Repair algorithm
- Repair duration

LOCKSS Auditing & Repair

Decentralized, peer-2-peer, tamper-resistant replication & repair

Regularity	Scheduled
Algorithms	Bespoke, peer-reviewed, tamper resistant
Scope	- Collection integrity - Collection repair
Trust model	- Publisher is canonical source of content - Changed contented treated as new - Replication peers are untrusted
Main threat models	- Media failure - Physical Failure - Curatorial Error - External Attack - Insider threats - Organizational failure
Key auditing limitations	- Correlated Software Failure - Lack of Policy Auditing, public/transparent auditing

DuraCloud Auditing & Repair

Storage replicated across cloud providers

Regularity	On-demand
Algorithms	Combination of bespoke algorithms and cloud provider
Scope	Object integrity only (no repair)
Trust model	- Content distributor (DuraCloud client) is completely trusted
Main threat models	- Media failure - Physical Failure
Key auditing limitations	- Limited range of threat models (e.g. software, curatorial failure). - Lack of scheduled auditing; collection integrity checks; policy auditing; repair.

NDSA

iRODS Auditing & Repair

Rules-based federated storage grid

Regularity	Scheduled, On-event
Algorithms	Bespoke, peer-reviewed
Scope	- Collection integrity - Collection repair - Micro-service policy auditing
Trust model	- Operator is implicitly trusted for content (by default) - More complex relationships possible through federation, microservices
Main threat models	- Media failure - Physical Failure - Policy implementation failure (auditing)
Key auditing limitations	- Limited range of threat models (e.g. software, curatorial failure) – some addressable through federation and microservices. = Lack of policy auditing, transparent/public auditing (by default)

NDSA

SafeArchive Auditing & Repair

TRAC-Aligned policy auditing as a overlay network

Regularity	Scheduled; Manual
Fixity algorithms	*Relies on underlying replication system*
Scope	- Collection integrity - Network integrity - Network repair - High-level (e.g. trac) policy auditing
Trust model	- External auditor, with permissions to collect meta-data/log information from replication network - Replication network is untrusted
Main threat models	- Software failure - Policy implementation failure (curatorial error; insider threat) - Organizational failure - *Media/physical failure through underlying replication system*
Key auditing limitations	Relies on underlying replication system, (now) LOCKSS, for fixity check and repair

Summary:

Micah Altman
<Micah_Altman@alumni.brown.edu>
Director of Research, MIT Libraries

Methods for Mitigating Risk

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

How can we choose?

- Clearly state decision problem
- Model connections between choices &outcomes
- Empirically calibrate and validate

The Problem

Keeping risk of object loss fixed

-- what choices minimize $?

“Dual problem”

Keeping $ fixed, what choices minimize risk?

Extension

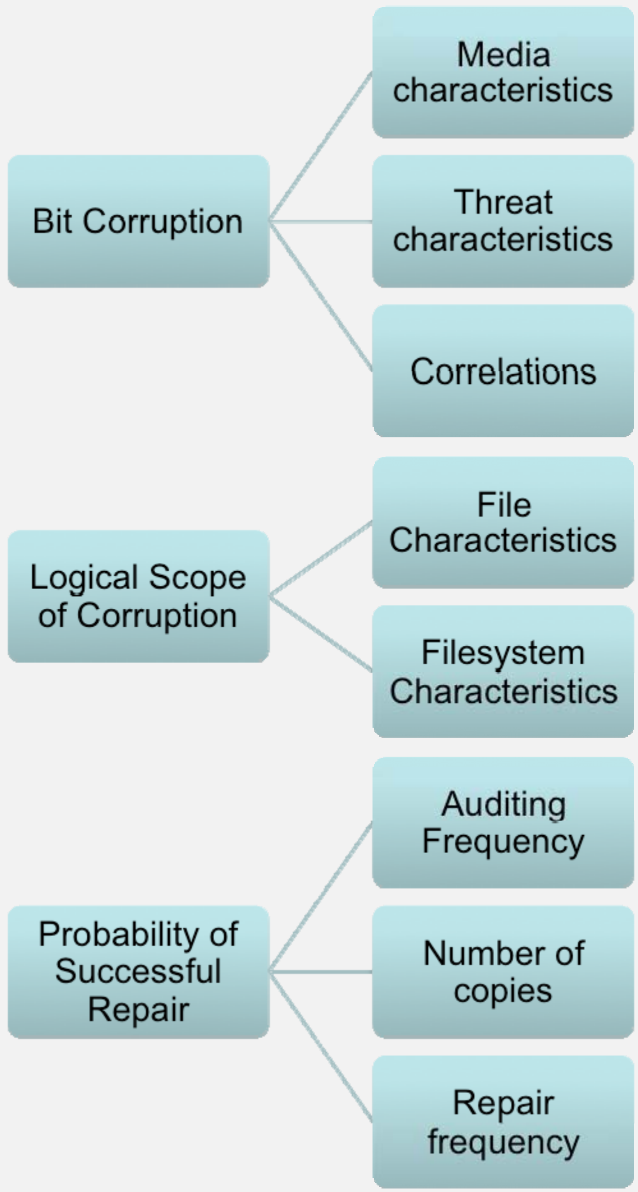
For specific cost functions for loss of object:

Loss(object_i), of all lost objects

What choices minimize:

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

Modeling



NDSA

Measurements

- Media – MBTF theoretical and actual
- File transformations:
  - compression ratio
  - partial recoverability
- Filesystem transformations:
  - Deduplication
  - Compression ratio
- Diversification
  - Single points of failure
  - Correlated failures
- Copies, Audit, Repair
  - Simulation models
  - Audit studies

NDSA

Questions*

What techniques are you using?

What models guide the "knobs"?

Contact the NDSA Infrastructure Working Group:

www.digitalpreservation.gov/ndsa/working_groups/

** Thanks to our moderator*:

Trevor Owens <trow@loc.gov>,Digital Archivist, Library of Congress