

Introduction to Digital Preservation

PASIG Mexico City

February 12, 2019

David Minor, Program Director for
Research Data Curation
UC San Diego Library

Tom Cramer, Chief Technology
Strategist
Stanford University Libraries



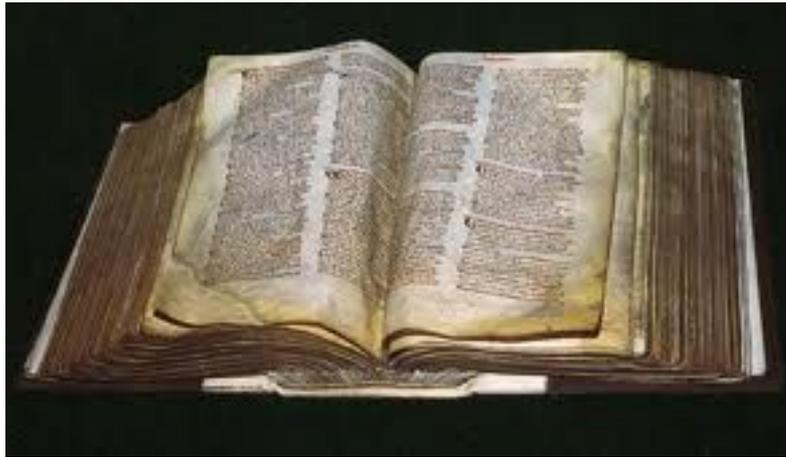
Agenda

- Introduction
 - Traditional vs. Digital Preservation
 - “Examples” of Loss
- Risks & Strategies
- Key Concepts:
 - OAIS
 - Authenticity
 - Trust
 - Sustainability
- Summary
- Resources



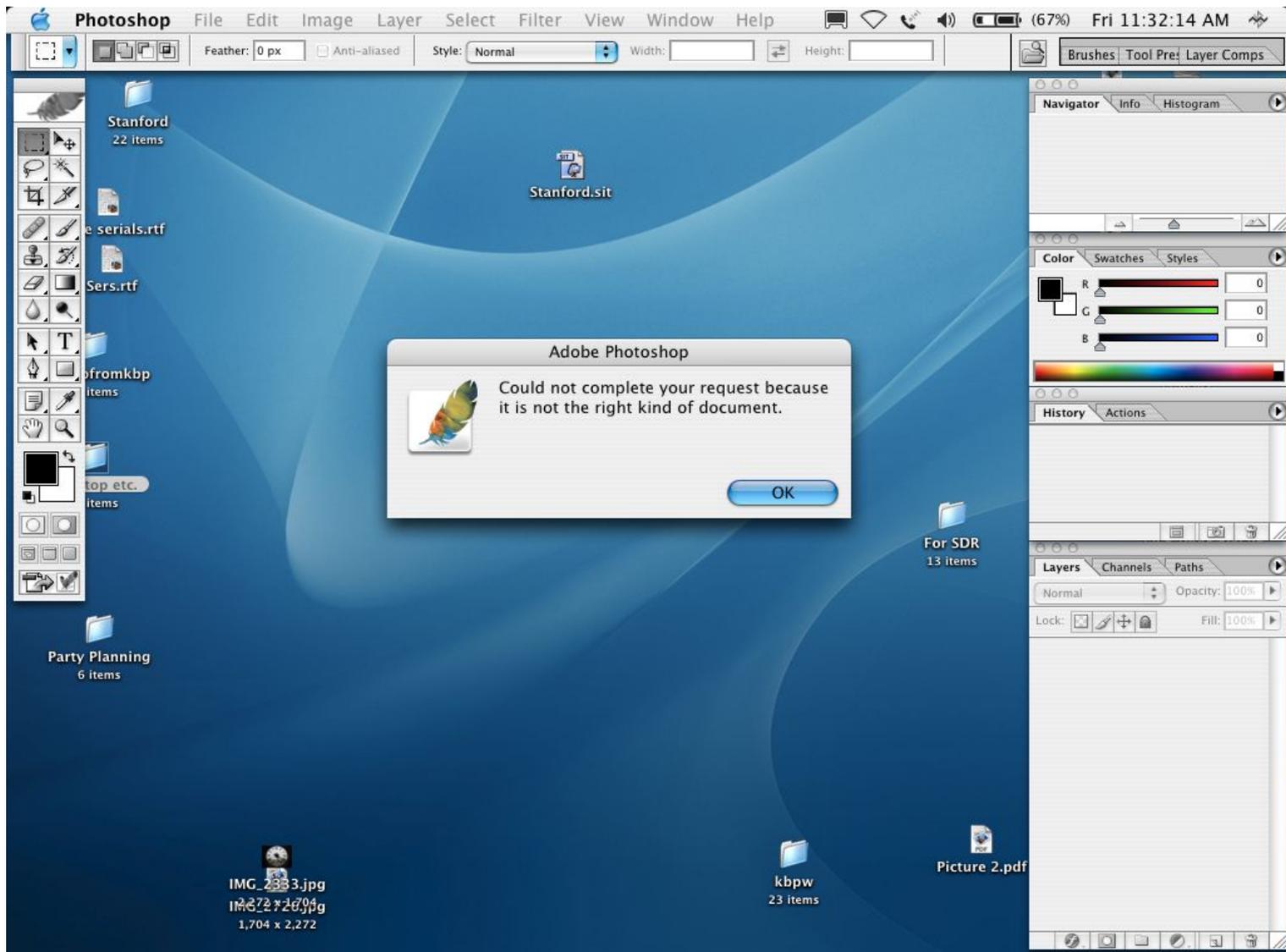
Bits Decay: Do Something Today
30th November 2017

The case of two formats...



Which of these will be usable in
100 years?





What Is Digital Information?

- Data that is encoded as a series of 0's and 1's...
- typically stored on magnetic (disk, tape) or optical (CD's, DVD's) media, that...
- requires specialized hardware to read (e.g., disk or tape drive, CD player), and...
- specialized software to operate (e.g., firmware, operating systems), and...
- still more software (applications) to interpret and render (e.g., Powerpoint) into usable form, and...
- contextual knowledge about how to operate that software in order to use it (e.g., *double click to open*)

What Is the Problem?

In short, access to digital information requires hardware, software and people...

But technology and people change, therefore creating potential barriers to re-use.



What Is Digital Preservation?

Digital preservation is the strategy and actions taken to promote the availability and usability of digital information over time.

The mission of the Preservation Directorate at the Library of Congress is to assure **long-term, uninterrupted access** to the **intellectual content** of the Library's collections.

<http://www.loc.gov/preservation/about/org.html>



(In)Famous Example of Loss

- The 1976 Viking Mars Landings
- Raw data still available, but...
 - Some not processed;
 - Some not documented;
 - Original software defunct.



- 1988: Processing 3,000 images from original tapes took 2 years of reverse engineering to build modern software.
- 2003: Researchers looking for biological data from the landing couldn't process the raw data in unknown formats. Tracked down printouts and hired students to re-key it all.

Less Well-Known Example of Loss

- **Distribute IT**, a web-hosting business in Australia - gone in 30 minutes

“...not only was the production data erased during the attack, but also key backups, snapshots and other information...” the company said in its final blog post.

The Register

June 11, 2011

(Aired on ABC May 20, 2013)

94% of companies suffering from a catastrophic data loss do not survive. 43% never reopen and 51% close within two years.

U. Texas Study

Risks to Digital Information

- Media decay (bit rot)
- Obsolescence
 - File Format
 - Software
 - Hardware
 - Media
- Technology Failure
 - Software
 - Hardware
 - Media
- Communication errors
- Lack of context
 - Data but no codebook
- Ambiguous IP State
 - Copyright
 - Licensing
- Natural disasters
- Information attack
- Economic failure
- Organizational failure
- Loss of will
- Human error

Strategies

- Replication
- Migration
 - Format
 - Media
 - Technology
- Emulation
 - Hardware
 - Software
- Encapsulation
- Redundancy and Heterogeneity
 - Technology
 - Location
 - Organization
- Succession Planning

Digital Preservation (aka Long-term Access)
is realized through a series of relays over time.

Digital Preservation is More Than Technology

10 Core Requirements for Digital Archives Center
for Research Libraries (CRL) / Trustworthy Repositories
Audit and Checklist (TRAC)

1. Mandate and Commitment to Digital Object Maintenance
2. Organizational Fitness
3. Legal and Regulatory Fitness
4. Efficient & Effective Policies
5. Adequate Technical Infrastructure

More Than Just Technology (continued)

6. Acquisition and Ingest
7. Preservation of Digital Object Integrity, Authenticity, & Usability
8. Metadata Management & Audit Trails
9. Dissemination
10. Preservation Planning and Action

Preservation & Archiving vs...

- Backup
- Disaster Recovery / Business Continuity
- Enterprise Content Management Systems
 - Document, Records, Web, Email
- Digital Asset Management
 - Images, Audio, Video
- Hierarchical Storage Management (HSM)



Key Concepts

- OAIS
- Authenticity
- Trust
- Sustainability



OAIS: Key Concepts & Definitions (ISO 14721)

- *Open* = developed in an open public forum
- *Archival Information System*: “an organization of people and systems that has accepted the responsibility to
 - preserve information and
 - make it available for a
 - Designated Community”

OAIS: Mandatory Responsibilities

- Accept content
- Obtain control (including necessary IP rights)
- Define user community
- Ensure that the preserved information is **independently understandable** to the user community
- Follow documented procedures to
 - **Preserve information against reasonable contingencies**
 - Enable dissemination of **authenticated copies**
- Make preserved information available

Authenticity

- Authenticity (traditional & digital) derives from:
 - Source
 - Chain of custody
 - Processing history
 - Fixity
 - Trust



Maintaining and disseminating authentic information is a primary mission for digital preservation systems.

Trust & Trustworthiness

- Trust is granted by a third party to a repository
- *Trustworthiness* is demonstrated by adherence to four principles:
 1. Documentation
 2. Transparency
 3. Adequacy
 4. Measurability
- Audits help establish Trustworthiness



Sustainability

- Long-term preservation, by definition, requires management of information over generations of change in;
 - Technology
 - Users & expectations
 - Staffing
 - Economic conditions
- Preservation is a journey, not a destination



Summary

- Minimize dependencies
 - Encapsulate your metadata with your objects
- Minimize correlated errors
 - Embrace redundancy
 - Embrace diversity
- Monolithic systems tend to serve poorly
 - Complex, expensive, inflexible
 - Migration costs can capsize you
- Keep it simple; have an exit plan for every component

Digital Preservation Resources

- Preservation and Archiving Special Interest Group (PASIG)
<http://www.preservationandarchivingsig.org/>
- Library of Congress (LoC) Designing Storage Architectures
Annual Meetings
(<http://www.digitalpreservation.gov/meetings/storage17.html>)
- Digital Preservation Coalition (DPC) <https://www.dpconline.org/>
- National Digital Stewardship Alliance (NDSA) <http://ndsa.org/>
- LOCKSS Mail List and Newsletter
www.lockss.org, artpasquinelli@stanford.edu