

Digital Preservation and Scholarly Communications

Oxford Festival of Open Scholarship 2023

March 9th 2023 – 14:00-15:30

Neil Jefferies

Head of Innovation

Open Scholarship Support

<https://orcid.org/0000-0003-3311-3741>



This session will be recorded



Slides will be made available after the session

Join us for more events at: <http://openaccess.ox.ac.uk/oxfos-23>



If you would like to tweet about this session please use: #OxFOS23

Session Structure

- Degradation of Resources
- Mitigation Strategies
- Relevance to Research
- The Importance of Standards

Degradation of Resources

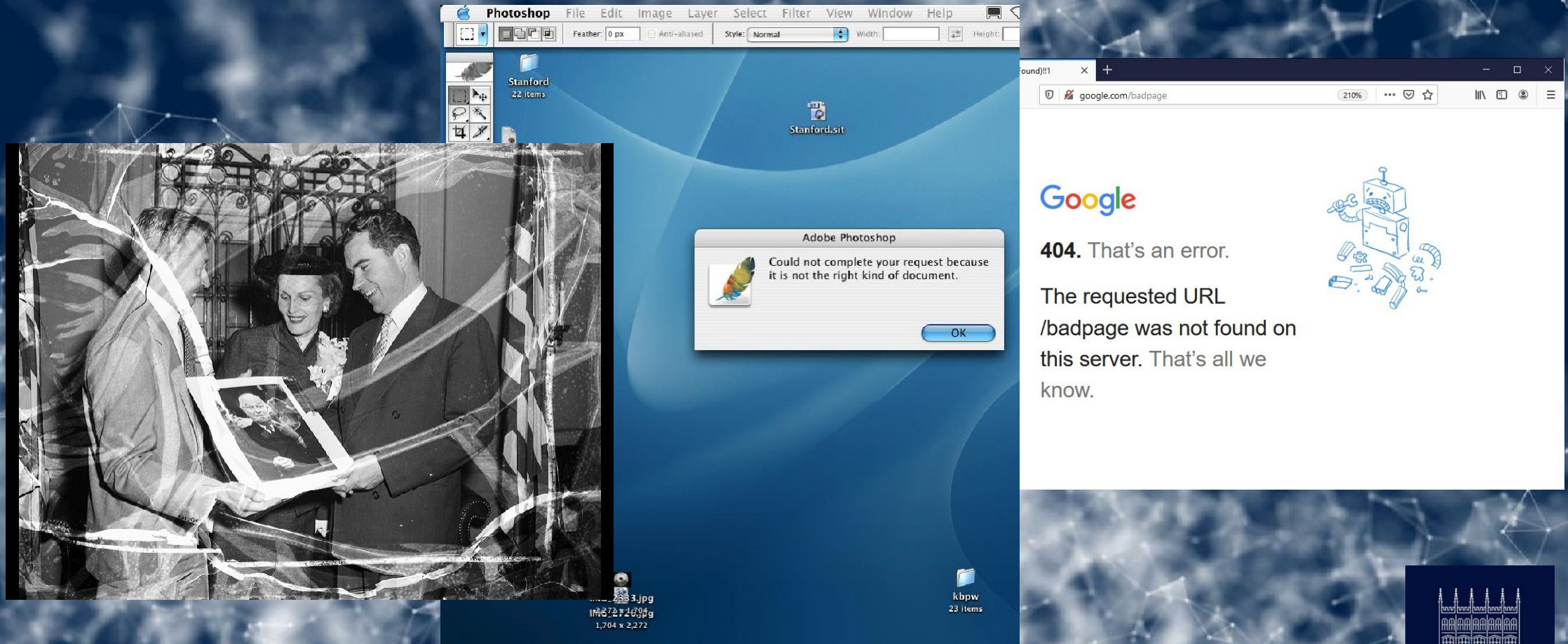
“Traditional” Preservation

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.

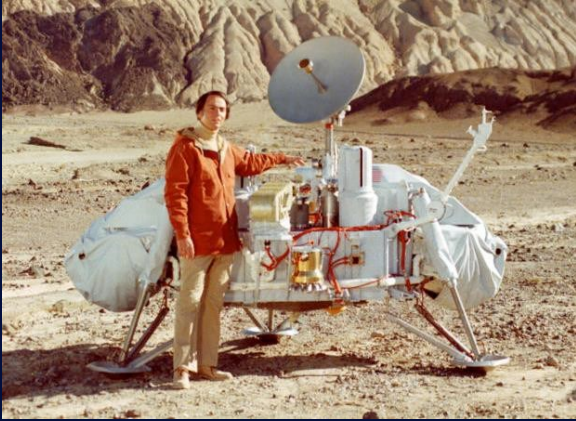
<http://www.loc.gov/preservation/about/org.html>

- Effective Strategies for Traditional Materials
 - Materials science: stable media
 - Physical Protection
 - Stable environmental conditions
 - Controlled use
 - Conservation, Repair and Reformatting

Degradation behaviour...



Examples...



- Viking Mars landings 1976
- 1988: Extracting 3000 images from tapes took 2 years of reverse engineering effort
- 2003: Biological researchers couldn't read the raw data format. Students rekeyed from printouts.

- The 1986 BBC Domesday project
- ~1M participants submitted images, maps, video, statistics and stories
- Two 300MB per side 12" Laser-discs, requires a Philips VP415 "Domesday Player"
- 1999-2003 – Two recovery attempts using emulation and migration
- 2090 – Final date copyright expires



Quantifying the Risks

- Bit rot (not really)
 - Media decay
 - Corruption
 - Transmission errors
- Obsolescence
 - File Format
 - Software
 - Hardware
 - Media
- Technology Failure
 - Software
 - Hardware
 - Media
- Loss of context
 - Data but no codebook
- Ambiguous IP State
 - Copyright
 - Licensing
- Disasters
 - Natural disasters
 - War
- Organizational failure
 - Loss of will
 - Human error
 - Sabotage
 - Economic failure

Mitigation Strategies

Sooner or later digital resources end up in a library or similar institution...or they disappear (me)

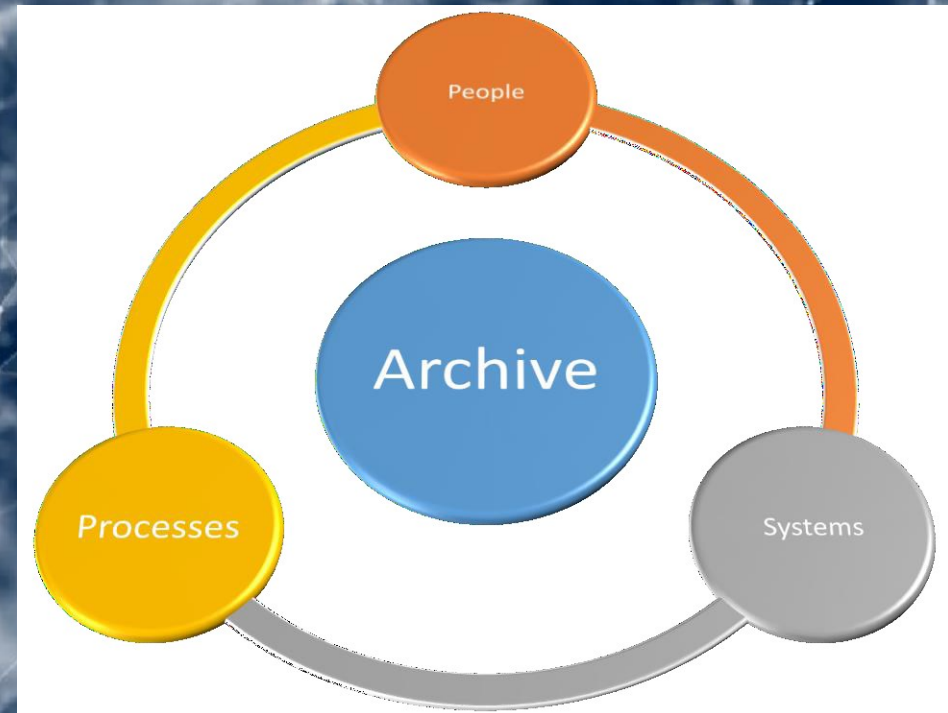
High Level Strategies

- Redundancy
 - Replication
 - “Never get down to your last 2 copies”
 - Heterogeneity
 - Location
 - Organisation
- Emulation
 - Hardware
 - “If some one wrote a game for a platform, then someone will write and emulator for it”
 - Software
 - “Appliance”
- Encapsulation
 - I prefer “Locality”
- Succession Planning
 - People
 - Organisations
 - Technical
 - File formats
 - Media
 - Hardware
 - Software
- Capability

Digital Preservation (long-term access) is realized as a series of relays over time.

The Digital Preservation Environment

- Ongoing access to digital information involves...
 - Systems – Hardware & Software
 - People & organisations
 - Processes & procedures
- All of which are subject to obsolescence and change



Being Realistic...



- Memory organisations are much smaller and poorly resourced than large corporations
 - Harder to take the long view
 - But tend to be much longer lived
- The economic justification for preservation is **long term access**
 - Hard to articulate if most of your users aren't born yet
 - And you have no idea what their requirements are...

TRAC Requirements for Digital Archives

- https://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf

- Mandate and Commitment to Digital Object Maintenance
- Organizational Fitness
- Legal and Regulatory Fitness
- Efficient & Effective Policies
- Adequate Technical Infrastructure

- Acquisition and Ingest
- Preservation of Digital Object Integrity, Authenticity & Usability
- Metadata Management & Audit Trails
- Dissemination
- Preservation Planning and Action

Relevance to Research

More specifically, Scholarly Communication

Role of Scholarly Outputs

- Scholarly Communications...

- Reaching peers and public
- Engaging in discourse
- Generating impact
- Discoverability and accessibility
- Reproducibility
- Re-usability
- Open-ness
- Ethical
- Timeliness
- Citation and Referencing

- Research Administration

- Funder and organisational policies
 - Funder and organisational metrics
 - Reporting
-
- FAIR – Findable, Accessible, Interoperable, Reusable
(<https://www.go-fair.org/go-fair-initiative/>)
 - CARE – Collective benefit, Authority to control, Responsibility, Ethics
(<https://www.gida-global.org/care>)

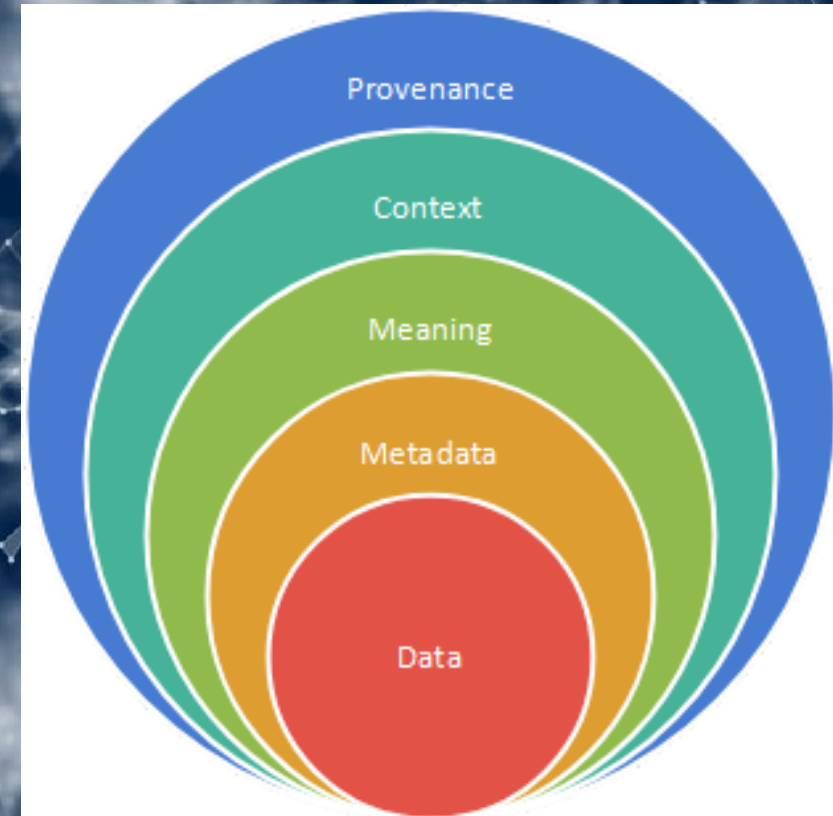
Diversity of Scholarly Outputs

- Analogues of physical outputs are augmented by:
 - Images and multimedia
 - Data
 - Software
 - Visualisations
 - Tagging
 - Annotations
 - Websites
 - Social media
 - etc.
- Scholarly Communications is now a multi-channel endeavour
- To meet their broader goals these all need to be (to a lesser or greater extent)...
 - **Persistent**
 - Discoverable and Understandable
 - **Accessible/usable**
 - Referenceable/citable
 - Versioned/timestamped
- ...over the long term (especially Humanities)
- ...not automatically the case for a stream of bytes somewhere on the internet

Digital Preservation (as discussed) only really addresses Persistence and Accessibility

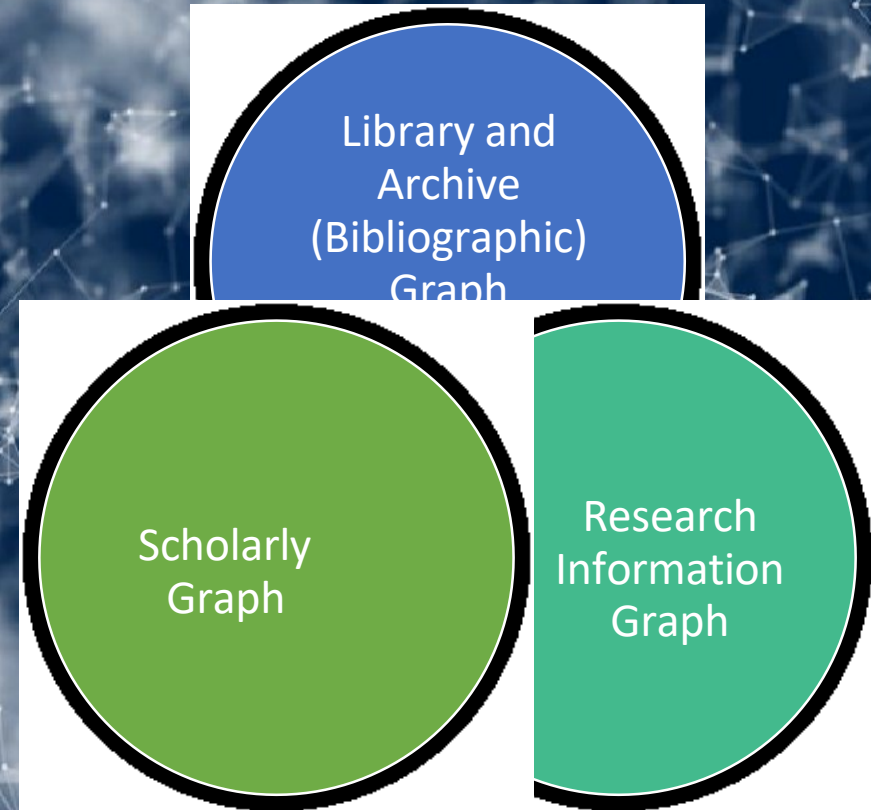
Discoverable and Understandable

- An artefact derives much of its meaning from attributes that are not intrinsic to the artefact itself
 - Context - the circumstances under which it was created
 - Provenance – how it came to be as it is now
- This is especially true for digital materials...
 - A file is a meaningless stream of bytes
 - Filenames can be readily changed
- The “metadata” can have more meaning than the “data”
 - Can we even unambiguously define metadata?
- A (digital) object should be considered a greater whole comprising several streams of information that can be arbitrarily labelled data or metadata but all of which contribute to the intellectual content of the object



The “Locality” problem

- Scholarly Outputs sit at the junction of three Knowledge Graphs (areas)
 - Each with their own standards
 - Generally requires linking...
- Libraries and Archives – catalogues of published materials and sources
- Research Information - funders and institutional administration
- Scholarly - scholarly process, taxonomies and practice (often domain specific)



Citability and Versioning

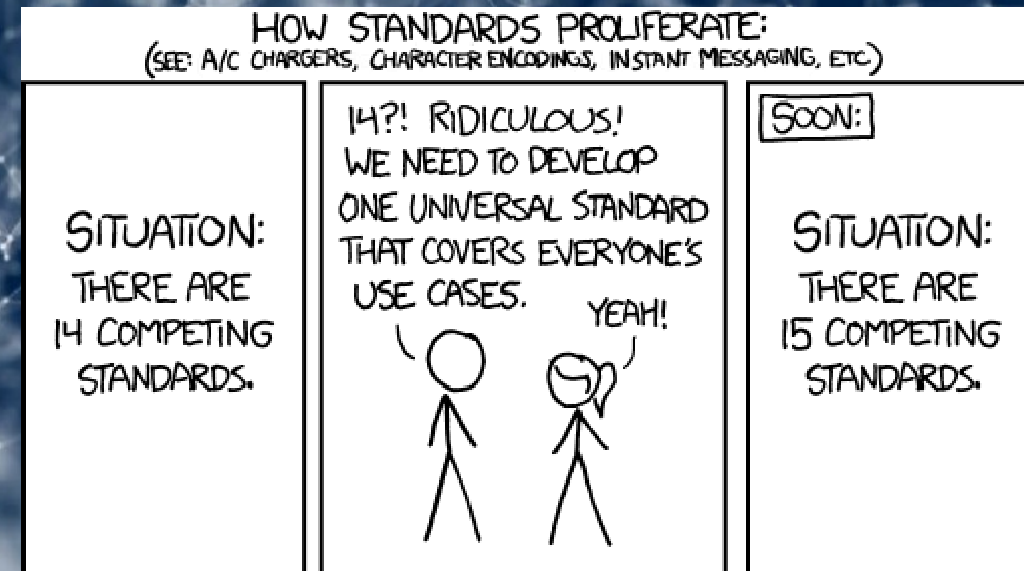
- Digital objects can and will be copied, updated and moved around
 - **Identifiers** provide a consistent way of referencing/citing them (avoid the dreaded 404 error)
 - A **resolver service** translates an identifier into information about the location of the object
 - **Expectation that they are persistent (PID's)**
- Generally, objects are collected in repositories maintained by memory organisations
 - Submission and acceptance to a repository generally results in the generation of a suitable identifier
- Some examples...*there should not be a real need to create any new systems*
 - DOI (Digital Object Identifier) for papers/articles
 - DataCite DOI for datasets
 - ORCID's for researchers/authors/creators (and ISNI for historical individuals
 - OrgID for organisations (institutions, companies, funders etc.)
 - RAID for research activities
- Versioning is important for **referential integrity** when digital objects can change

Standards

If you want to go fast, go alone. If you want to go far,
go together. (provenance unclear, poss. African
<https://andrewwhitby.com/2020/12/25/if-you-want-to-go-fast/>)

Standards

- “The wonderful thing about standards is that there are so many of them” - Tannenbaum/Hopper/Olsen/Seybold
 - Question standards not used by a significant community
 - Question standards if implementations cannot actually interoperate
- Perverse incentives – doing something different gets you published
 - Short term gain – this limits accessibility and reuse...
- Investigating existing standards for data and metadata should go alongside a literature search prior to research
 - Sustainability section of a data management plan



Permanent link to this comic: <https://xkcd.com/927/>

Technologies are Not Objective

- They embody the priorities, aims and biases of their creators. For example:
 - Preparing material for others to use is hard
 - Still a mindset around “final format” publication
- (Domain) specific vs generic approaches
 - Do one thing well or many things less well
 - Standards often suffer from “scope-creep”
- Unstructured vs structured
 - Documents, narratives and texts vs tabular datasets
- Flexibility
 - Do you expect the material to expand or evolve?
- One size does not fit all!
 - You may (should?) have to use more than one



A Special Note on Software

- Software is hard to develop and preserve well
 - What is it meant to do? (Specification)
 - Can we prove it is working correctly? (Tests)
 - How do we keep it working? (Preservation)
- In 2016, this happened...
 - “There could be a very serious problem with the past 15 years of research into human brain activity, with a new study suggesting that a bug in fMRI software could invalidate the results of some 40,000 papers.”
- Always sanity check your output
 - Alternative code you can use to cross check?
 - Is there a specification/test set?



Tips

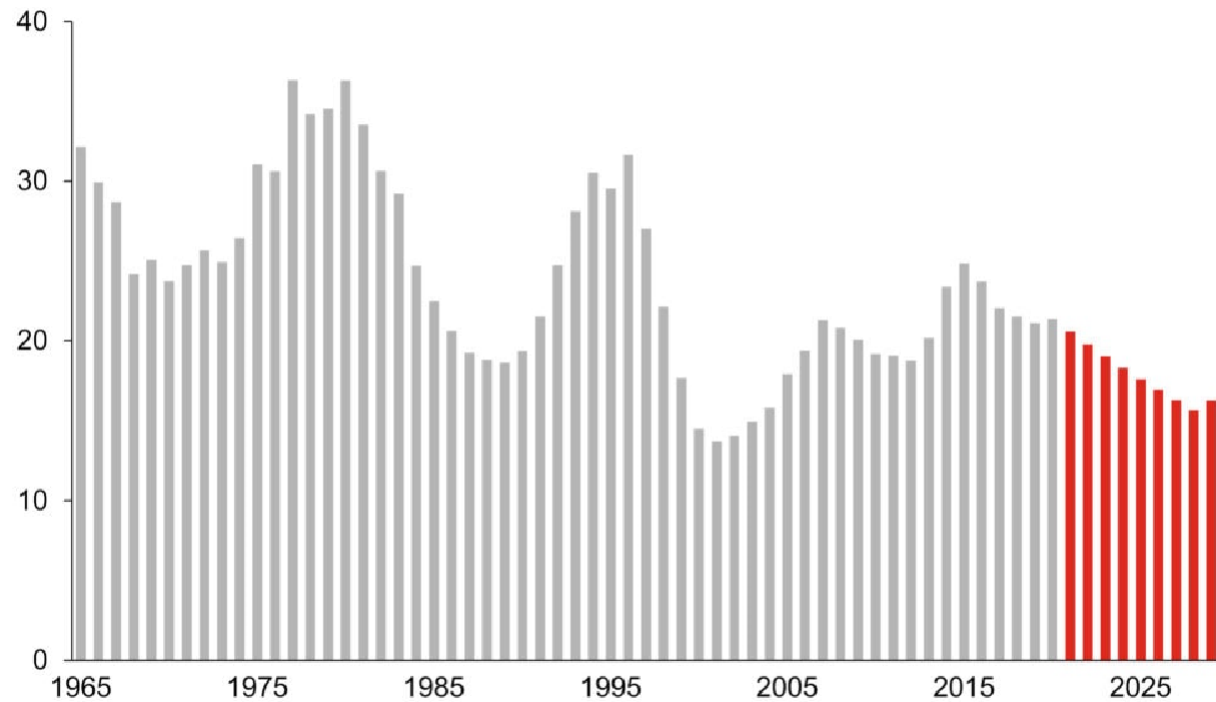
- Data management plans should cover **sustainability**.
 - In the absence of ongoing funding this equates broadly to **preservability**.
- Look for **established standards** for data and metadata at the outset
 - Consider their “preservation readiness”.
- **Consider how you might preserve all your research outputs.**
 - Data, publications and software should go into a repository and be given a PID
 - A project website might be submitted to the Internet Archive.
 - You might not bother with Twitter – that is a perfectly valid curatorial decision
- **Metadata and links to contextual objects are crucial.** Links will go outside your domain area.

Cloud-based Preservation

Is nothing new – it is the same hardware and software
but with one important new risk factor...

Expect to Migrate!

**Chart 1: Average company lifespan on S&P 500 Index in years
(rolling 7-year average)**



Data: Standard & Poor's; Innosight analysis based on public S&P 500 data sources. See endnote on methodology.

A perpetual motion machine: The preserved digital scholarly record

Published: 2022-09-21

<https://doi.org/10.1002/leap.1494>