arkivum

Bringing archived data to life

Digital Preservation Technology Futures

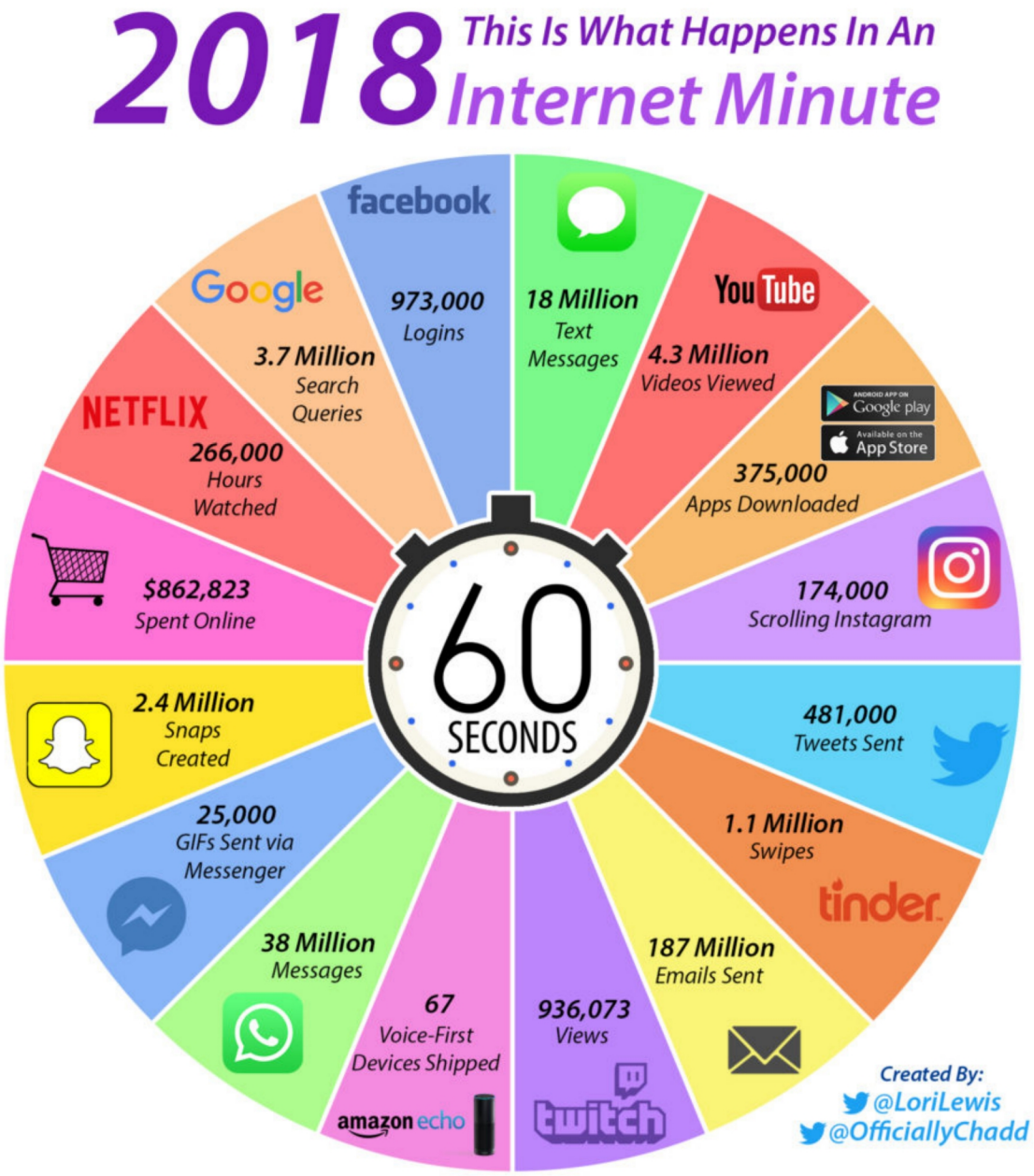
Matthew Addis - Arkivum

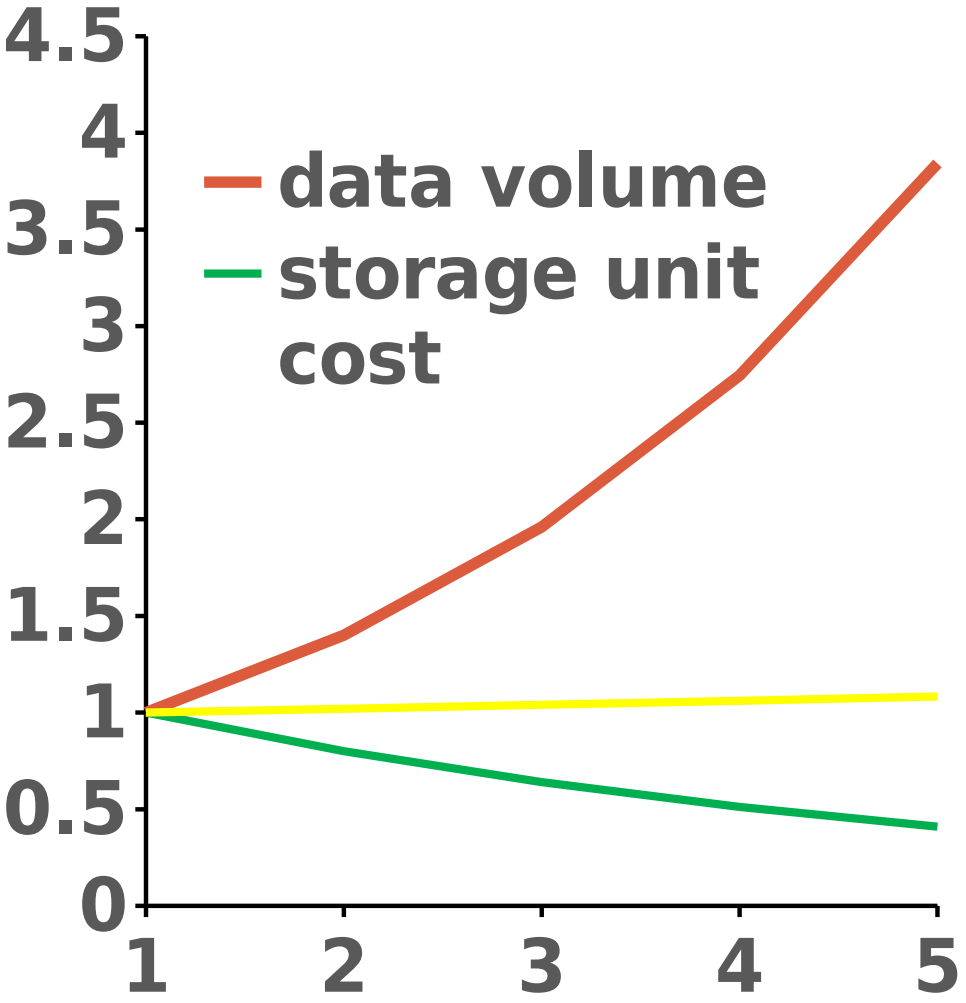
25 June 2017

Digital Preservation Technology … Where Next?

Kat B. CC BY-NC-ND 2.0, https://flic.kr/p/rhuN6B

1. Volume

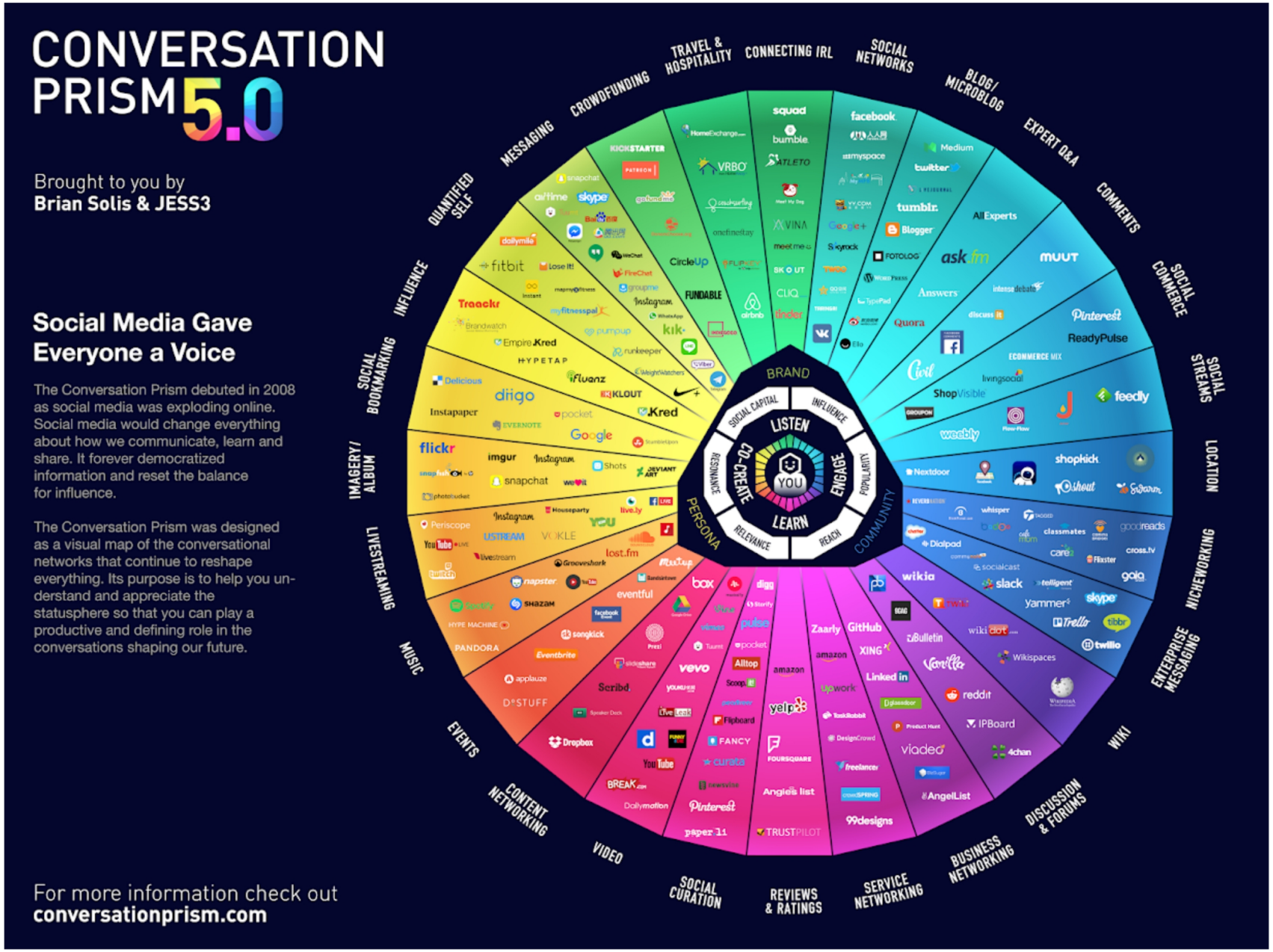

2018 This Is What Happens In An Internet Minute
facebook
973,000 Logins
18 Million Text Messages
YouTube
4.3 Million Videos Viewed
375,000 Apps Downloaded
174,000 Scrolling Instagram
481,000 Tweets Sent
1.1 Million Swipes
tinder
187 Million Emails Sent
936,073 Views
twitch
67 Voice-First Devices Shipped
amazon echo
38 Million Messages
25,000 GIFs Sent via Messenger
2.4 Million Snaps Created
$862,823 Spent Online
NETFLIX
266,000 Hours Watched
Google
3.7 Million Search Queries
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd

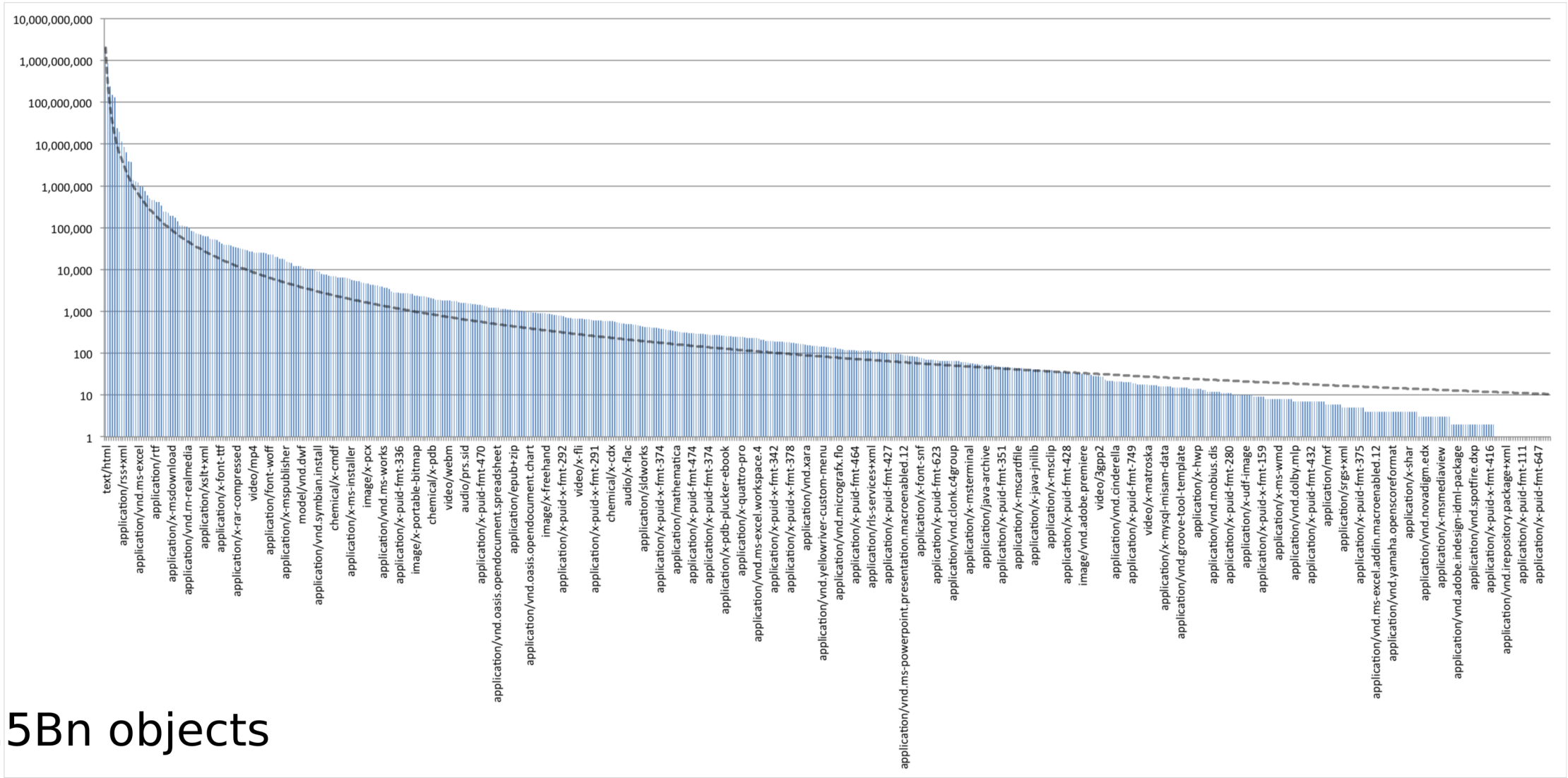


- Keep everything is not an answer
- Falling storage costs is not enough

- Automation
- Appraisal
- Selection
- Disposition

2. Variety


CONVERSATION PRISM 5.0
Brought to you by
Brian Solis & JESS3
Social Media Gave Everyone a Voice
The Conversation Prism debuted in 2008 as social media was exploding online. Social media would change everything about how we communicate, learn and share. It forever democratized information and reset the balance for influence.
The Conversation Prism was designed as a visual map of the conversational networks that continue to reshape everything. Its purpose is to help you understand and appreciate the statusphere so that you can play a productive and defining role in the conversations shaping our future.
For more information check out
conversationprism.com
TRAVEL & HOSPITALITY
CONNECTING IRL
SOCIAL NETWORKS
BLOG/ MICROBLOG
EXPERT Q&A
COMMENTS
SOCIAL COMMERCE
SOCIAL STREAMS
LOCATION
NICHE/NETWORKING
ENTERPRISE MESSAGING
WIKI
DISCUSSION & FORUMS
BUSINESS NETWORKING
SERVICE NETWORKING
REVIEWS & RATINGS
SOCIAL CURATION
VIDEO
CONTENT NETWORKING
EVENTS
MUSIC
LIVESTREAMING
IMAGERY / ALBUM
SOCIAL BOOKMARKING
INFLUENCE
QUANTIFIED SELF
MESSAGING
CROWDFUNDING
BRAND
INFLUENCE
POPULARITY
REACH
RELEVANCE
RESONANCE
PERSONA
SOCIAL CAPITAL
LISTEN
ENGAGE
LEARN
CO-CREATE
YOU
COMMUNITY



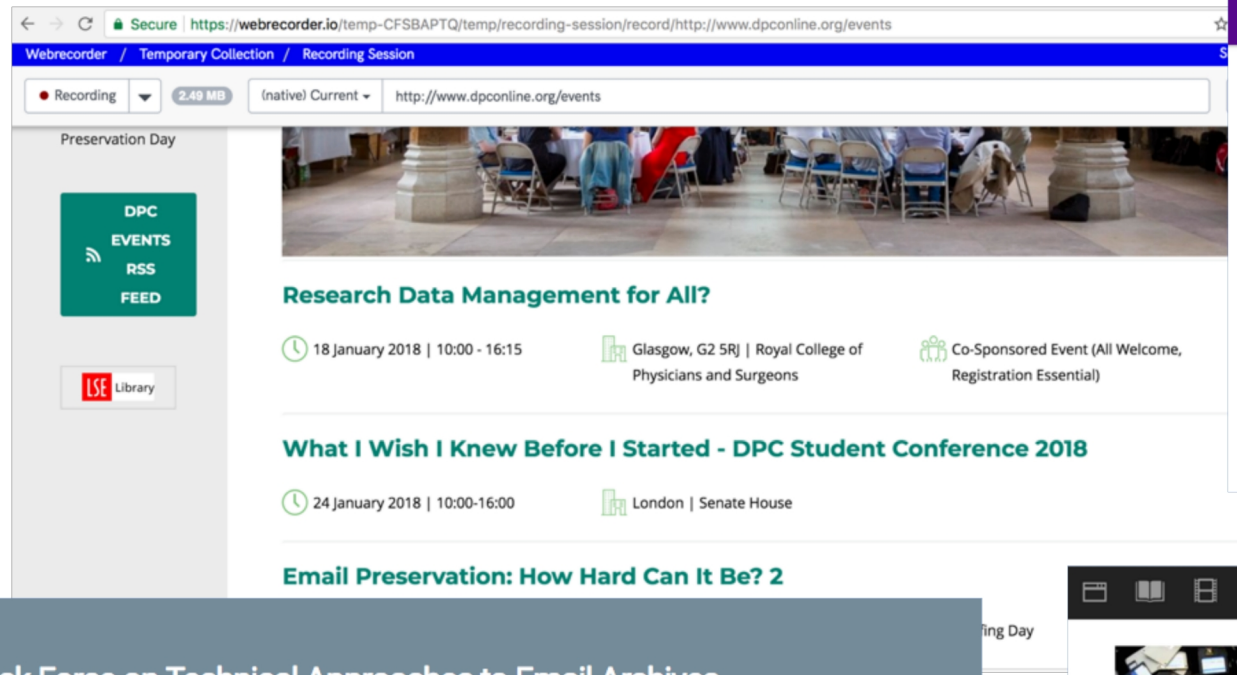
3.5Bn objects

99.82% in just 20 formats (html, xml, text, image, video, MS office)

19th is ‘unknown format’ (5,015 extensions)

RHIZOME

https://webrecorder.io/



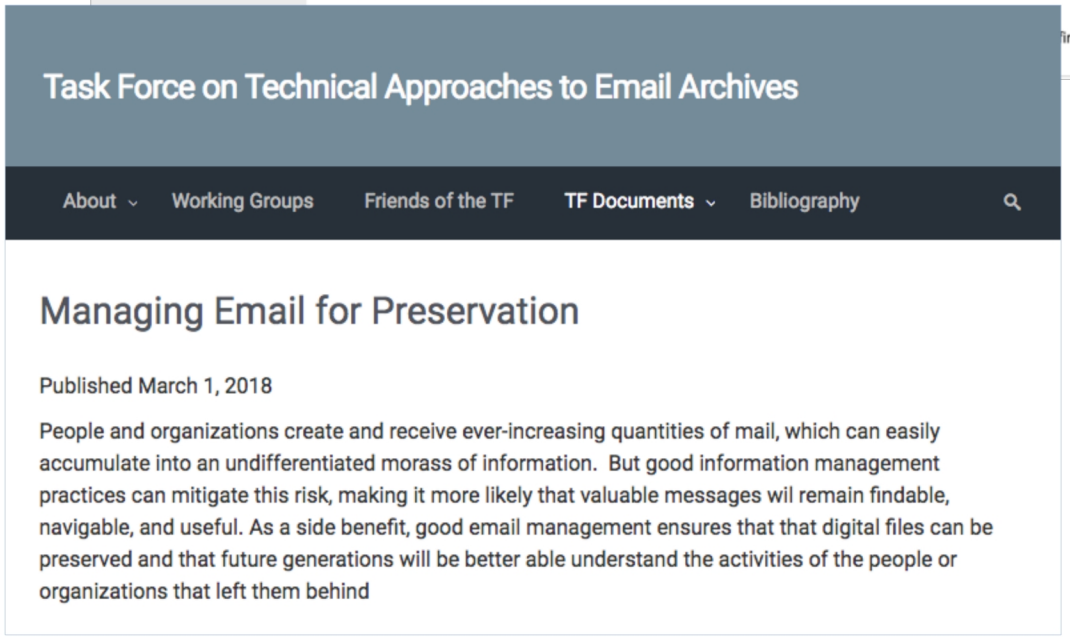
ReproZip | About | News | Documentation | GitHub Project | Examples

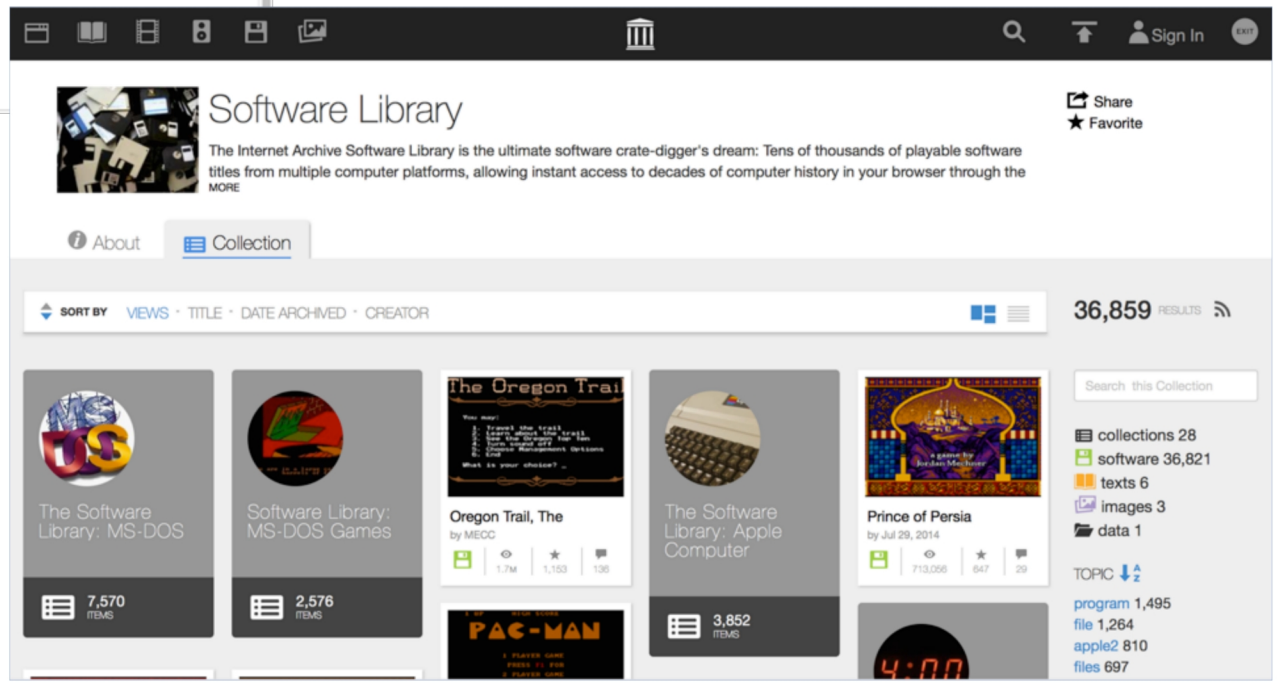
Automatically pack your research to be run elsewhere!

ReproZip allows you to pack your research along with all necessary data files, libraries, environment variables and options.

Then anybody can reproduce the research on a different machine, without tracking down and installing the dependencies, or even having to run the same operating system!

https://www.reprozip.org/





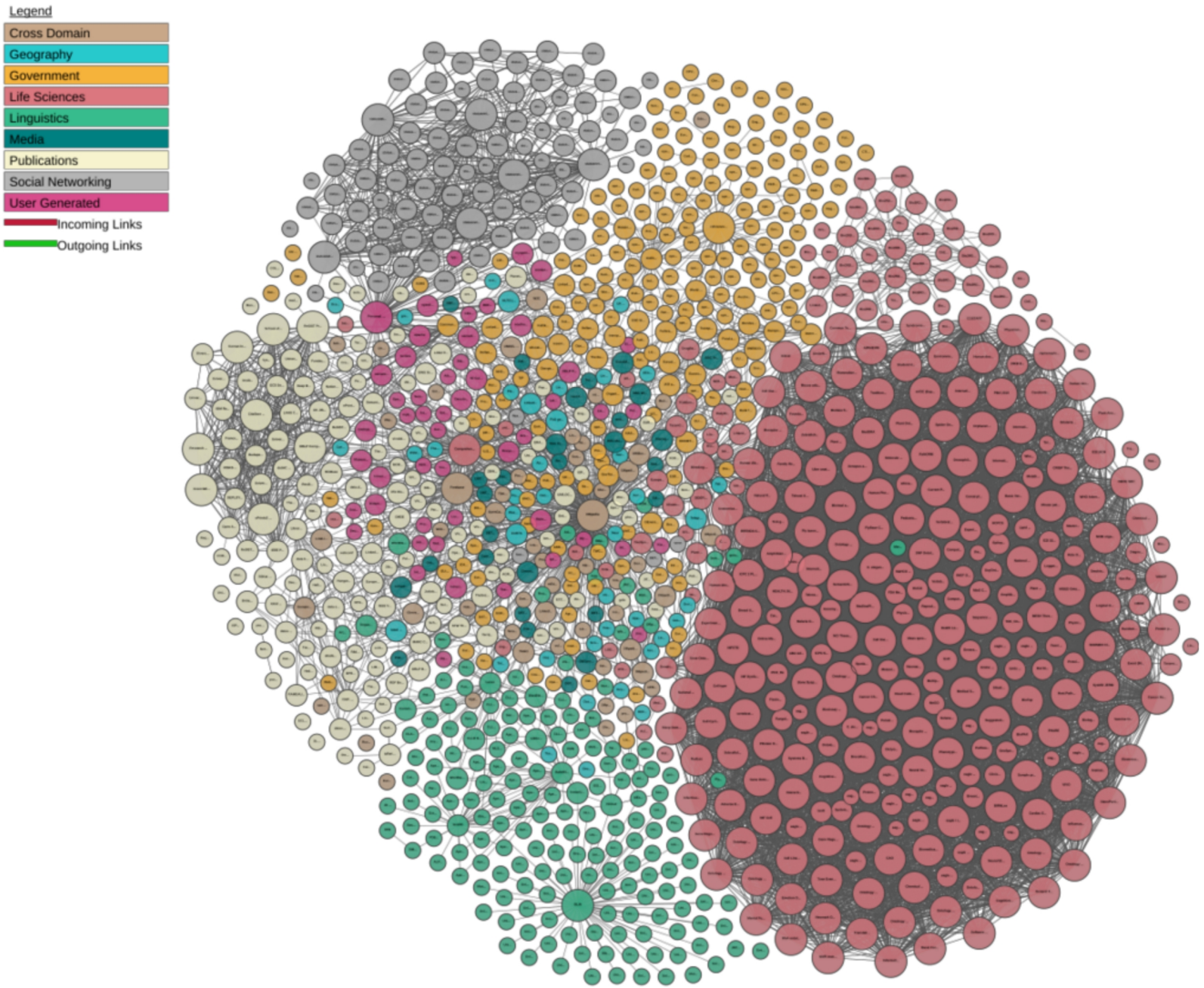
arkivum
Bringing archived data to life

http://www.emailarchivestaskforce.org/

https://archive.org/details/softwarelibrary

3. Connectedness

As of September 2008


Legend
Cross Domain
Geography
Government
Life Sciences
Linguistics
Media
Publications
Social Networking
User Generated
Incoming Links
Outgoing Links

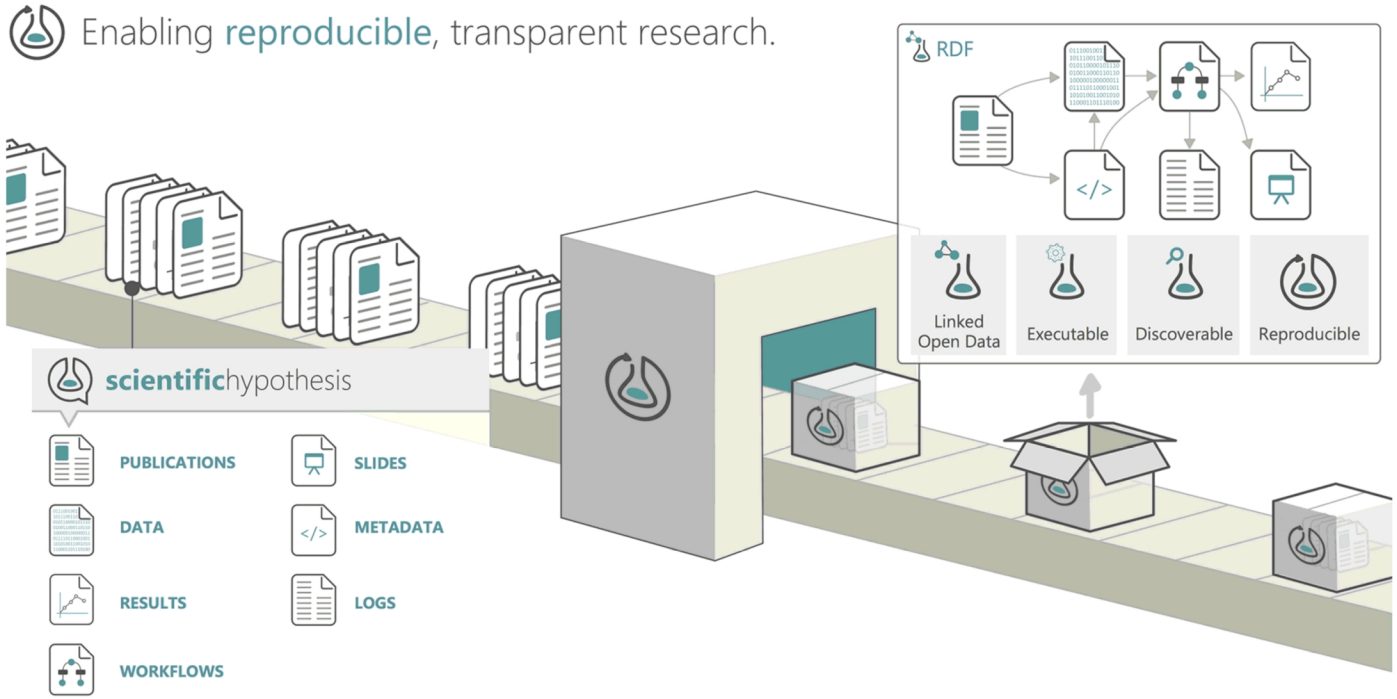
4. User experience

Fiona Wen Hui C, CC-BY-2.0, https://flic.kr/p/8pzadr

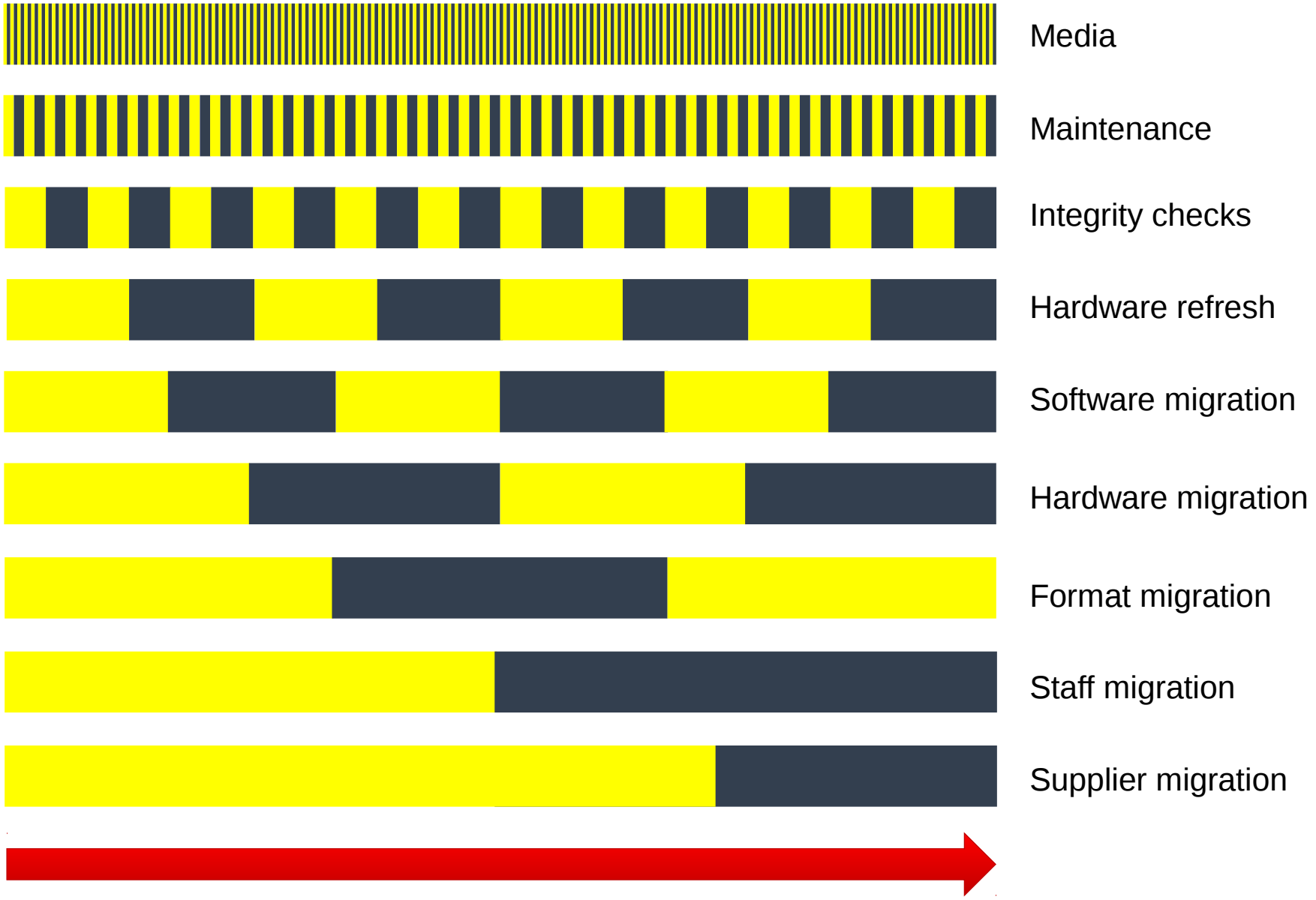
arkivum
Bringing archived data to life

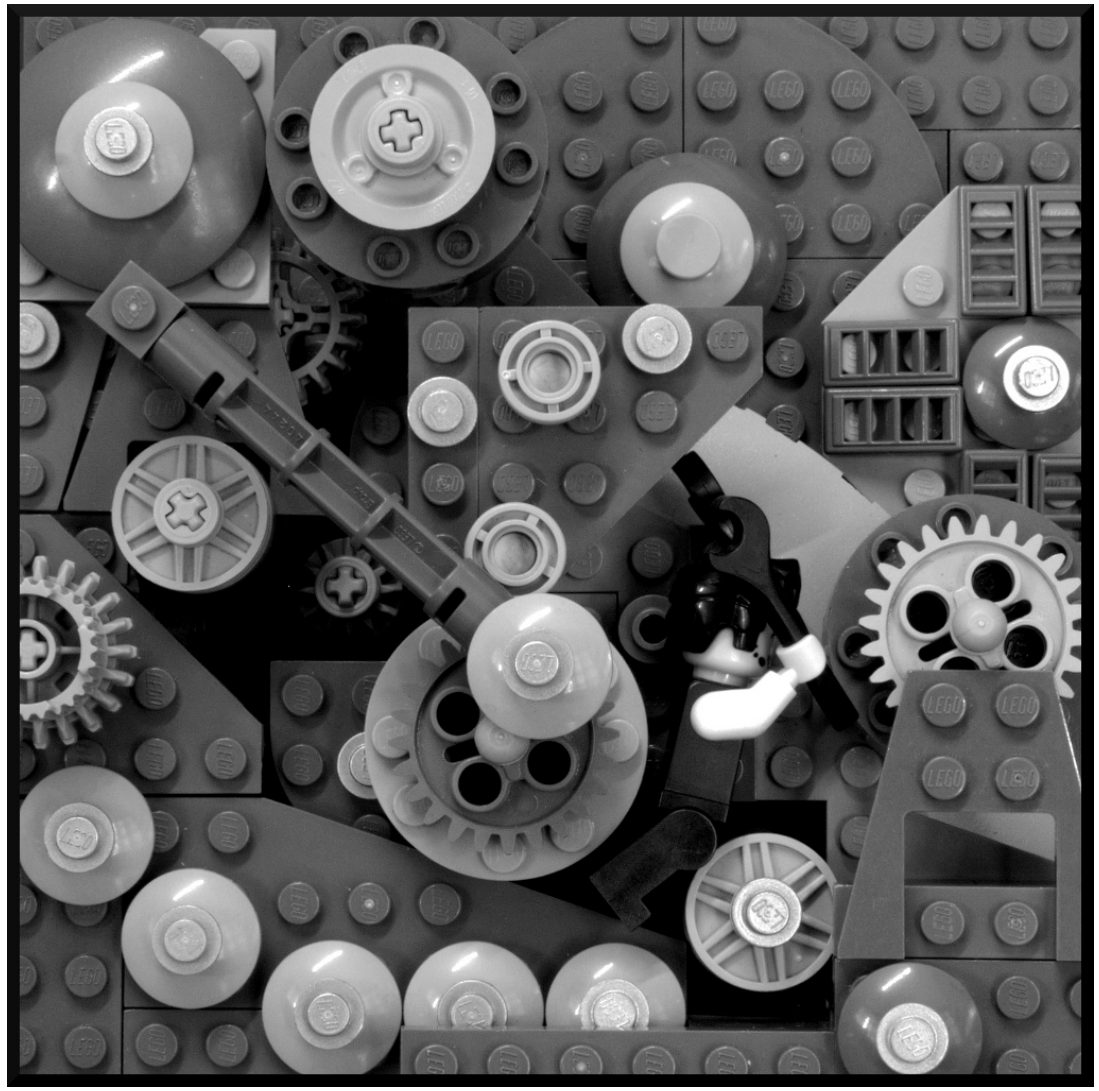
DRs Kulturarvsprojekt, CC BY-SA 2.0, https://flic.kr/p/aUg7Ma

- Record ‘what it was like at the time’
- Ability to recreate an experience, event or outcome
- Evidential weight and authenticity
- Access environments that allow users to interact with and use digital objects
- Packaging and linking of data, metadata, software, workflows, provenance



5. Change


Media
Maintenance
Integrity checks
Hardware refresh
Software migration
Hardware migration
Format migration
Staff migration
Supplier migration



- Preservation systems need to be modular
- Choose, integrate, migrate components

- Transparency
- Open Standards
- Open implementations and open source
- Open APIs
- Exit and migration strategies

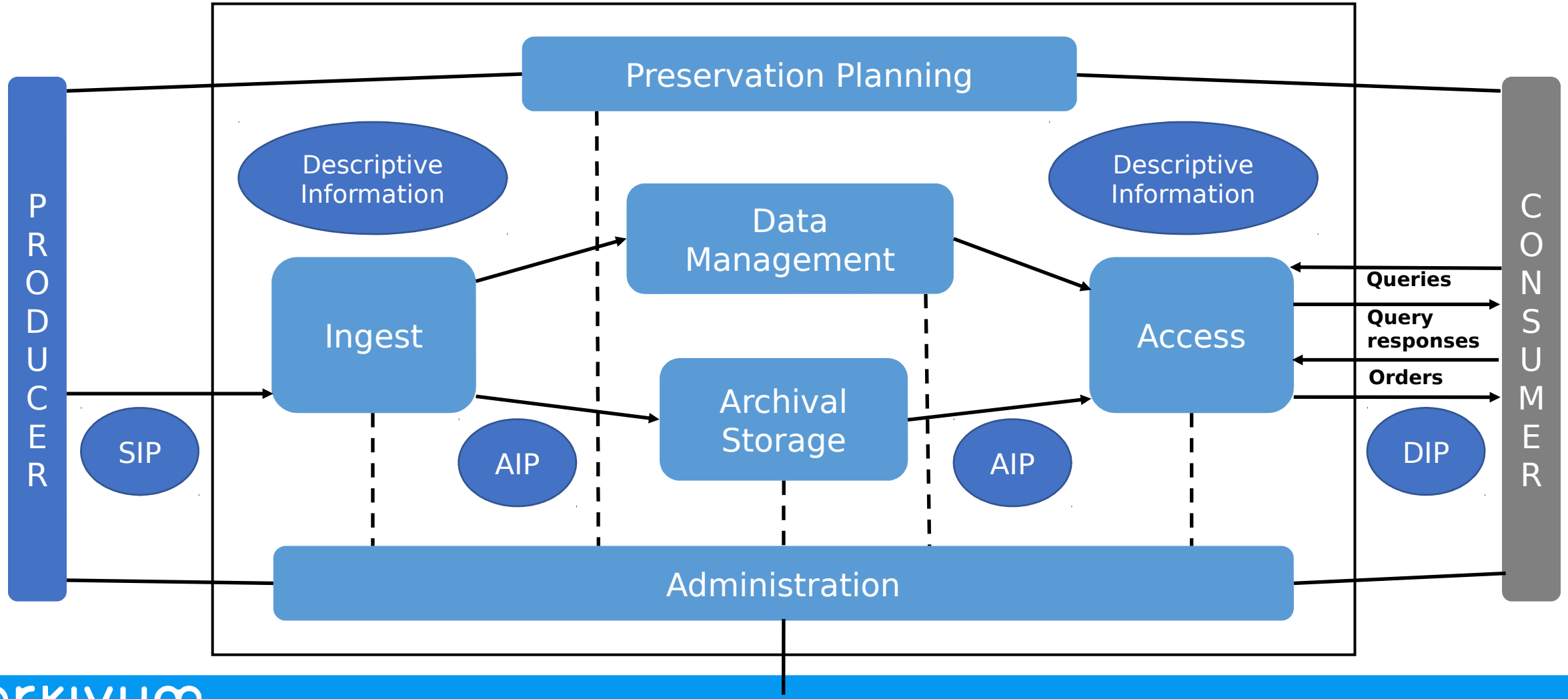
Neil Coulter, CC BY-SA 2.0, https://flic.kr/p/eBrhdn

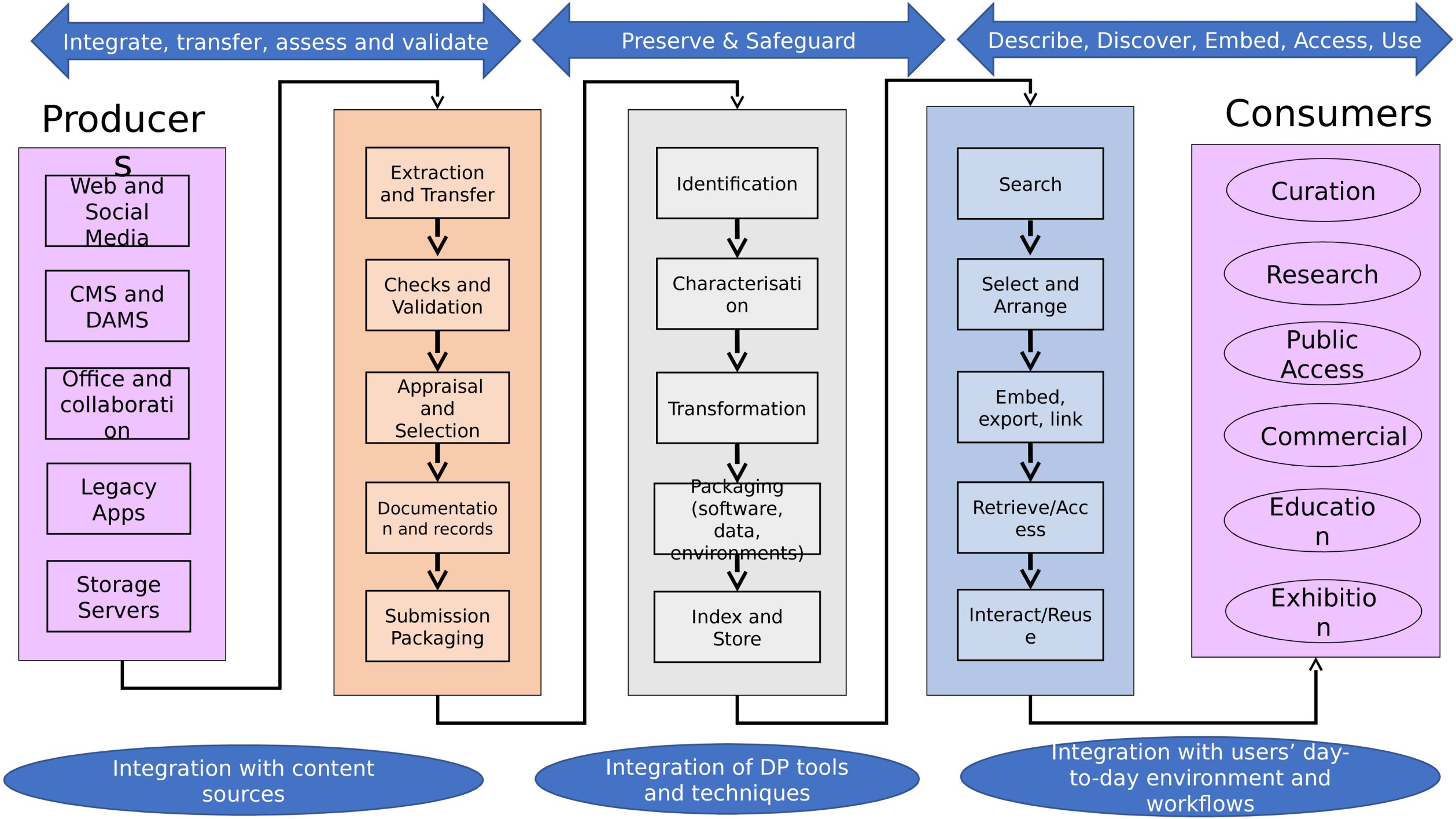
A digital preservation future

- Digital Preservation happens through integrated systems
  - Integration with content sources
  - Integrated tools for different content types
  - Integration with systems for content discovery and use
- Digital Preservation becomes more automated
  - Automated processing
  - Shared Knowledge
  - An ‘engineering’ approach
- Users can experience digital content in different contexts

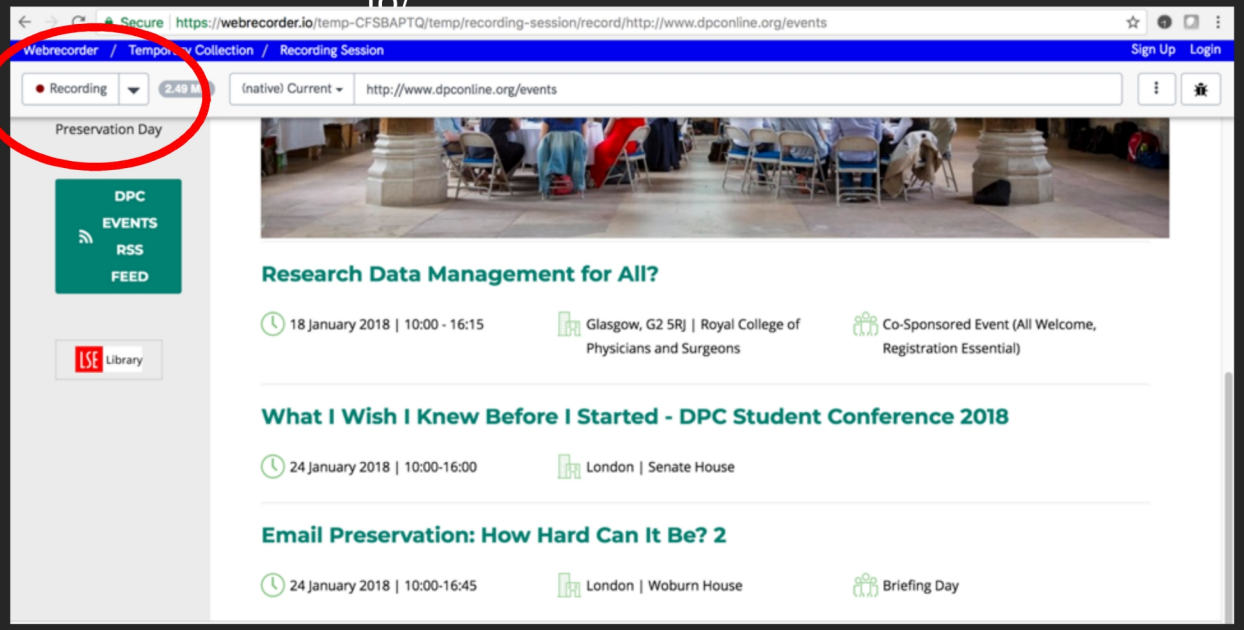
Open Archive Information System (OAIS)

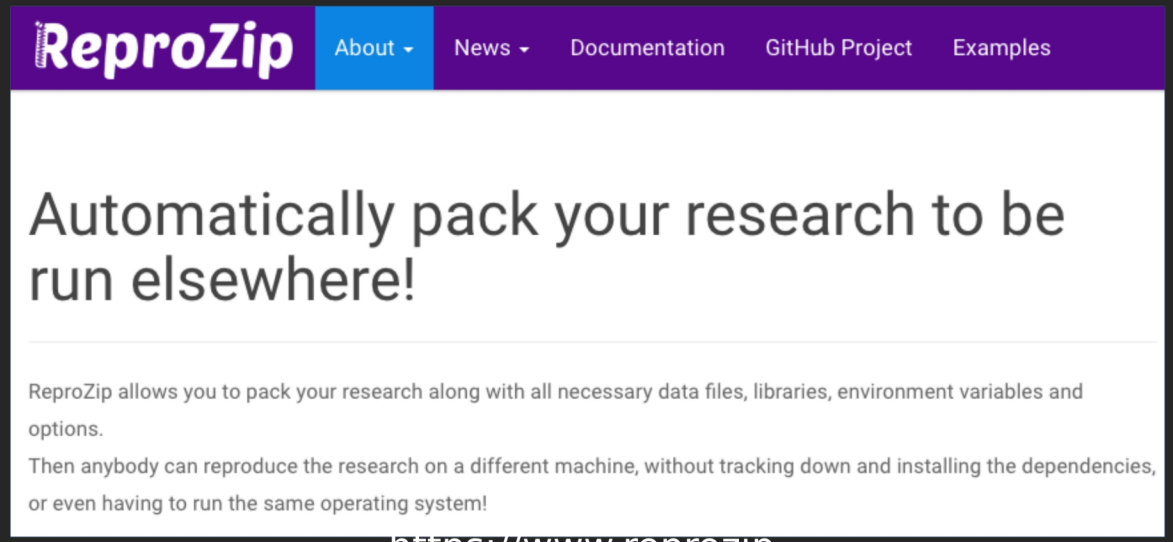
ISO 14721:2012 - OAIS



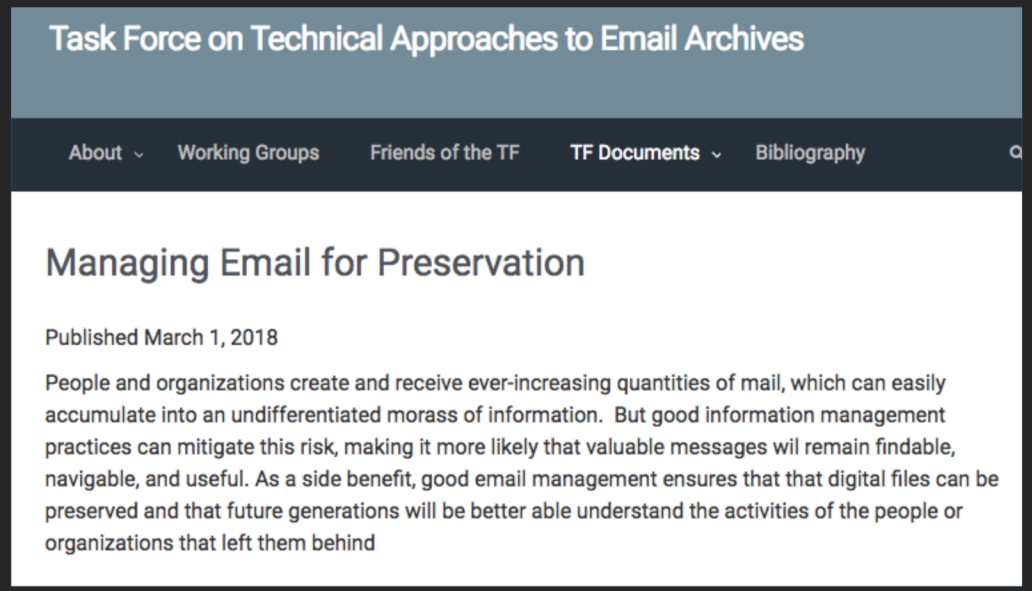

Integrate, transfer, assess and validate
Preserve & Safeguard
Describe, Discover, Embed, Access, Use
Producers
Web and Social Media
CMS and DAMS
Office and collaboration
Legacy Apps
Storage Servers
Extraction and Transfer
Checks and Validation
Appraisal and Selection
Documentation and records
Submission Packaging
Identification
Characterisation
Transformation
Packaging (software, data, environments)
Index and Store
Search
Select and Arrange
Embed, export, link
Retrieve/Access
Interact/Reuse
Consumers
Curation
Research
Public Access
Commercial
Education
Exhibition
Integration with content sources
Integration of DP tools and techniques
Integration with users' day-to-day environment and workflows

https://webrecorder
io/

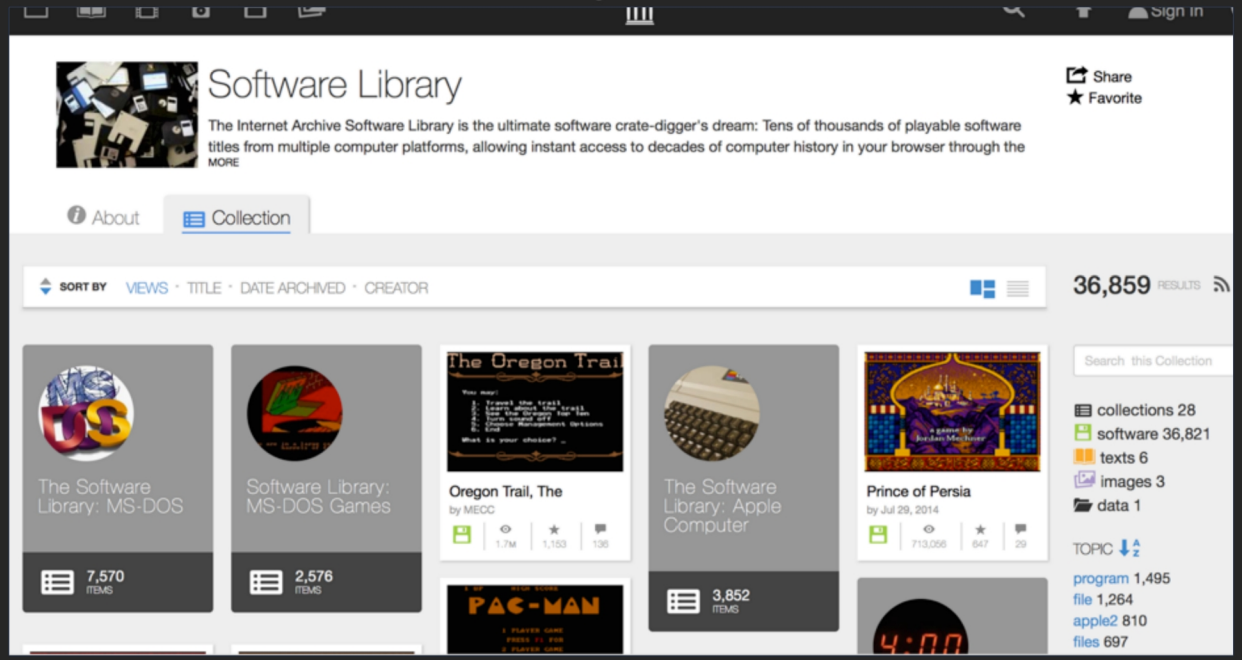




https://www.reprozip.
org/

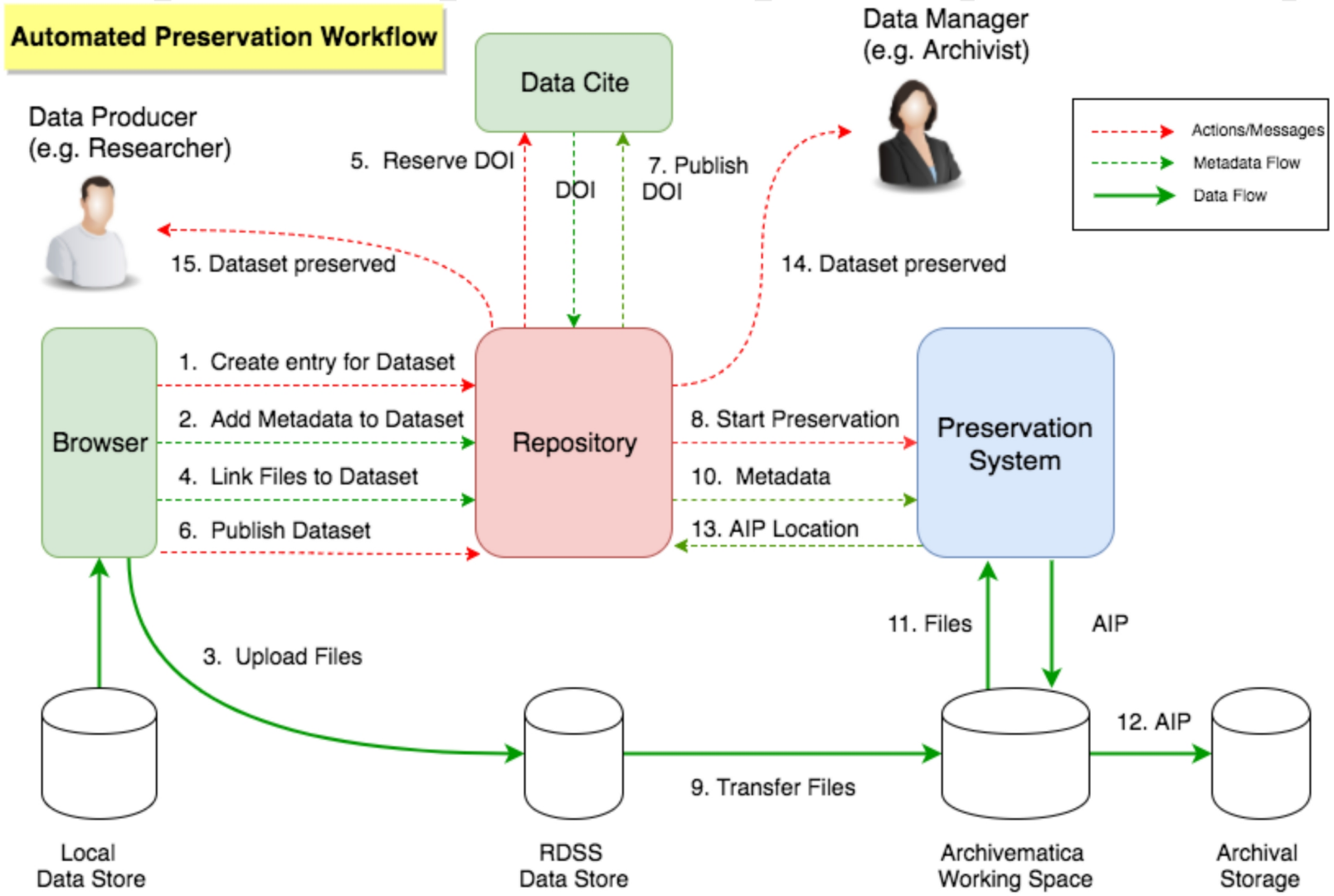


http://www.emailarchivestaskfor
ce.org/



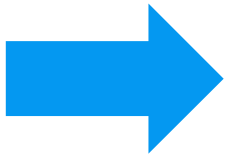
https://archive.org/details/softwarelibrary

Case Study: Jisc Research Data



Making Digital Preservation easier in the RDSS

- Embed preservation into research publication workflows
- Not something that happens at the ‘end of the line’
- Reduced manual intervention and specialist
- Preservation ‘disappears’ into day-to-day life

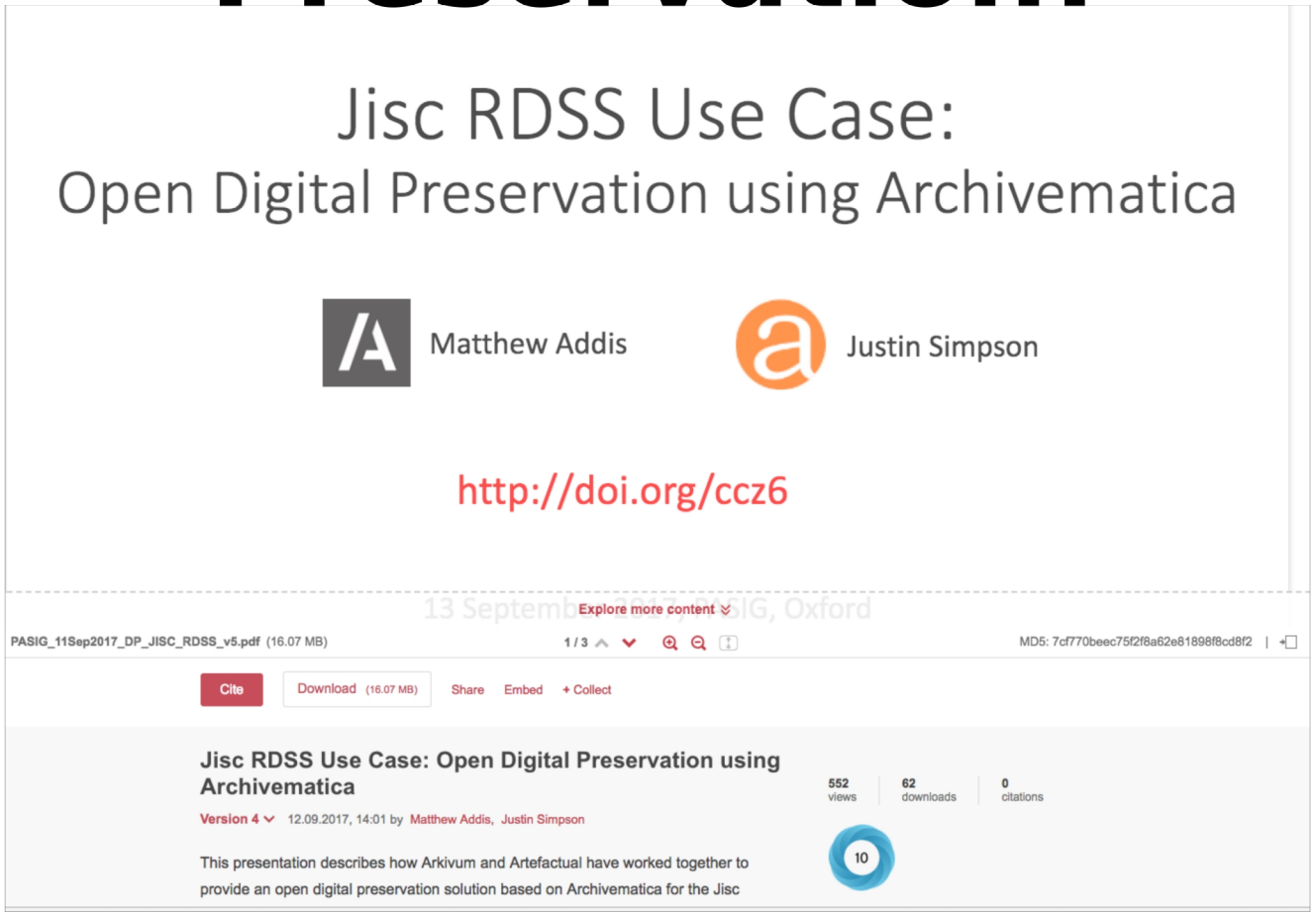


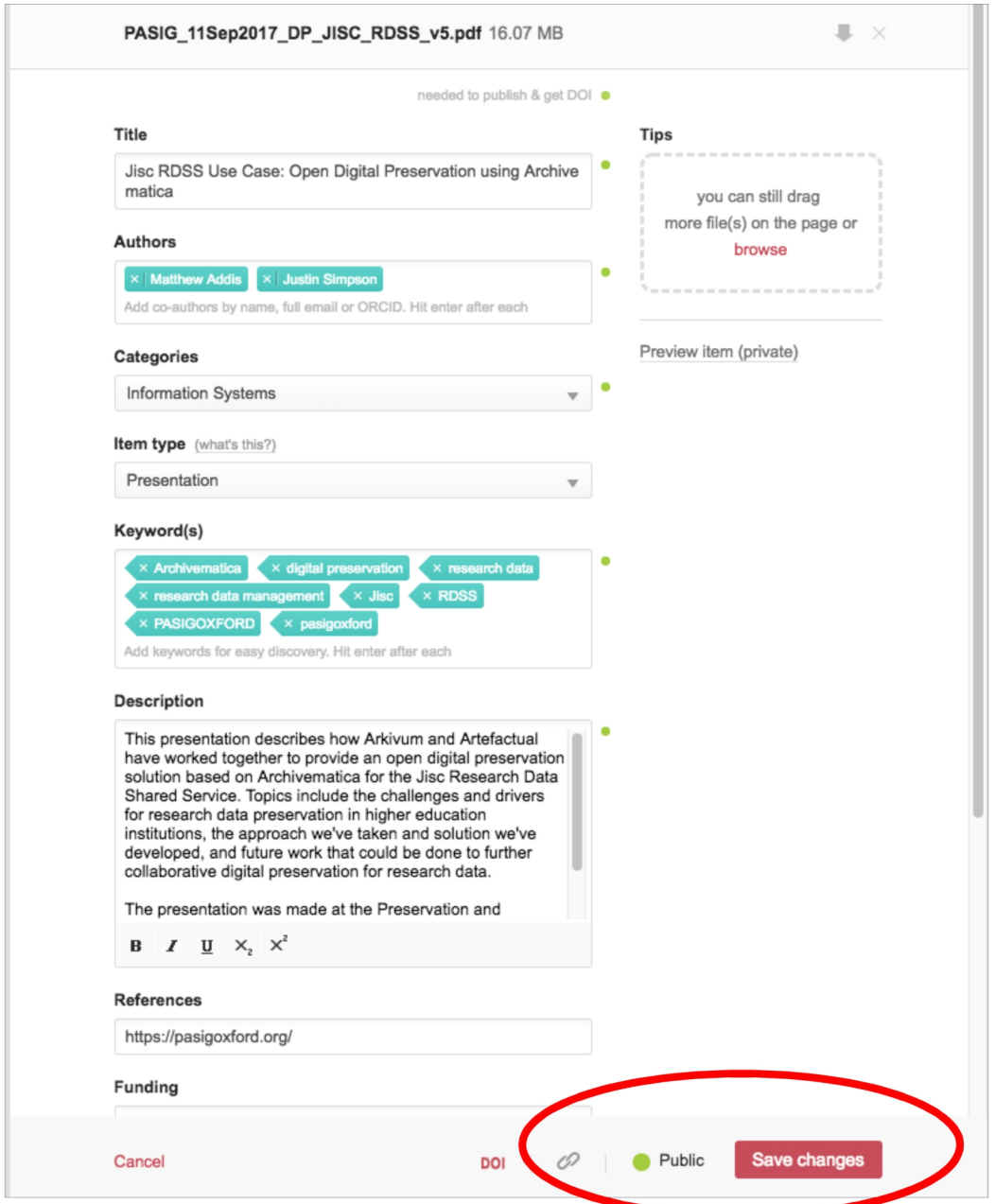
- Lower costs
- Increased throughput
- Reduced problems
- Less ‘slips through the net’

- 220 UK higher education institutions
- 95,000 research staff
- 450 PB of data
- 5,000+ different file types

- DP and storage budget: 1-2%
- 0.5 – 5 dedicated staff per institution
- Cover everything: outreach, training, support, guidelines, policies, workflows, infrastructure, procurement …

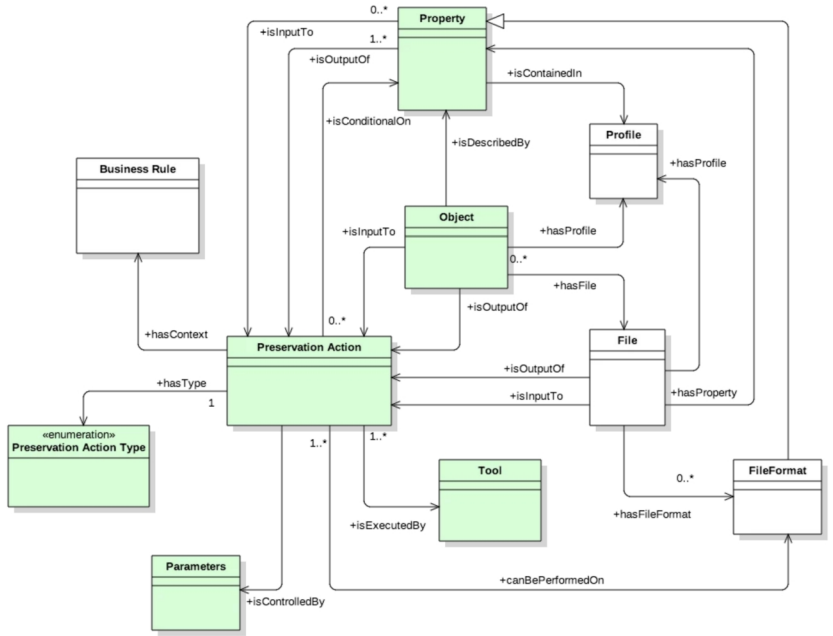
Hidden Preservation!

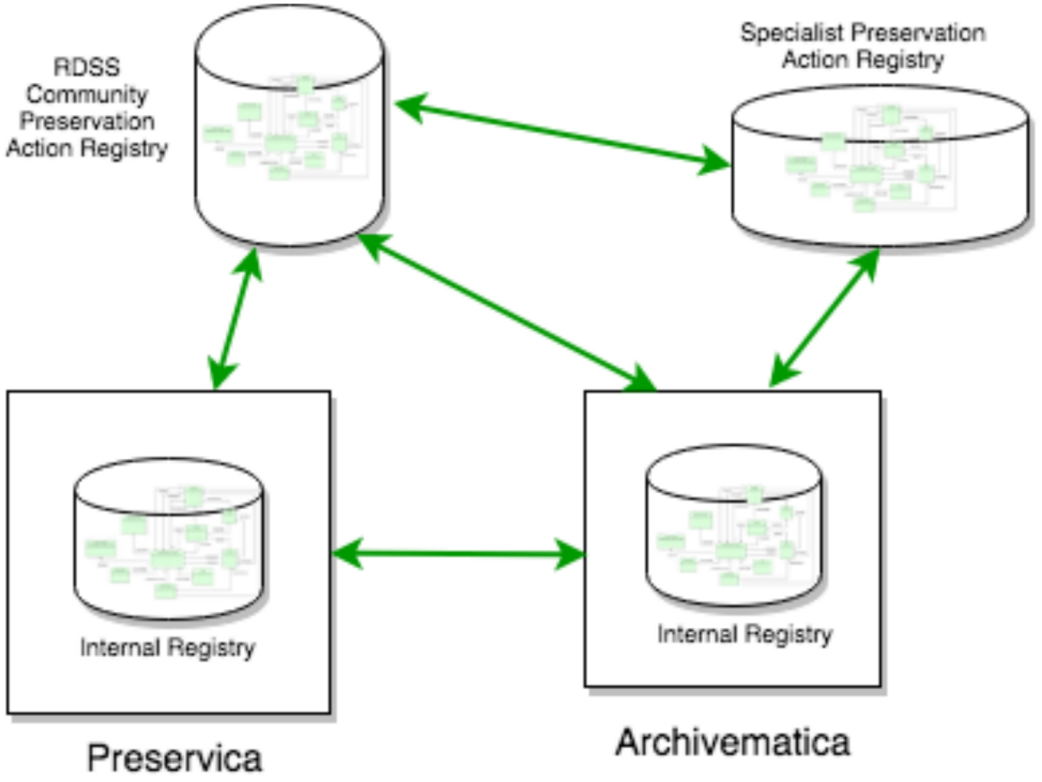




Preservation Actions Registry

- What preservation actions to take
- How to implement them
- Capture and share knowledge
- Trusted set of rules and tools
- Executable by preservation systems



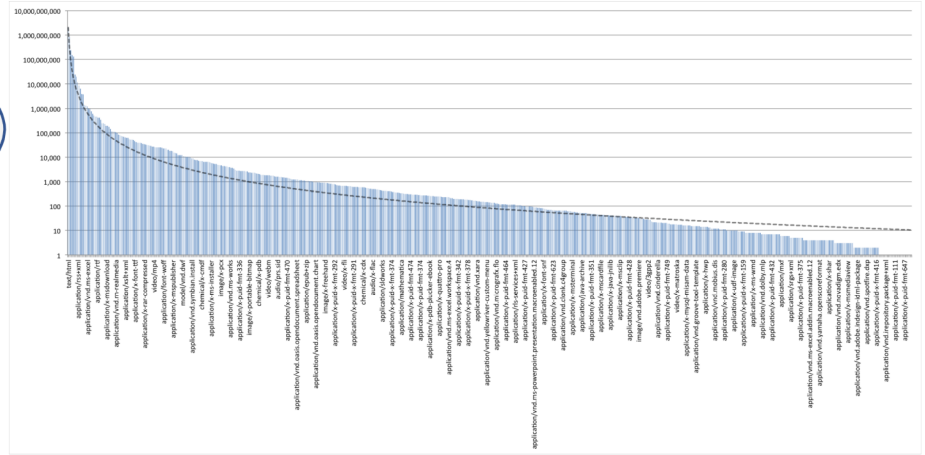


An ‘engineering approach’ to Digital Preservation

Automatic Preservation

Semi-automated Preservation

Interactive Preservation



Volume of content

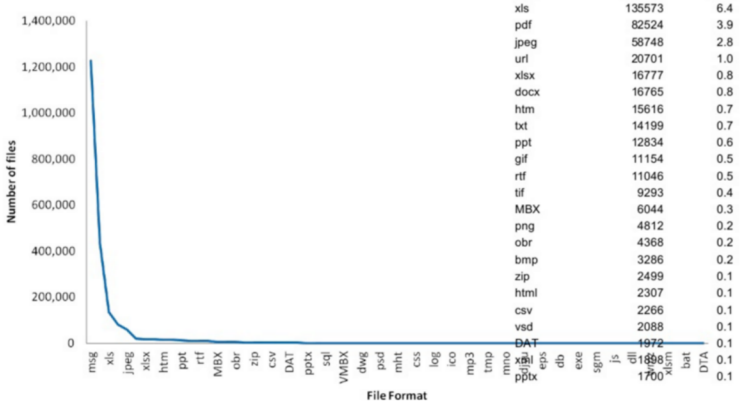
Majority of content, low cost

Minority of content, high cost

File types

The File format Long Tail

- Our own administrative records



File Extension	Count	%
msg	1225790	57.6
doc	437803	20.6
xls	135573	6.4
pdf	82524	3.9
jpeg	58748	2.8
url	20701	1.0
xlsx	16777	0.8
docx	16765	0.8
htm	15616	0.7
txt	14199	0.7
ppt	12834	0.6
gif	11154	0.5
rtf	11046	0.5
tif	9293	0.4
MBX	6044	0.3
png	4812	0.2
obr	4368	0.2
bmp	3286	0.2
zip	2499	0.1
html	2307	0.1
csv	2266	0.1
vsd	2088	0.1
DAT	1972	0.1
xml	1898	0.1
pptx	1700	0.1

Andrew Jackson, British Library. https://speakerd.s3.amazonaws.com/presentations/a2fd5875bd0346a4be56f14d7900f015/dpc-searching-for-obsolescence.pdf

Tim Gollins, The National Archive. http://www.slideshare.net/repofringe/repo-fringe-v0-01-gollins

Experiences, authenticity, environments for re-use

Mark Hesseltine, CC BY-ND 2.0, https://flic.kr/p/e1dvcq

A digital preservation future

- Digital Preservation will happen through integrated systems
  - Easier to capture, preserve, find and use content
  - Combine different techniques, tools and workflows
  - Open and flexible, avoid lock-in, easier to sustain
- Digital Preservation will become more automated
  - Lower costs, scale-up
  - Focus valuable staff resources on the ‘hard and valuable’ content
  - Community sharing of knowledge and expertise
- Users can interact with digital content in different contexts

arkivum

Bringing archived data to life

matthew.addis@arkivum.com

orcid.org/0000-0002-3837-2526

www.arkivum.com