Data Preservation

Techniques, Traps and Processes

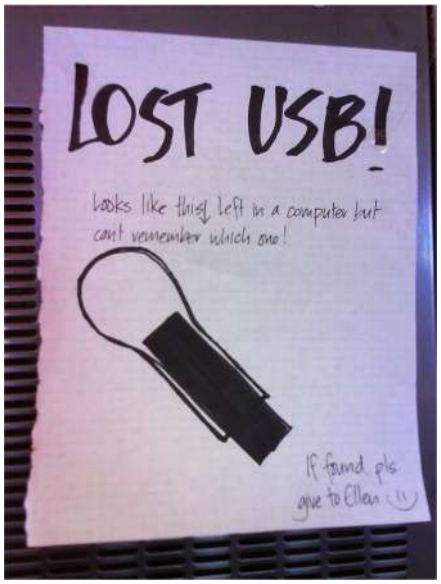
Shaun de Witt, United Kingdom Atomic Energy Authority

RECAP

- Why do we preserve data
  - Requirement of funders, support claims, speed research, protect against data loss
- Data Loss scenarios
  - Media loss
  - Format loss
  - Bad hats
  - Funding loss
  - Link loss

Media loss

HDD Failure
Often get 'soft' warnings
Failure rates 1-8% p.a.
Data recovery often impossible

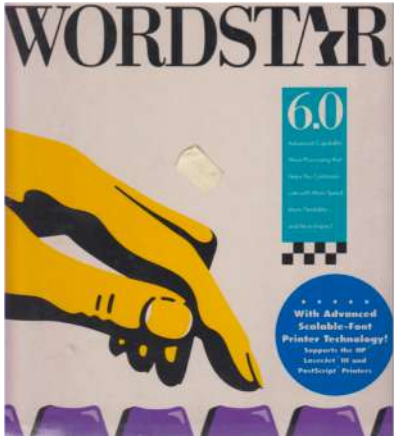


Human Failure
No warning, moderate rate
Data recovery often impossible

Tape Failure
Very reliable
Increasing chance of failure with number of reads
High chance of partial data recovery

Planning Failure
Data still on media that can't be read
Data recovery often impossible

Format Loss



Commercial Obsolescence

But what about Science?

- *In-house formats still used (ipx in fusion)*
  - *But more and more 'open' formats are being adopted*
- *Myriad of formats*
  - http://justsolve.archiveteam.org/index.php/Scientific_Data_formats
  - *13 just for Astronomy!*
- *What about databases?*
  - *Still direct access to databases and poorly understood database design*

SYBASE® An SAP Company

Its not just word processors!

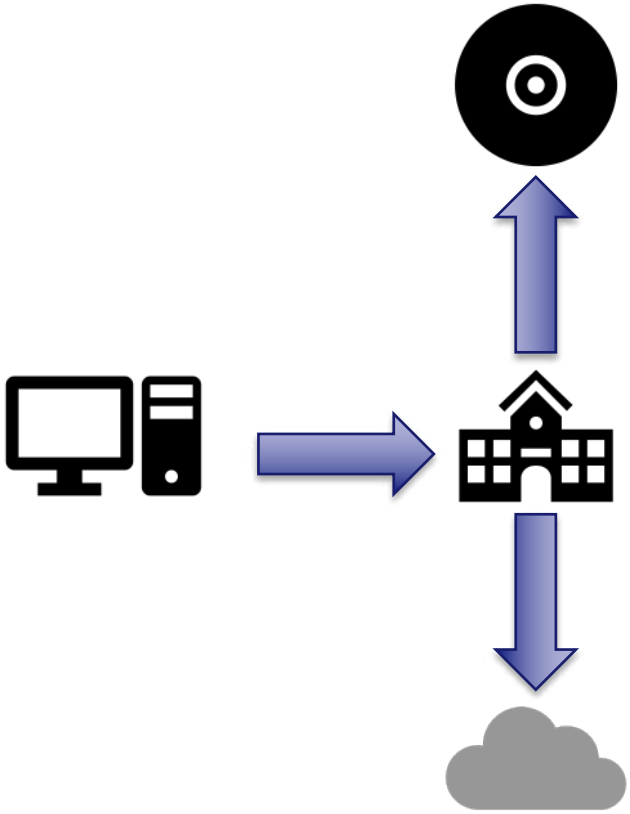
Bad Hats and accidents

- Bad hats can get in and delete, modify, or add to your data
  - In the 1990's Gigabytes of porn was dumped on University servers at Reading University
  - "Climategate"
- Sometimes people make mistakes
  - 600TB of particle physics data was accidentally deleted while trying to correct another problem (about 2/3 of all its data)
  - "I needed more space so I deleted this file called vmunix?"
  - Using *rm –rf ** as root (find directory ..)

Link Rot

- 2014 study showed 50% of the URLs in U.S. Supreme Court opinions no longer link to the original information
  - doi:10.1017/1472669614000255
- August 2015 Weblock analyzed more than 180,000 links from references in the full-text corpora of three major open access publishers and found that overall 24.5% of links cited were no longer available.
  - https://web.archive.org/web/20160304081204/https://weblock.io/report?id=all
- 2016-17: Study of Yahoos! Directory links showed a half life of 2 years

Data Preservation Techniques - Backup



Traditional Backup to Media

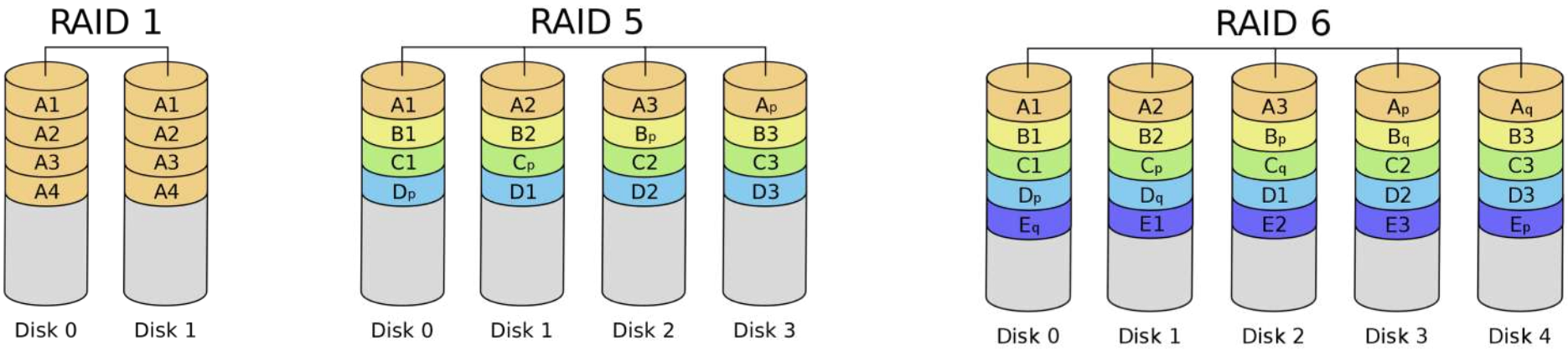
- *Still subject to media failure*
- *Only done once a day*
- *Effortful to recover*

Backup to 'cloud'

- ***Can be** backed up more frequently*
- *Recovery **can be** costly and time consuming*
- *Subject to commercial Terms and Conditions*
- *Expensive to use for long term preservation*
  - *E.g. UKAEA cloud backup is only 30 days*
  - *For longer term preservation we still use n in-house system*

- *In both cases it may be difficult to recover version history and is not suitable for large volumes of data*
  - *It can take > 1 day to back up scientific data sets*

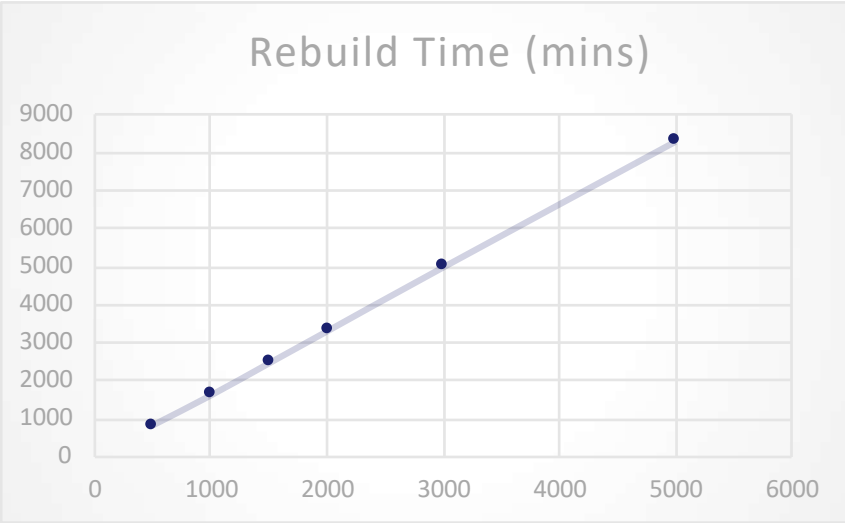
Data Preservation - RAID

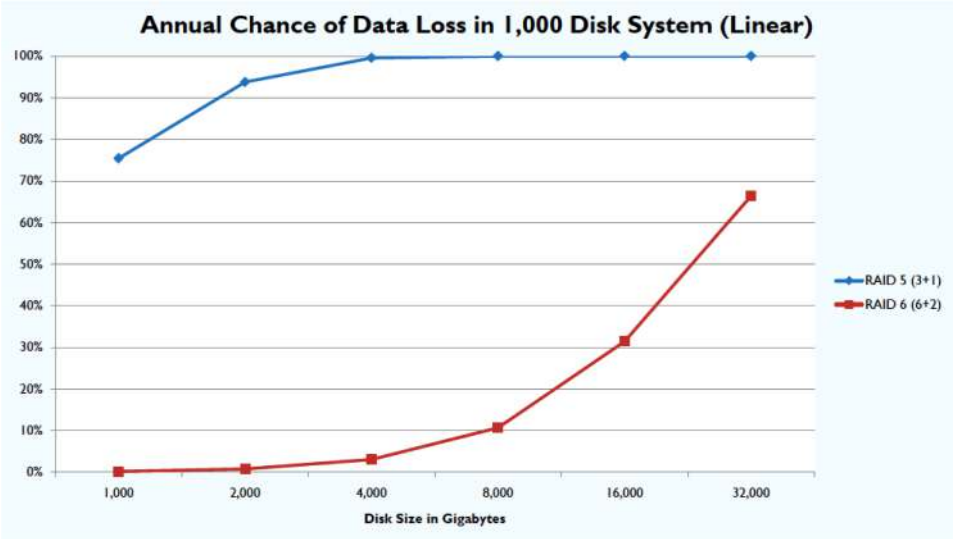


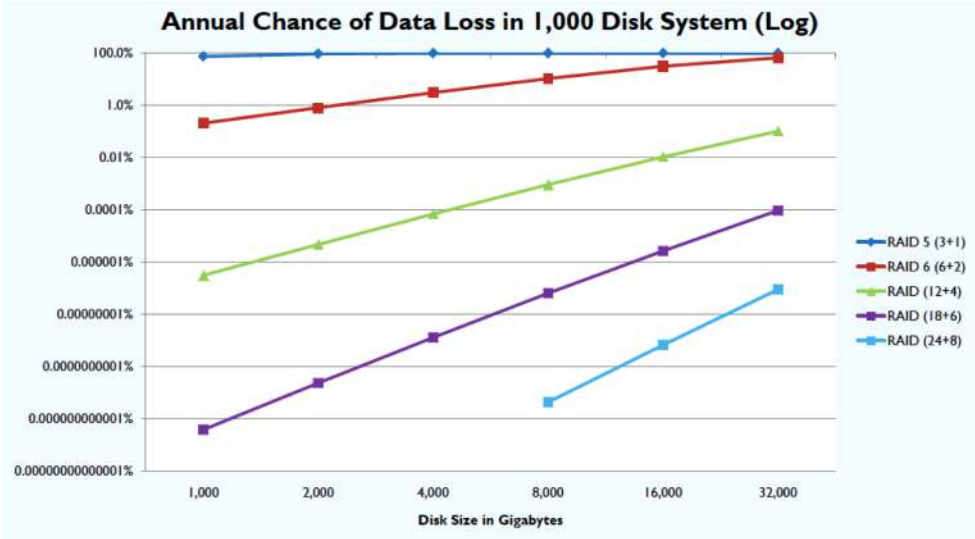
RAID Level	Description	Min. Drives	Space Loss	Fault Tolerance
RAID 0	Striped	2	1	None
RAID 1	Mirroring	2	1/n	n-1 drive failures
RAID 1+0	Striped + mirrored	2	1/n	n-1 drive failures
RAID 5	Striped + 1 parity	3	1-1/n	1 drive failure
RAID 6	Striped + 2 parity	4	1-2/n	2 drive failures

The problems with RAID

- Data rebuild times increase with disk size
- This graph is theoretical
  - Reality is double or triple this
- Typically 1 or two ‘hot spares’ are reserved to cover disk failures, so more space is lost than nominal
- And you still have the single site problem

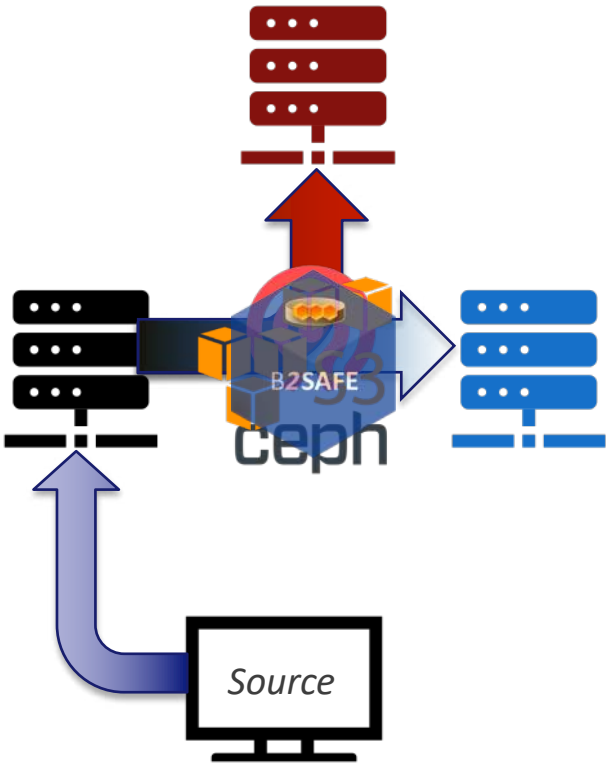






© J. Resch, Cleversafe, Storage Developers Conference, 2010

Data Replication



- Commonly used within **object stores** with 3x replication
  - Normally replicas held at different places
- Also common on **cloud services** to give very high resilience
- Can be used distribute data to partner operations
  - E.g. data generated in UKAEA could be replicated to PSNC (another fusion research establishment) and CINECA (which hosts the community HPC facility)
- In most cases data integrity is maintained by **checking-summing** each object
  - Either as a background task or on retrieval
  - Many tools can self-heal in case of failure

The Good and the Bad of Checksums

- EXERCISE

Summary of Results

Original Image MD5sum: 9bceeae7b0b4c19df35052b9018c9e8d

Bit-flipped Image MD5sum: d46866f79200465f94baede9450928e0

To Whom it May Concern:

Alice Falbala fulfilled all the requirements of the Roman Empire intern position. She was excellent at translating roman into her gaul native language, learned very rapidly, and worked with considerable independence and confidence.

Her basic work habits such as punctuality, interpersonal deportment, communication skills, and completing assigned and self-determined goals were all excellent.

I recommend Alice for challenging positions in which creativity, reliability, and language skills are required.

I highly recommend hiring her. If you'd like to discuss her attributes in more detail, please don't hesitate to contact me.

Sincerely,

Julius Caesar

May, 22, 2005

Order:

Alice Falbala is given full access to all confidential and secret information about GAUL.

Sincerely,

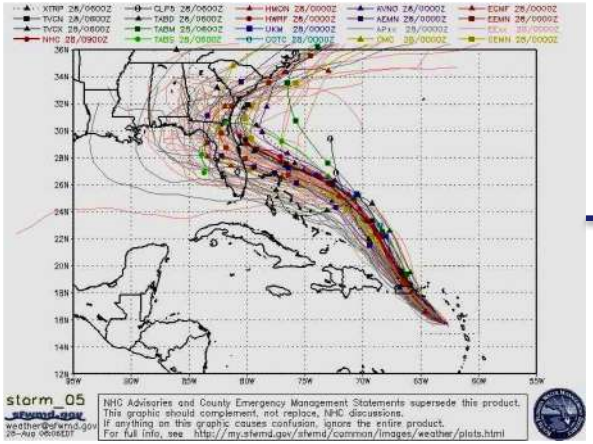
Julius Caesar

A Letter of Recommendation? Or Security Clearance?

A Panda? Or a Bucket?

Extending FAIR

- Making data **open** or **FAIR** is generally a good thing
  - Faster Science, More verifiable results, Better impact assessment, driving the data and digital economies, etc, etc…
- BUT… There are **risks**
  - **Accidental and Malicious**
  - **Lets look at some examples**



More Examples – The Reproducibility Crisis

Gene name errors are widespread in the scientific literature

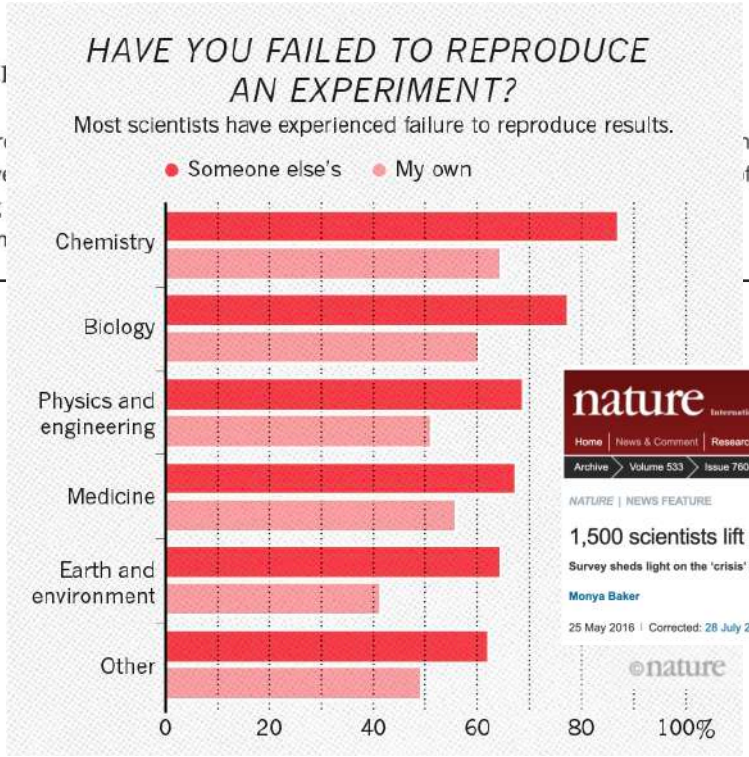
Mark Ziemann, Yotam Eren and Assam El-Osta

Genome Biology 2016 17:177 | DOI: 10.1186/s13059-016-1044-7 | © The Author(s). 2016

Published: 23 August 2016

Abstr

The spr
to conv
leading
supplen



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Vide

Archive > Volume 533 > Issue 7604 > News Feature > Article

NATURE | NEWS FEATURE

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

Monya Baker

25 May 2016 | Corrected: 28 July 2016

Ioannidis et al. Nature Genetics, 41, 2010
doi:10.1038/ng.295

nature genetics

ANALYSIS

56% of analyses could not be repeated,
of which 30% were because of software issues.
50% did not state software version, 39% did not provide raw data.
Only 11% could be reproduced satisfactorily.

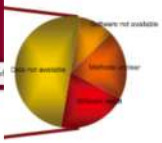
Repeatability of published microarray gene expression analyses

John P A Ioannidis[1–3], David B Allison[4], Catherine A Ball[5], Issa Coulibaly[4], Xiangqin Cui[4], Aedín C Culhane[6,7], Mario Falchi[8,9], Cesare Furlanello[10], Laurence Game[11], Giuseppe Jurman[10], Jon Mangion[11], Tapan Mehta[4], Michael Nitzberg[5], Grier P Page[4,12], Enrico Petretto[11,13] & Vera van Noort[14]

Given the complexity of microarray-based gene expression studies, guidelines encourage transparent design and public

CAMBRIDGE JOURNAL OF ECONOMICS

Issues | JEL | More Content | Submit | Purchase | About



Volume 38, Issue 2
March 2014

Next >

Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff

Thomas Herndon, Michael Ash, Robert Pollin

Cambridge Journal of Economics, Volume 38, Issue 2, March 2014, Pages 257–279,
https://doi.org/10.1093/cje/bet075

Published: 24 December 2013 **Article history**

Cite | Permissions | Share

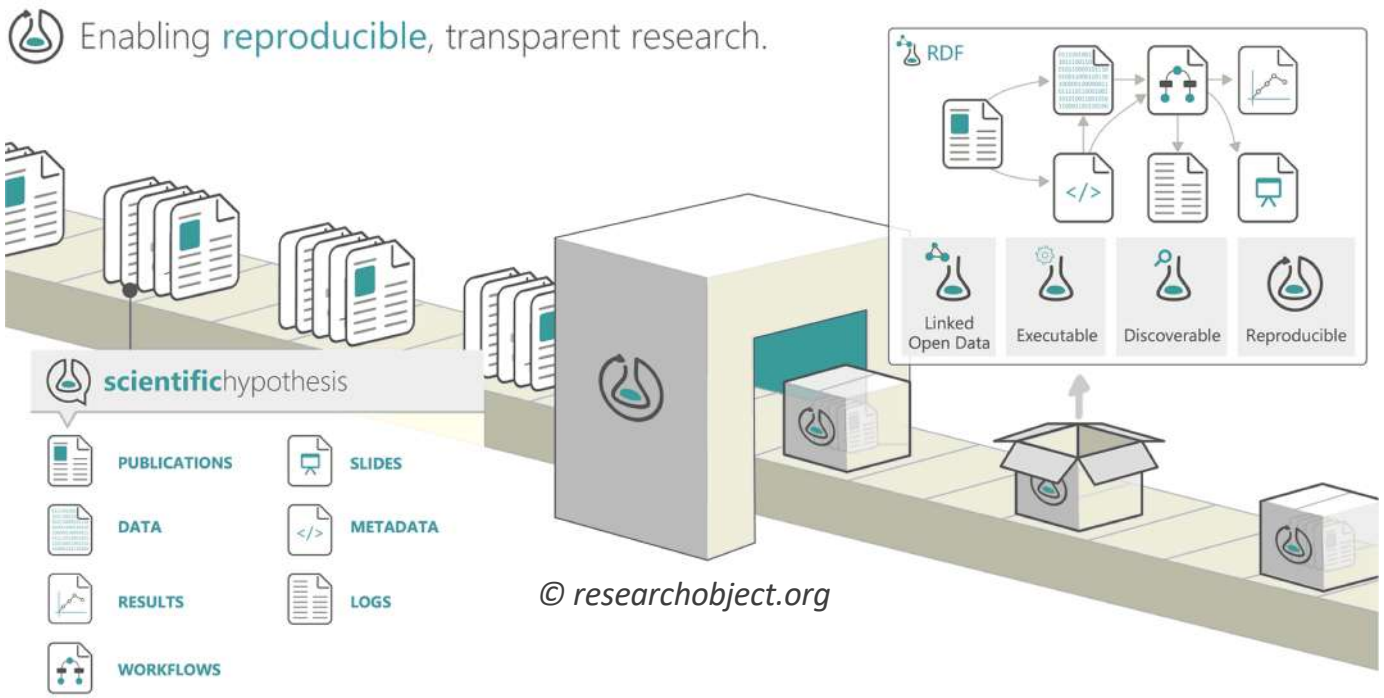
Abstract

We replicate Reinhart and Rogoff (2010A and 2010B) and find that selective exclusion of available data, coding errors and inappropriate weighting of summary statistics lead to serious miscalculations that inaccurately represent the relationship between public debt and GDP growth among 20 advanced economies. Over 1946–2009, countries with public debt/GDP ratios above 90% averaged 2.2% real annual GDP growth, not −0.1% as published. The published results for (i) median GDP growth rates for the 1946–2009 period and (ii) mean and median GDP growth figures over 1790–2009 are all distorted by similar methodological errors, although the magnitudes of the distortions are somewhat smaller than with the mean figures for 1946–2009. Contrary to Reinhart and Rogoff's broader contentions, both mean and median GDP growth when public debt levels exceed 90% of GDP are not dramatically different from when the public debt/GDP ratios are lower. The relationship between public debt and GDP growth varies significantly by period and country. Our overall evidence refutes RR's claim that public debt/GDP ratios above 90% consistently reduce a country's GDP growth.

JEL: E60 - General, E62 - Fiscal Policy, E65 - Studies of Particular Policy Episodes

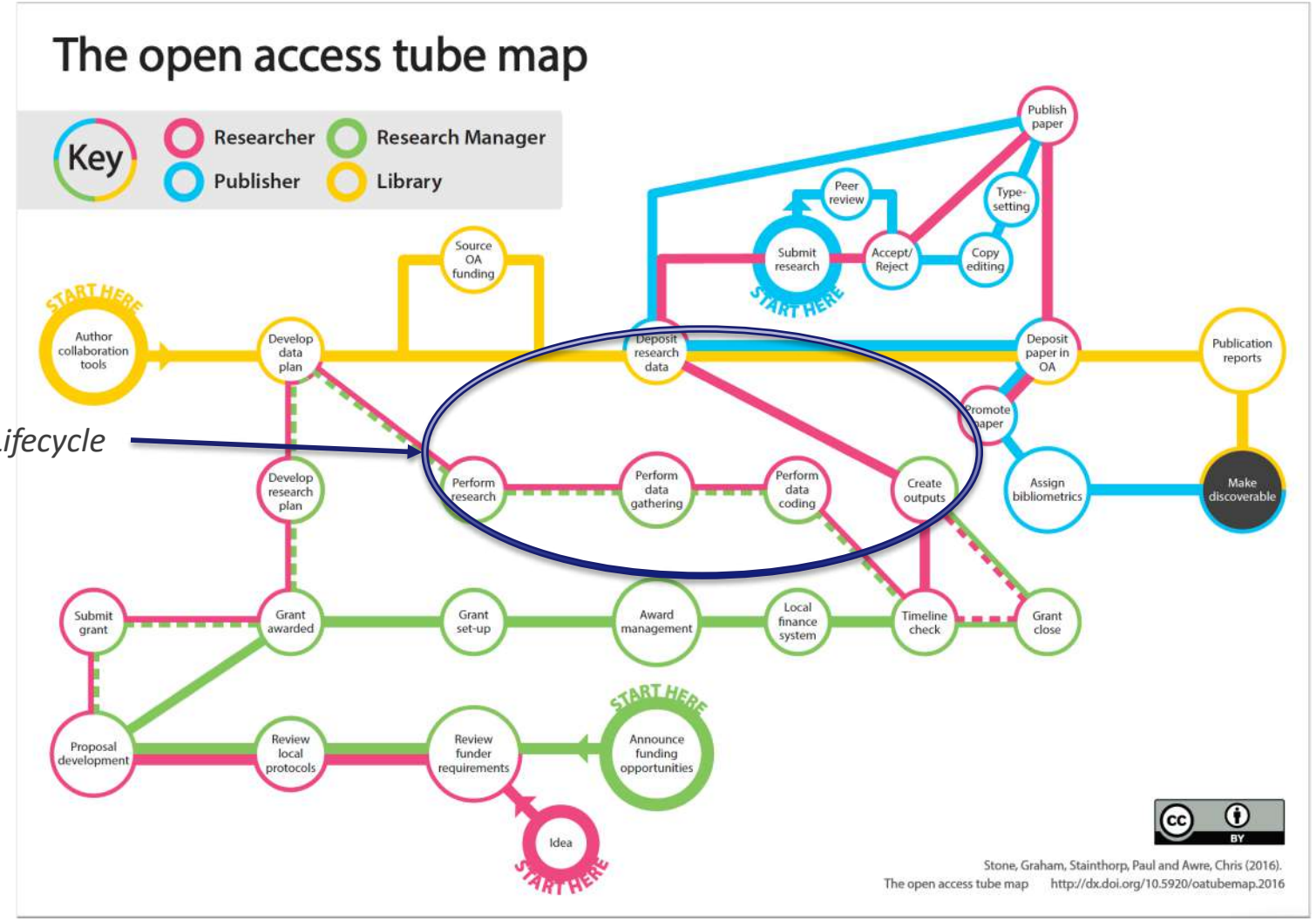
Issue Section: Article

Overcoming the Problem – 1. The Research Object Paradigm



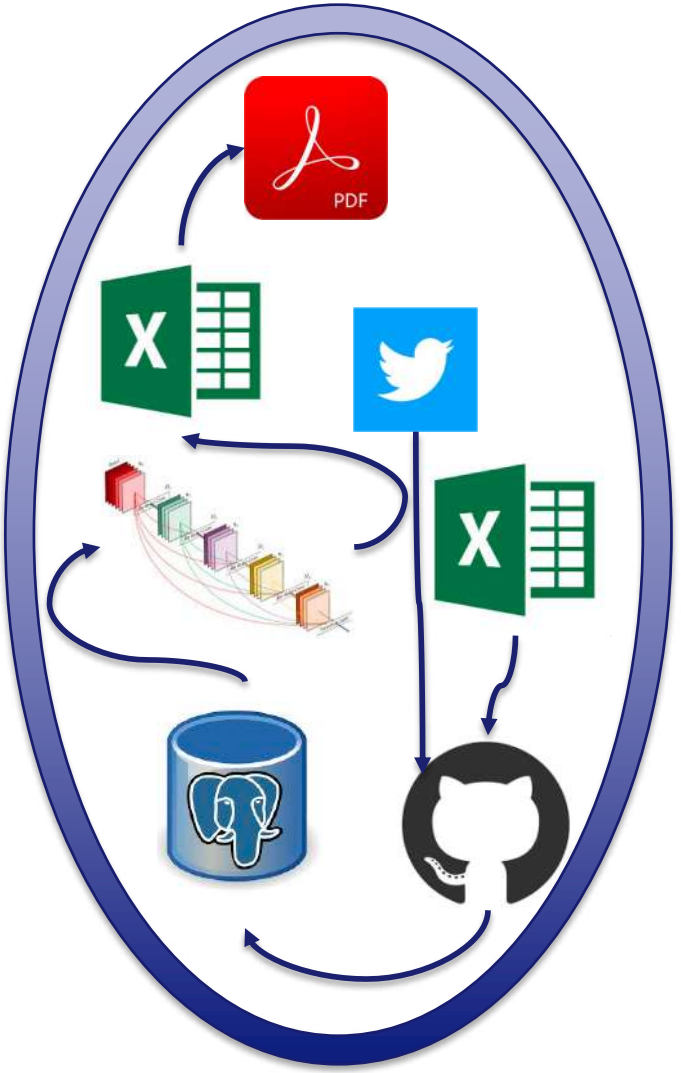
Taking the RO to the Next Level

The Data Lifecycle



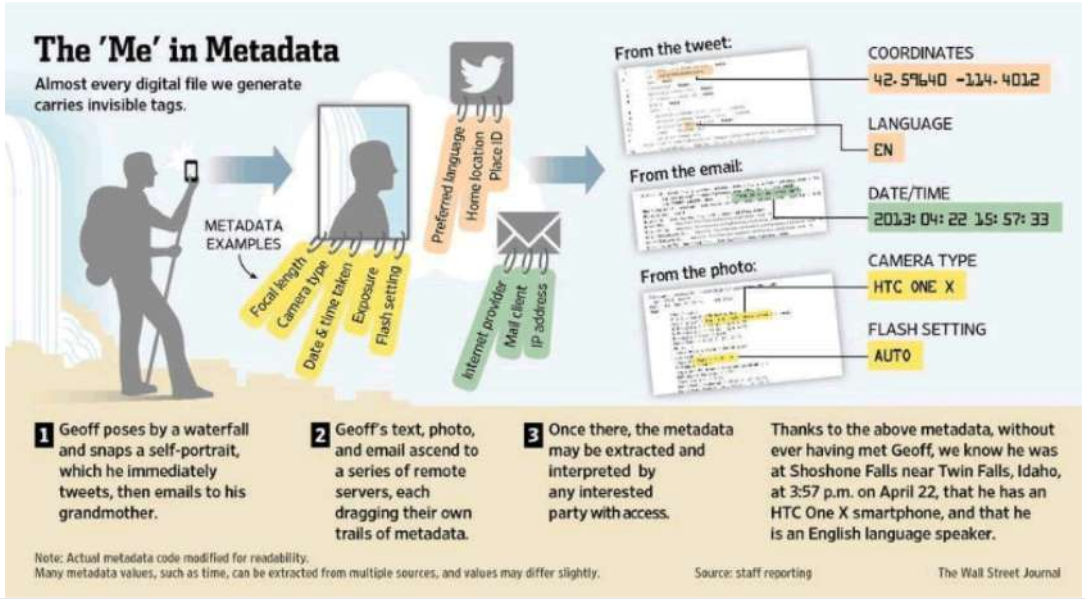
Making the RO Usable

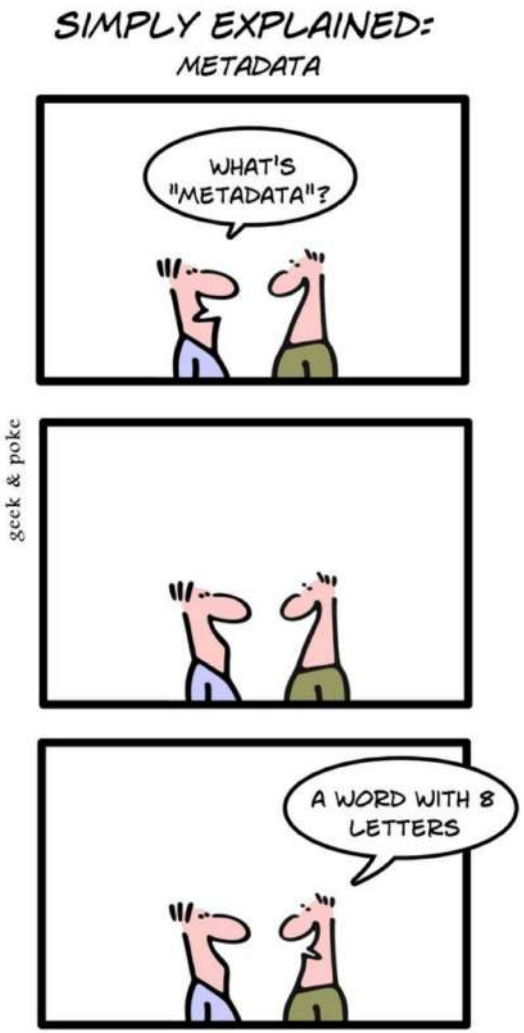
- Linked Open Data and RDF == **PROVENANCE**
- Open Questions…
  - Q: How much detail should I go into
  - A: How reproducible do you want your results to be?
  - Q: This is a lot of effort, why should I do it?
  - A: How else can a user trust your data?
  - Q: My observational outputs are not reproducible – do I still need provenance
  - A: Yes – need to know how you gathered your results
- Exposing Provenance
  - **Make it a part of your metadata**



Metadata – What is it?

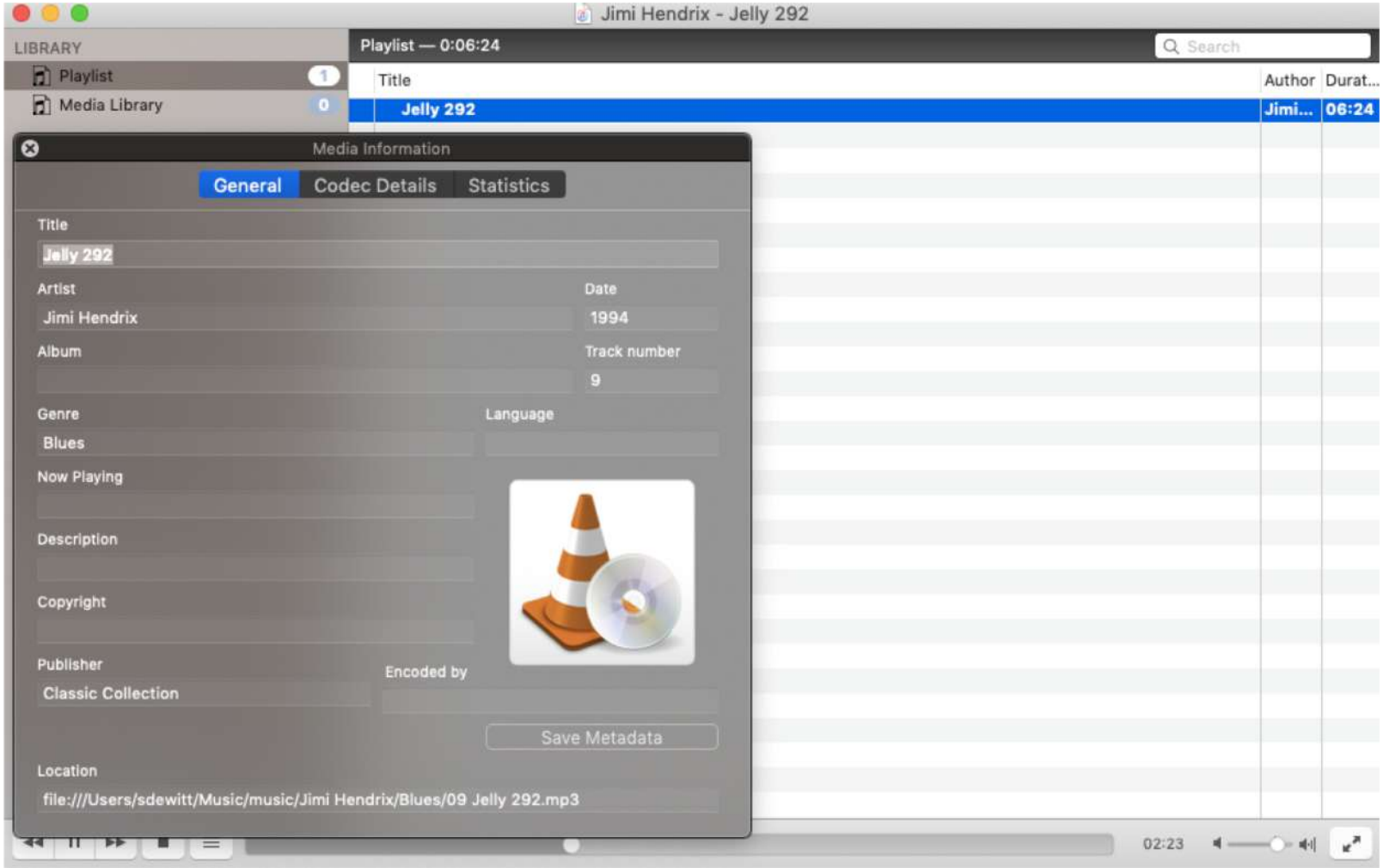
- Data about Data
  - => Metadata **IS** data
- A way of describing a research artefact in a structured, machine readable way
- Something my community or funder insist I have
- "A love note to the future"
- A way of tracking your every movement…





Metadata in Life

Metadata in Life



Metadata Standards - Simple

Dublin Core

- *Title*
- *Subject*
- *Description*
- *Creator*
- *Publisher*
- *Contributor*
- *Date*
- *Type*
- *Format*
- *Identifier*
- *Source*
- *Language*
- *Relation*
- *Coverage*
- *Rights*

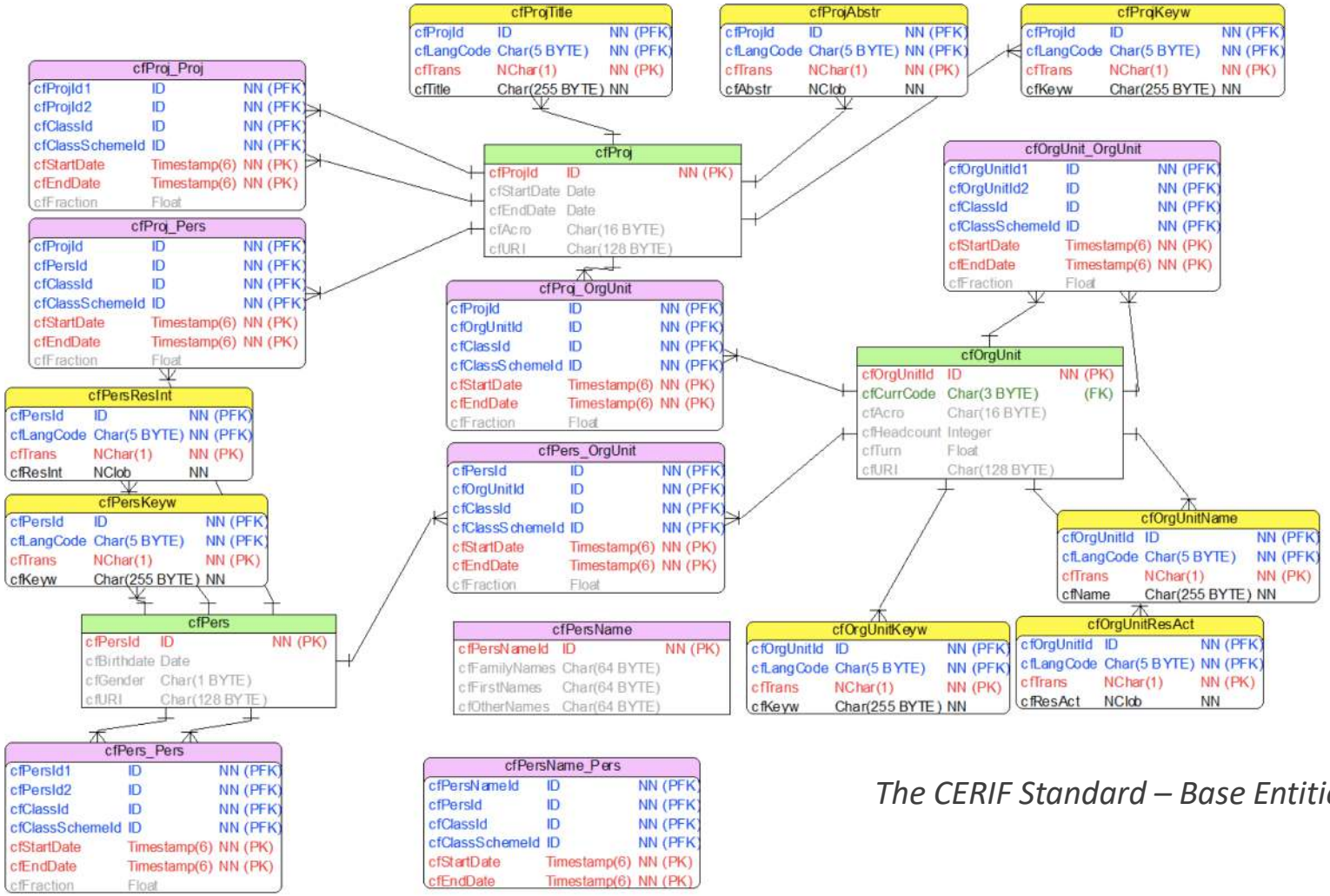
DataCite

- ***Title***
- ***Creator***
- ***Publisher***
- ***Identifer***
- ***Publication Year***
- ***Resource Type***
- ***Subject***
- ***Contributor***
- ***Date***
- ***Related identifier***
- ***Description***
- ***Geolocation***
- *Language*
- *Alternate identifier*
- *Size*
- *Format*
- *Version*
- *Rights*
- *Funding Reference*

EDMI

- ***Name***
- ***Description***
- ***Identifier***
- ***url***
- ***Creator***
- ***Date Created***
- ***license***
- ***Data Standard***
- ***Date Modified***
- ***Access URL***
- ***Access Interface***
- ***Structure***
- ***Included In***
- ***Measurement Technique***
- ***Keywords***
- ***Variable Measured***
- ***Format***
- ***Scientific Type***
- ***Includes***
- ***Content Type***
- ***Size***
- ***Authentications***
- *Version*
- *Metric*
- *Same as*
- *Spatial Coverage*
- *Temporal coverage*
- *Citation*
- *Reference citation*
- *compression*

Metadata Standards - Complex



The CERIF Standard – Base Entities

Exercises in metadata – Guess the film!

Subject	*None*
Description	*Three men searching for stolen gold*
Creator	*Age & Scarpelli, Luciano Vincenzoni*
Publisher	*Produzioni Europee Associate*
Date	*23 December 1966*
Type	*Spaghetti Western*
Format	*35mm anamorphic*
Identifier	*None*
Source	*Original Work*
Language	*Italian, English,Spanish*
Relation	*Part of trilogy*
Coverage	*US Civil War*
Rights	*Produzioni Europee Associate*

Exercises in metadata – Guess the film!

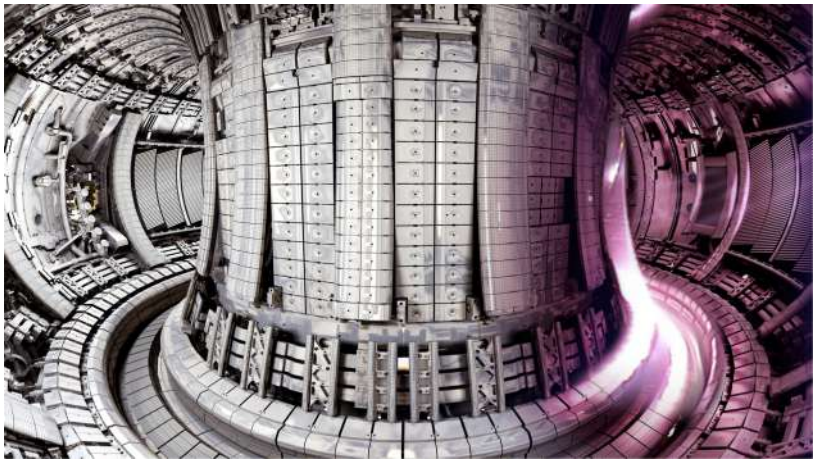
Series	*None*
Cast	*Marlon Brando, Charlie Sheen, Robert Duvali, Dennis Hopper, Harrison Ford*
Credits	*Director: Francis Ford Coppola, Writer: John Millus*
Country	*USA*
Format	*35mm film*
Length	*4205m*
Duration	*2hrs 27 mins*
Language	*English, French, Vietnamese*
Year	*1979*
Identifier	*Italian, English,Spanish*
Genre	*Part of trilogy*
Relation	*US Civil War*
Source	*Produzioni Europee Associate*

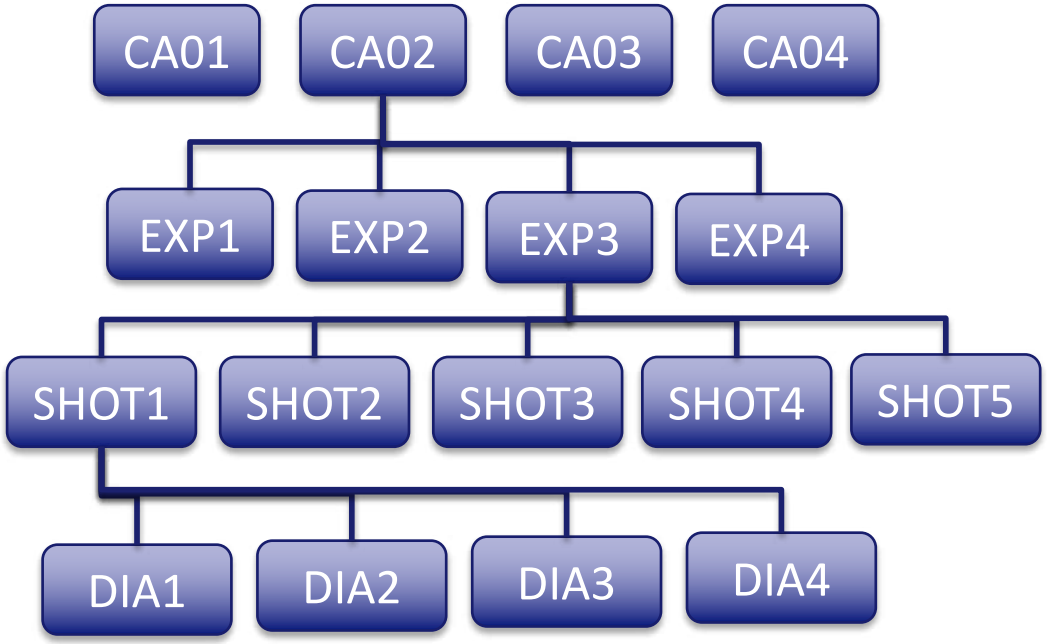
Exercises in Metadata

- http://nsteffel.github.io/dublin_core_generator/
- Use the Simple metadata generator to start with and to and add metadata about yourself!
- If you get time – have a look at Advanced Generator but not essential

Metadata – the Last Word(s)

- FAIR
  - 13 of the 15 FAIR principles mention metadata
  - **It is as important as the data**
- Problems
  - Defining granularity
  - Multi-lingualism
  - Integrating existing schemas
  - Older Standards
- Metadata from different sources can be aggregated
  - Improve interdisciplinary science





Persistent Identifiers

- A way of giving your data a location independent link
- Means if someone cites data, the citation remains the same even if the data underneath moves
- Cans be used to link the data with the metadata (in a PK-FK type relationship)
- Various forms….
  - Archival Resource Key (ARK):
    - http://bnf.fr/ark:/13030/tf5p30086k
  - Persistent Uniform Resource Locators (PURL) – not to be confused with personalised URL
    - https://archive.org/services/purl/purl/Redford-Physics-of-God
  - DOI
    - https://doi.org/10.1109/5.771073
  - International Standard Name Identifier
    - http://isni.org/isni/000000012146438X
- But they all have the same purpose – give a **unique**, **consistent**, **permanent** and **resolvable** id to something… e-mails, addresses etc are ephemeral

Last name of author(s), initial(s) of first name(s) with period. Put & before last author. Separate authors with a comma.

Publication year in brackets.

Use sentence-case for title of article: capitalize first word of title and subtitle only. Capitalize proper nouns, such as companies or place names.

Use italics and capitalize all long words in the journal/source title.

Spitz, D., & Hunter, S. (2005). Contested codes: The social construction of Napster. *Information Society,*

21(3), 169-180. doi: 10.1080/01972240490951890

Digital object identifier (doi) labelled with doi: and placed at end of citation. If needed, remove blue hyperlink.

Italics for volume number, issue number in brackets.

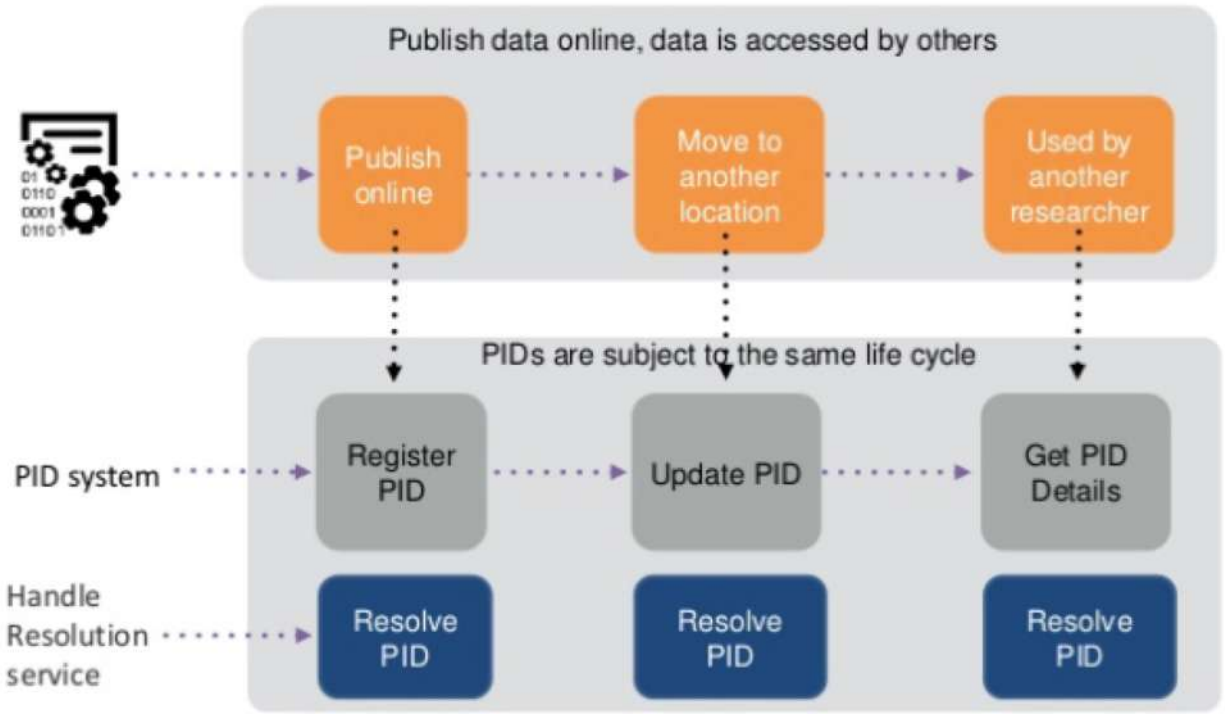
First-page to last -page of article

Great – I can find the paper

Oh – there's o
And they don'

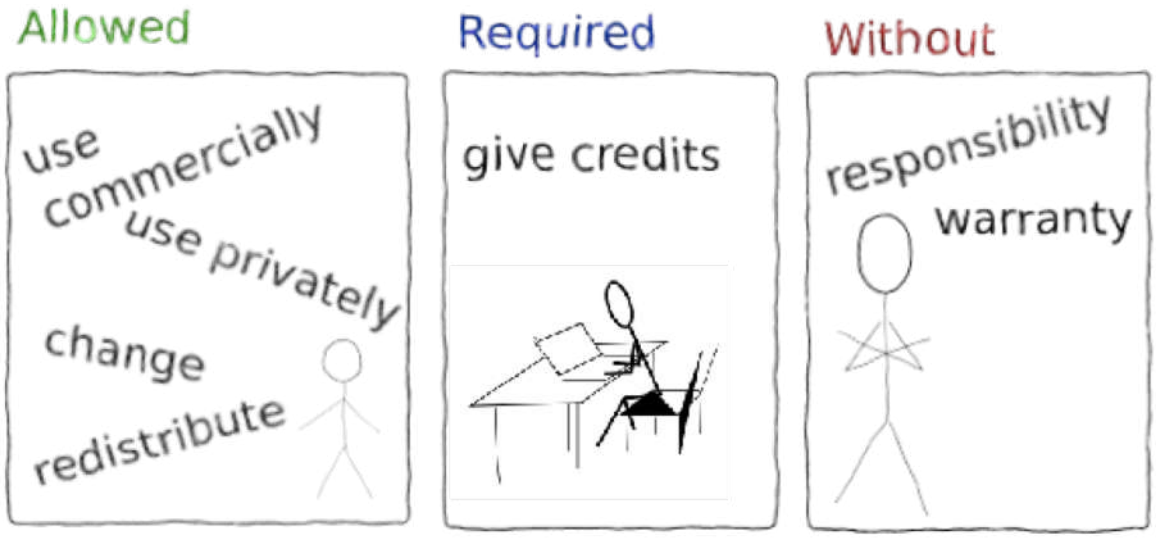
PIDs and Indirections

- All technologies rely on a resolver, and the resolution has to be kept updated – archivist/data manager/ data steward



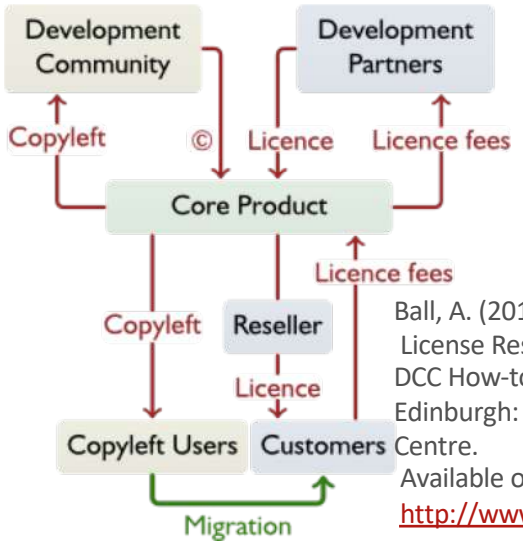
Licensing – the last but most important thing you will do

- Why license
  - Legal protection
  - Allow users to understand what they can and con not do with the data
  - Make sure you get credit
- But which license?
  - B2SHARE can help – it has a tool to help you select your license



Examples of licenses commonly used for open data

- Open licenses
  - CC-0
  - WTFPL
  - Unlicense
  - PDDL
- Permissive Licenses
  - BSD
  - MIT
  - Creative Commons
  - EUPL v1.2
- Copyleft license
  - GPL, LGPL, AGPL
  - Common Development & Distribution License
- Specific Licenses
  - IES Restricted Data License
  - MetaShare no Redistribution License

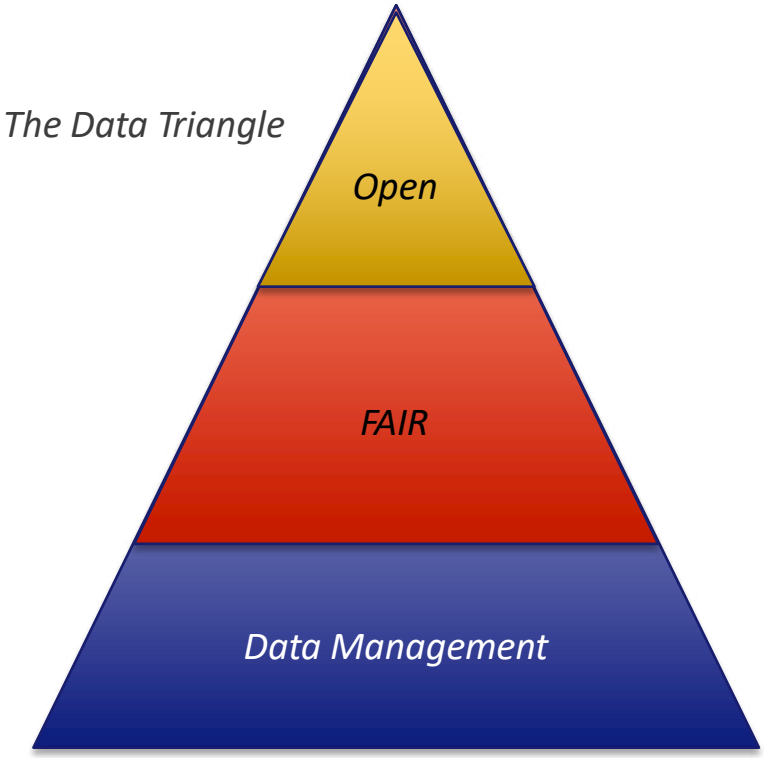


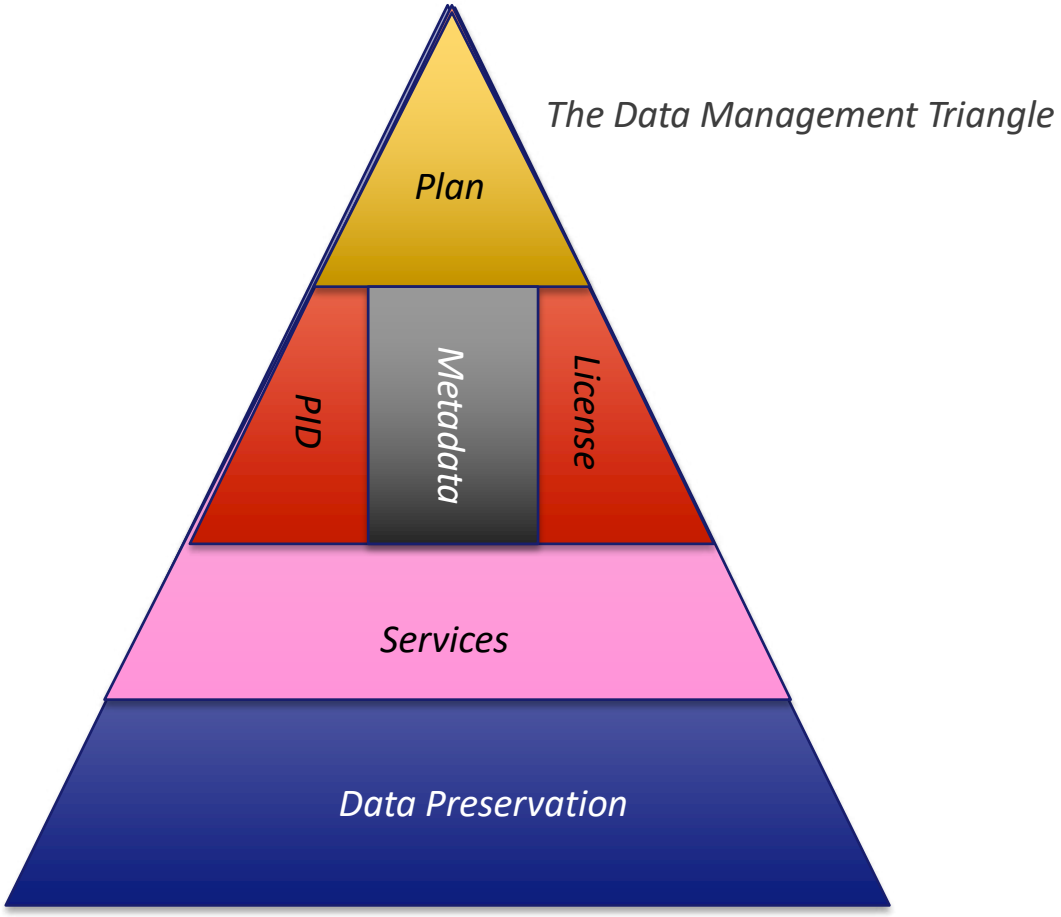
Ball, A. (2014). 'How to License Research Data'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

Summary

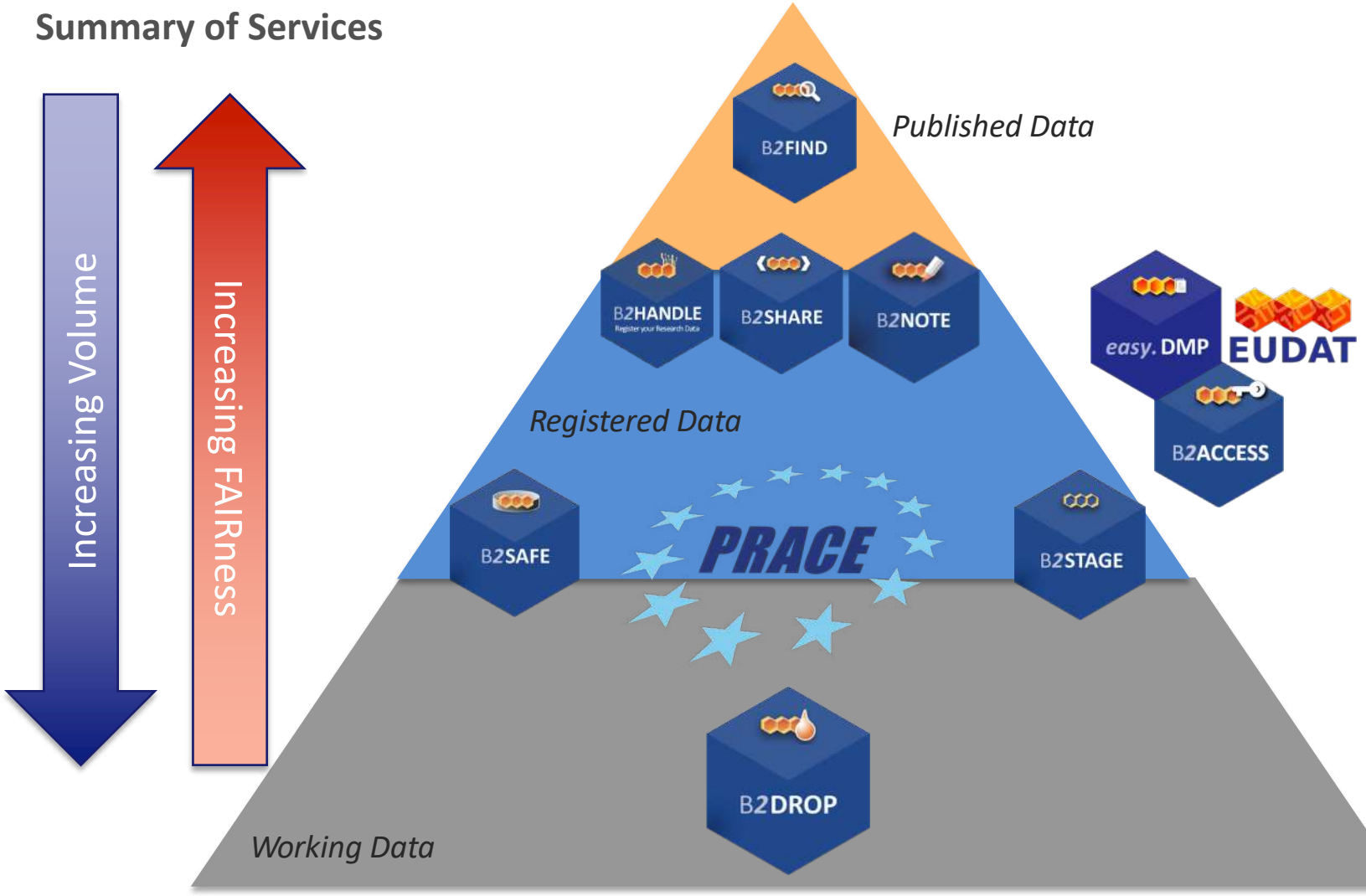
- **It is ridiculously easy to lose data**
- **Data Preservation** is a subset of **Information Preservation**
  - With overlap when **common standards** are used
- Various **Technologies** and **Procedures** can help you with data preservation
  - **None** are foolproof
  - AWS offer **99.999% availability** annual failure rate (AFR) of between **0.1%-0.2%.**
    - "We have no liability whatsoever for any damages, liabilities, losses (including any corruption, deletion, or destruction or loss of data, applications or profits)…"
- **Checksumming** (a.k.a. hashing or fixity checking) used to detect changes in files)
  - Very good but can be subject to malicious attack
- **Metadata** and **Persistent Identifiers** needed for both data preservation and information preservation
  - **Provenance** for repeatability and trust

Summary…


The Data Triangle
Open
FAIR
Data Management


The Data Management Triangle
Plan
PID
Metadata
License
Services
Data Preservation

Summary of Services



The Rest of The Week

- Introduction to some of the services…
  - B2DROP and B2SHARE for sharing your data and making it public
  - B2FIND and making your data discoverable
  - GridFTP – the interface between EUDAT and PRACE
- Taking you through a real use case from ENES Climate Analysis Service (ECAS)