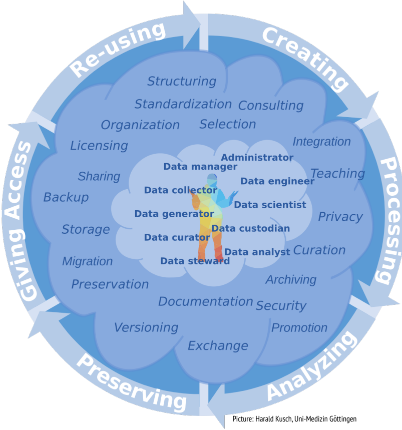
Research Data Management

HPC Day for Physicists



Picture: Harald Kusch, Uni-Medizin Göttingen

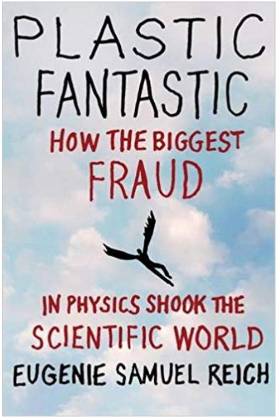
Jörg Steinkamp

Zentrum für **Datenverarbeitung**

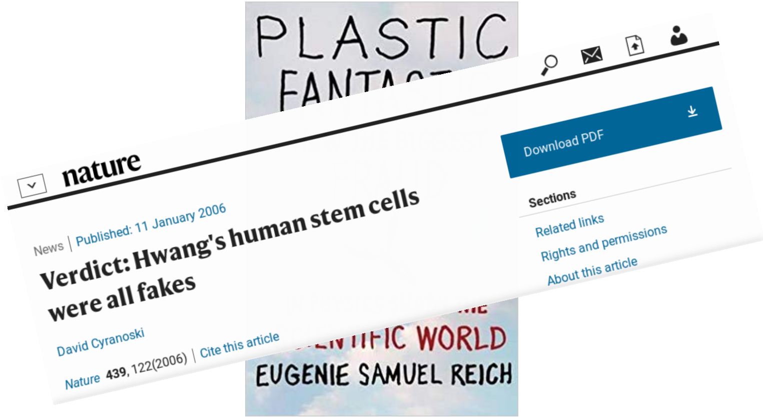
Content

1. Motivation
2. Archiving
3. Data life cycle
4. FAIR principles
5. Tools for RDM

Why care about RDM



Why care about RDM



nature

News | Published: 11 January 2006

Verdict: Hwang's human stem cells were all fakes

David Cyranoski

Nature **439**, 122(2006) | Cite this article

Download PDF

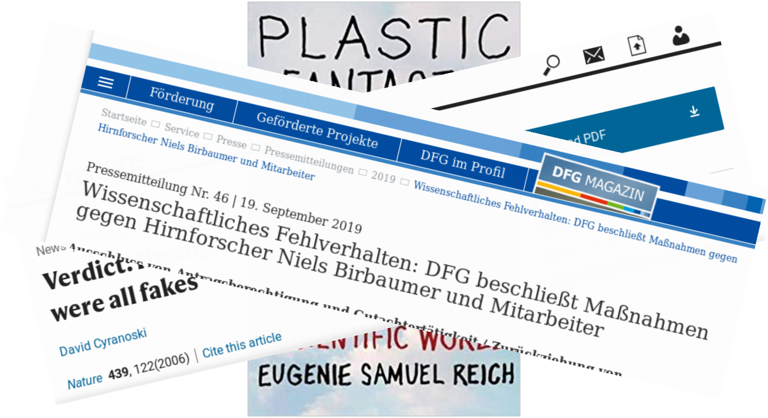
Sections

Related links

Rights and permissions

About this article

Why care about RDM



PLASTIC FANTASTIC

SCIENTIFIC WORLD

EUGENIE SAMUEL REICH

Startseite □ Service □ Presse □ Pressemitteilungen □ 2019 □ Wissenschaftliches Fehlverhalten: DFG beschließt Maßnahmen gegen Hirnforscher Niels Birbaumer und Mitarbeiter

Pressemitteilung Nr. 46 | 19. September 2019

Wissenschaftliches Fehlverhalten: DFG beschließt Maßnahmen gegen Hirnforscher Niels Birbaumer und Mitarbeiter

Ausschluss von Antragsberechtigung und Gutachtertätigkeit / Zurückziehung von

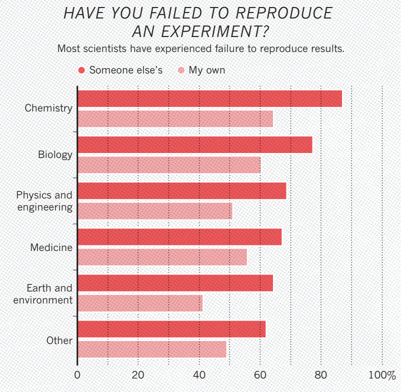
News

Verdict: were all fakes

David Cyranoski | Cite this article

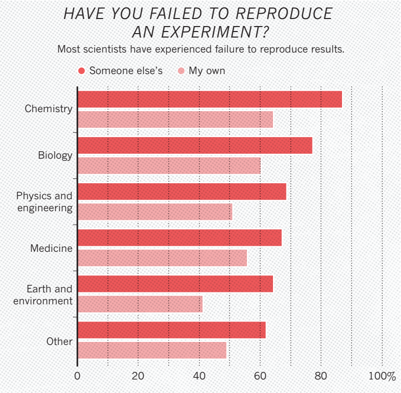
Nature 439, 122(2006)

Why care about RDM

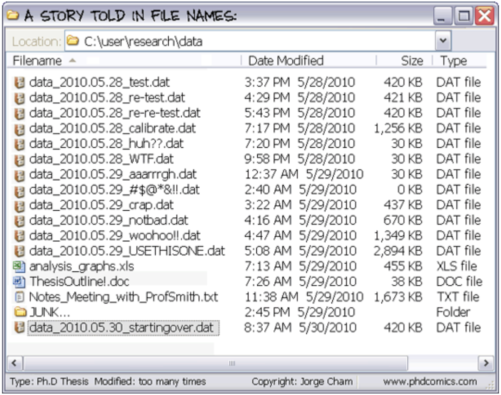


Baker (2016), Nature.

Why care about RDM



Baker (2016), Nature.



Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

Jorge Cham www.phdcomics.com

Why care about RDM

Everybody can handle order, but only a genius can master chaos, ...

Why care about RDM

Everybody can handle order, but only a genius can master chaos, ...

... but what happens when the genius leaves?!

Why care about RDM

Everybody can handle order, but only a genius can master chaos, ...

... but what happens when the genius leaves?!

Why care about RDM

Researchers

- minimize risk of data loss
- maximize ...
  - ... efficiency
  - ... sustainability
  - ... reproducibility
- facilitate ...
  - ... teamwork
  - ... follow-up projects
- increase reputation
- beeing nice to following scientists

Why care about RDM

Researchers

- minimize risk of data loss
- maximize ...
  - ... efficiency
  - ... sustainability
  - ... reproducibility
- facilitate ...
  - ... teamwork
  - ... follow-up projects
- increase reputation
- beeing nice to following scientists

System Administrators

- minimize ...
  - ... risk of data loss
  - ... costs
- maximize ...
  - ... efficiency
  - ... sustainability
- avoid patchwork rug
- provide central on-site storage
- simplify scientific workflows
- beeing nice to all scientists

Archiving

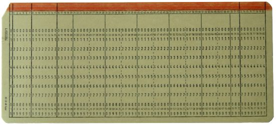
2 Archiving
- Backup vs. Archive
- Money, money, money

3 Data life cycle

4 FAIR principles

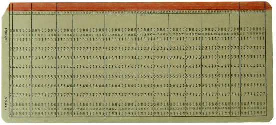
5 Tools for RDM

Backup isn’t Archiving and Archiving isn’t Backup



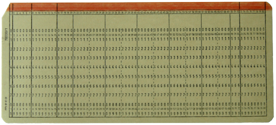
	Backup	**Archive**
target	active data	final data
purpose		
timeframe		
metadata		

Backup isn’t Archiving and Archiving isn’t Backup



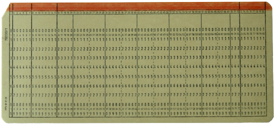
	Backup	Archive
target	active data	final data
purpose	protection and recovery	preservation of information
timeframe		
metadata		

Backup isn’t Archiving and Archiving isn’t Backup



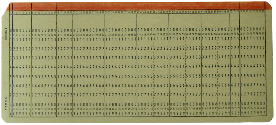
	Backup	Archive
target	active data	final data
purpose	protection and recovery	preservation of information
timeframe	short term	long term
metadata		

Backup isn't Archiving and Archiving isn't Backup



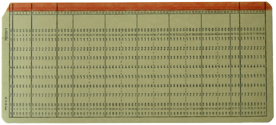
	Backup	Archive
target	active data	final data
purpose	protection and recovery	preservation of information
timeframe	short term	long term
metadata	mainly ACL and file attributes	data-specific

Archiving medium



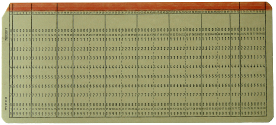
Medium	logevity	HW cost per PB	Total cost per year

Archiving medium



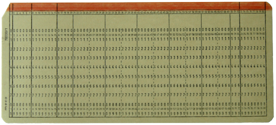
Medium	logevity	HW cost per PB	Total cost per year
SSD	100 yr	100.000€	690.000$

Archiving medium



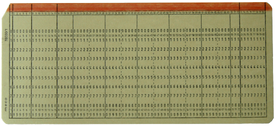
Medium	logevity	HW cost per PB	Total cost per year
HDD	3–5 yr	32.000€	555.000$
SSD	100 yr	100.000€	690.000$

Archiving medium



Medium	logevity	HW cost per PB	Total cost per year
Tape	30 yr	6.000€	65.000$
HDD	3–5 yr	32.000€	555.000$
SSD	100 yr	100.000€	690.000$

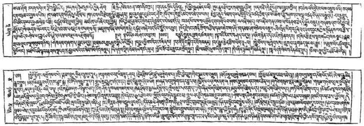
Archiving medium

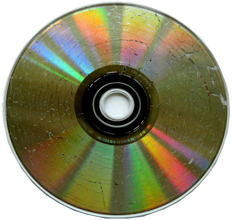


Medium	logevity	HW cost per PB	Total cost per year
Tape	30 yr	6.000€	65.000$
HDD	3–5 yr	32.000€	555.000$
SSD	100 yr	100.000€	690.000$
DVD	10–25 yr	60.000€	–

Archiving: Disc in Desk

Image stored on a CD-R

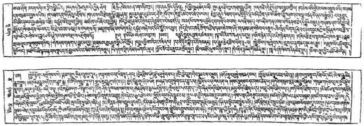




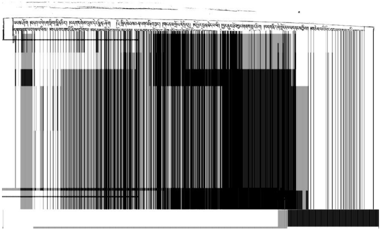
@EdithHalvarsson

Archiving: Disc in Desk

Image stored on a CD-R



Still readable, but ...

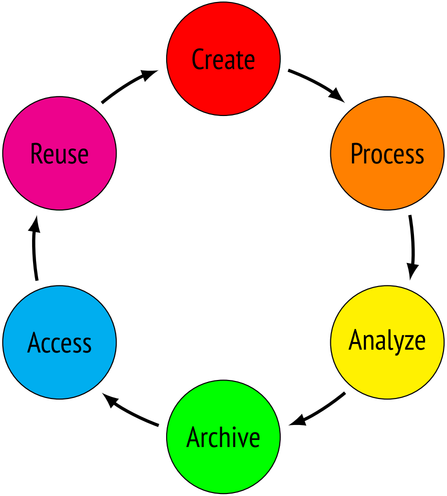


@EdithHalvarsson

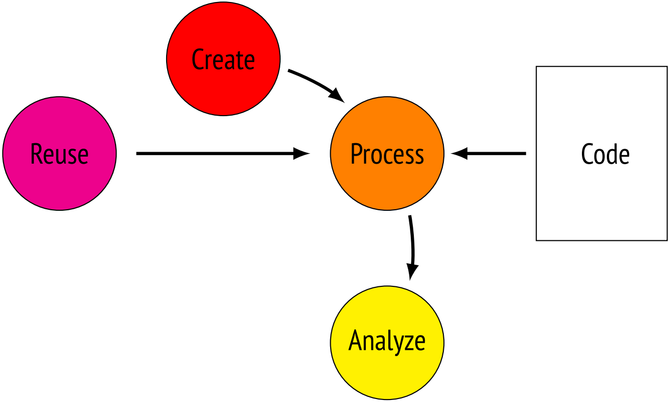
Data life cycle

1. Motivation
2. Archiving
3. Data life cycle
   - Scientific segment
   - Administrative segment
4. FAIR principles
5. Tools for RDM

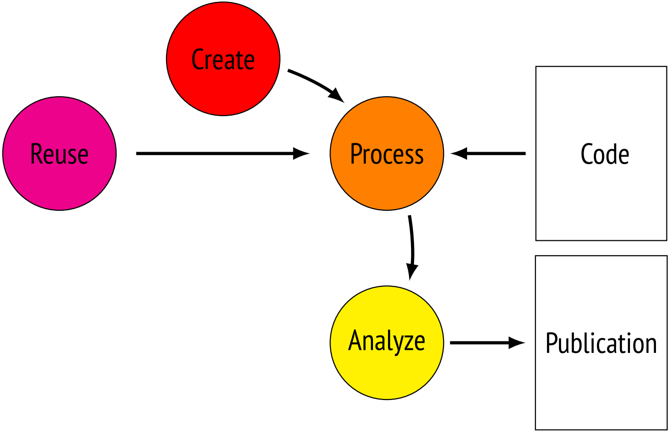
Classic data life cycle



Scientific part of the data life cycle



Scientific part of the data life cycle



Create

Reuse

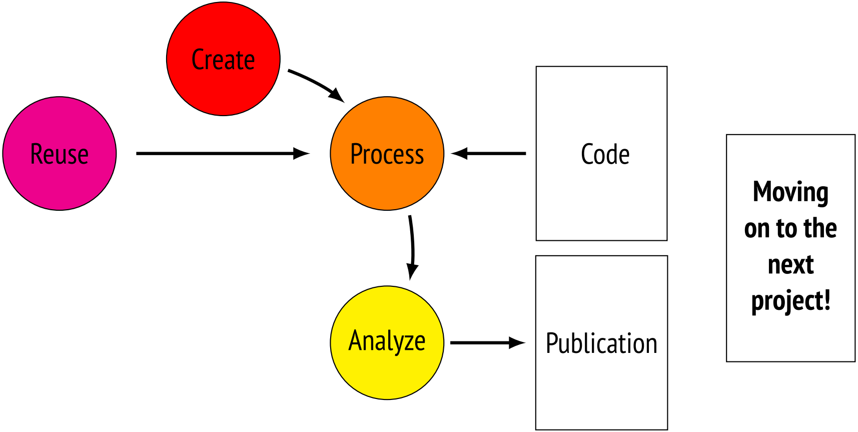
Process

Code

Analyze

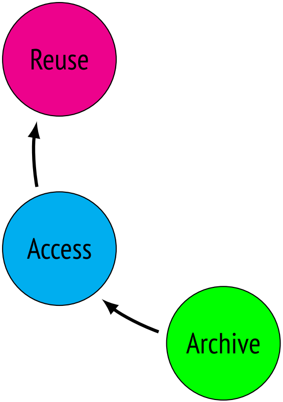
Publication

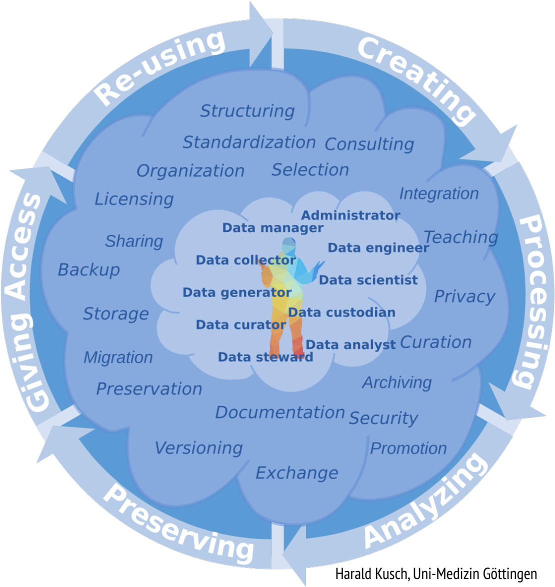
Scientific part of the data life cycle



Moving on to the next project!

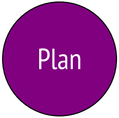
Getting lost in the Research Data Management Cloud





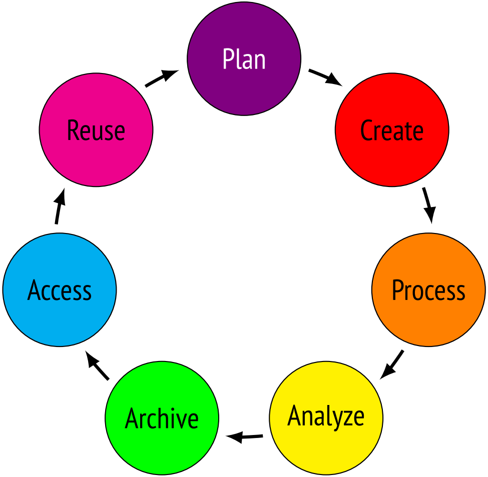
Harald Kusch, Uni-Medizin Göttingen

Don’t forget the Planning

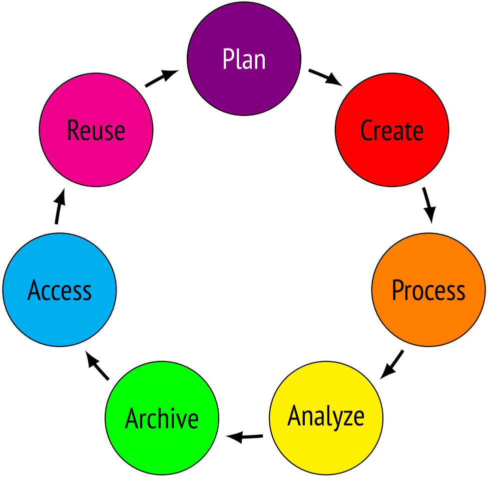


Plan

Don’t forget the Planning



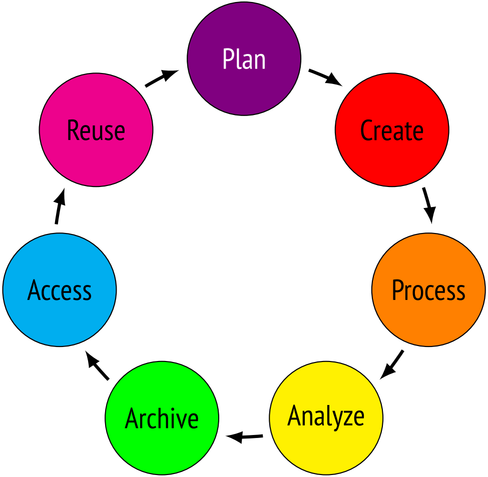
Don’t forget the Planning



Data Management Plan

- dynamic document

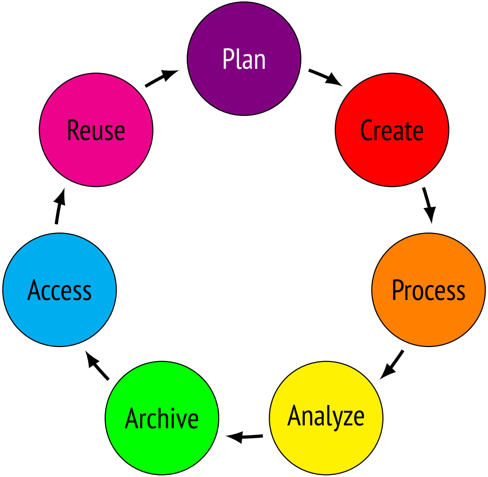
Don’t forget the Planning



Data Management Plan

- dynamic document
- JGU-RDMO ↗

Don’t forget the Planning



Data Management Plan

- dynamic document
- JGU-RDMO
- other Tools:
  - DMPonline
  - DMPTool

Consultation

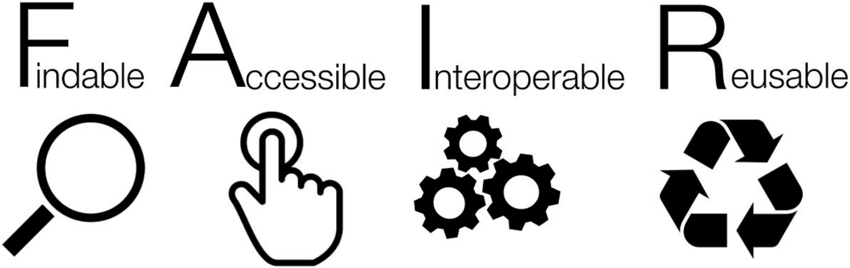
- HPC: Christian Meesters ✉
- ZDV: Jörg Steinkamp/ Sarah Wettermann ✉
- Administrativ: Anne Vieten ✉
- Other Universities

FAIR principles

1. Motivation
2. Archiving
3. Data life cycle
4. FAIR principles
   - Keywords
   - Publish
   - Standardized
   - Reusable
5. Tools for RDM

geralt@Pixabay.com

What FAIR stands for



Findable by keywords

Sufficient rich metadata

Minimum set

- Creator
- Title
- Date
- Location
- Publisher
- Keywords
- ...

Findable by keywords

Sufficient rich metadata

Minimum set

- Creator
- Title
- Date
- Location
- Publisher
- Keywords
- ...

Extended metadata

- NFDI

Findable by keywords

Sufficient rich metadata

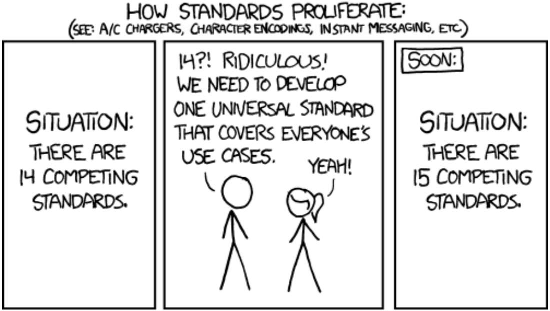
Minimum set

- Creator
- Title
- Date
- Location
- Publisher
- Keywords
- ...

Extended metadata

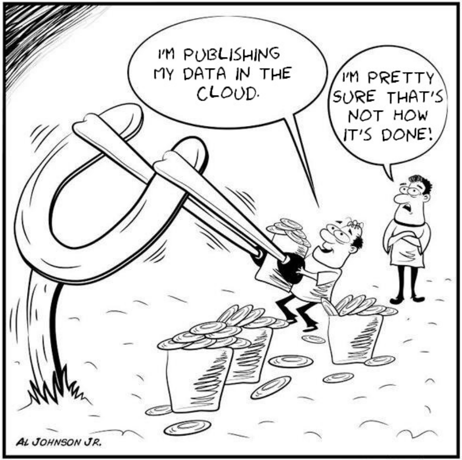
- NFDI
- Other/general standards:
  - RADAR ↗
  - Dublin Core ↗
  - Data Cite ↗
  - Disciplinary Metadata ↗
  - Research Data Management Toolkit: Metadata Standards ↗
  - Schema.org ↗
  - Bioschemas.org ↗

Findable by keywords


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

XKCD

Let others Access your (meta-)data


I'M PUBLISHING MY DATA IN THE CLOUD.
I'M PRETTY SURE THAT'S NOT HOW IT'S DONE!
AL JOHNSON JR.

modified after @Cloudways

©CloudTweaks.com

⚠ **The Cloud is not an archive**

- ~~Dropbox/Box~~
- ~~Google Drive~~
- ~~Owncloud/Nextcloud~~
- ~~Seafile~~
- ~~Amazon Drive~~
- ~~Microsoft Azure~~
- ...

3rd party repositories

General repositories

- DRYAD
- figshare
- + zenodo

3rd party repositories

General repositories

- DRYAD
- figshare
- + zenodo

Lists of subject specific repositories

- Subject specific at SCIENTIFIC DATA
- Search engine re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES
- ...

Interoperable: Let others be able to read your data

General

- unencrypted
- (uncompressed)
- open, documented standard
- non-proprietary non-patented

Interoperable: Let others be able to read your data

General

- unencrypted
- (uncompressed)
- open, documented standard
- non-proprietary non-patented

However, ...

- docx, xlsx are acceptable
- raw device data

Reusable: Allow/Enable others to use it

General

- Provide a license
- add subject-relevant metadata
- describe the source/origin of the data

Tools for RDM

3 Data life cycle

4 FAIR principles

5 Tools for RDM
- Archiving

Archiving

Archiving

TSM (Tape Library)

- Two copies at different location
- Encrypted Tapes
- No need to care about ...
  - ... access control
  - ... reuse/-ability
  - ... metadata

Archiving

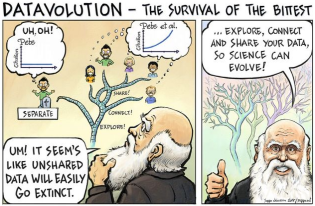
TSM (Tape Library)

- Two copies at different location
- Encrypted Tapes
- No need to care about ...
  - ... access control
  - ... reuse/-ability
  - ... metadata

iRODS

- Two copies at different location
- Encrypted HDDs
- You benefit from ...
  - ... access control
  - ... attached metadata
  - ... publication possibility

publicdomainvectors.org


DATAVOLUTION – THE SURVIVAL OF THE BITTEST
UH, OH!
Pete
Citation
Pete et al.
Citation
SEPARATE
SHARE!
CONNECT!
EXPLORE!
UM! IT SEEM'S LIKE UNSHARED DATA WILL EASILY GO EXTINCT.
... EXPLORE, CONNECT AND SHARE YOUR DATA, SO SCIENCE CAN EVOLVE!

Cartoon: Seppo Leinonen, www.seppo.net

iRODS: **i**ntegrated **R**ule-**O**riented **D**ata **S**ystem

What is iRODS?

- Virtual Filesystem

iRODS: **i**ntegrated **R**ule-**O**riented **D**ata **S**ystem

What is iRODS?

- Virtual Filesystem
- Attached searchable Metadata

iRODS: **i**ntegrated **R**ule-**O**riented **D**ata **S**ystem

What is iRODS?

- Virtual Filesystem
- Attached searchable Metadata
- Publishing

iRODS: **i**ntegrated **R**ule-**O**riented **D**ata **S**ystem

What is iRODS?

- Virtual Filesystem
- Attached searchable Metadata
- Publishing
- Fine grained Access Control

iRODS: **i**ntegrated **R**ule-**O**riented **D**ata **S**ystem

What is iRODS?

- Virtual Filesystem
- Attached searchable Metadata
- Publishing
- Fine grained Access Control
- Workflow automation

iRODS: **i**ntegrated **R**ule-**O**riented **D**ata **S**ystem

What is iRODS?

- Virtual Filesystem
- Attached searchable Metadata
- Publishing
- Fine grained Access Control
- Workflow automation
- Commandline tools

iRODS: **i**ntegrated **R**ule-**O**riented **D**ata **S**ystem ↗

What is iRODS?

- Virtual Filesystem
- Attached searchable Metadata
- Publishing
- Fine grained Access Control
- Workflow automation
- Commandline tools
- (WebUI)

Upcoming courses and further information

Further links

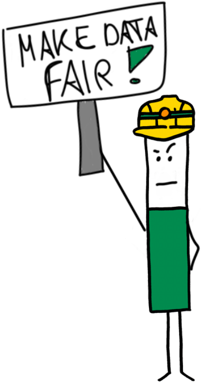
- Kompetenzteam FDM
- Archiving @ Mogon-Docs

Courses

- HPC-related courses
- Git Courses Registration

Thank you

Questions?



Sketchnotes & Sketches – FranziMachtDas