



CC BY 4.0

FAIR Principles for Research Data Management and Stewardship

Pinar Alper



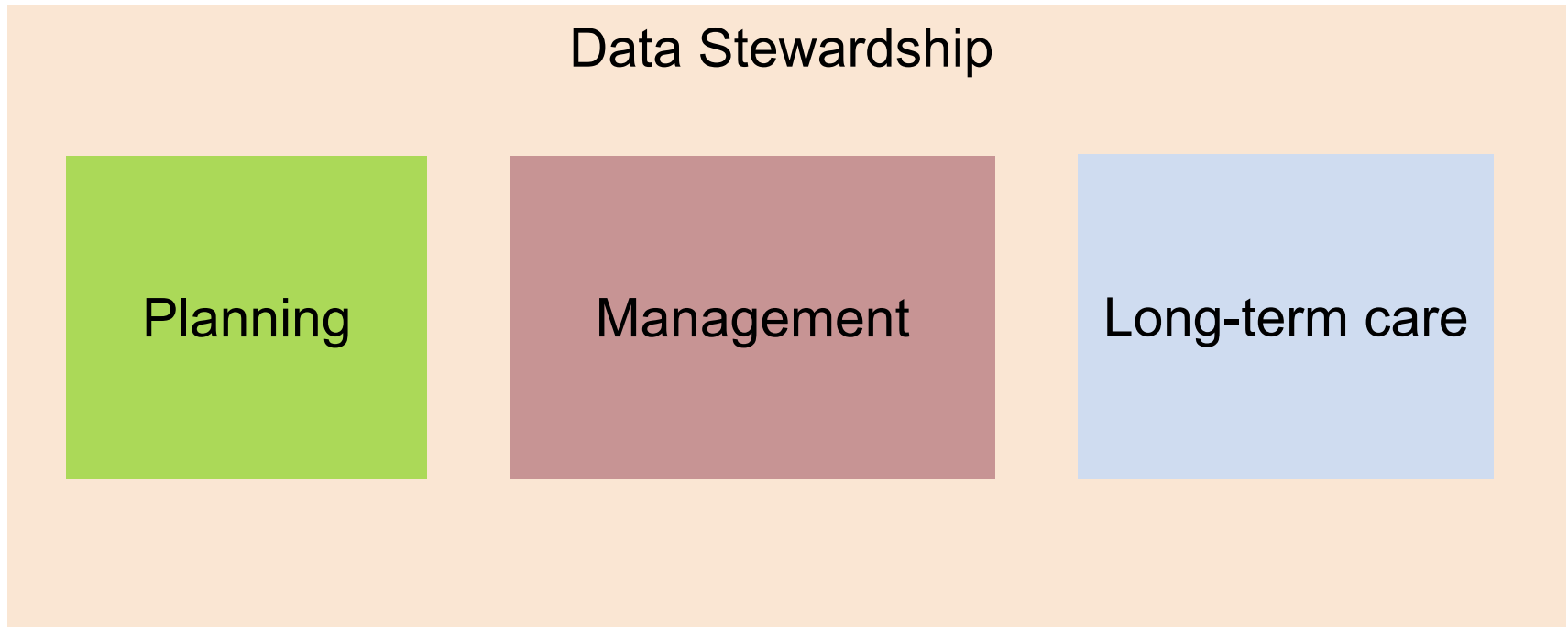
*Training on Research Data Management
25-26 June 2019
Luxembourg Learning Centre*

Definitions

" Research data management (RDM) concerns the organisation of data, from its entry to the research cycle through to the dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and permits new and innovative research built on existing information. "

- Examples:
 - Day to day data handling during project , e.g. using consistent file naming conventions.
 - Preserving and sharing after the project completion e.g. depositing the data in a community repository.

Data Stewardship = RDM ++



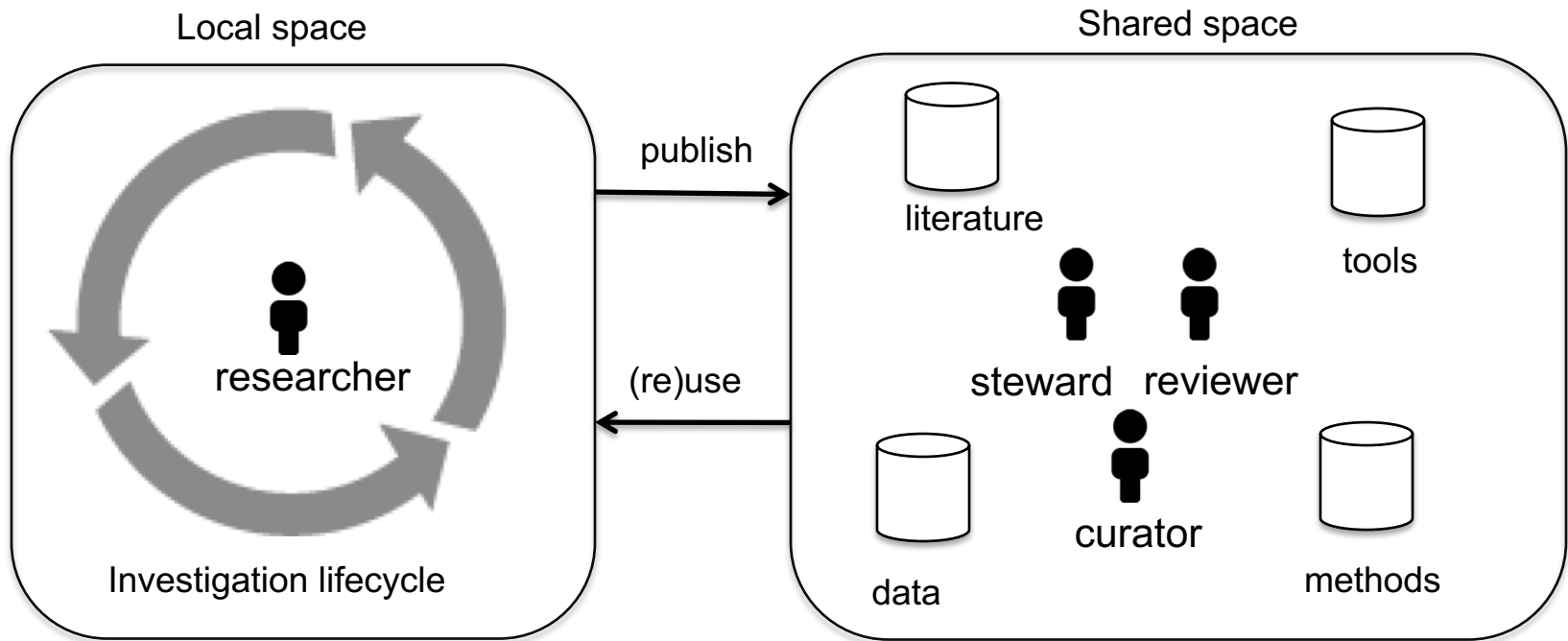
Researcher



Data specialist

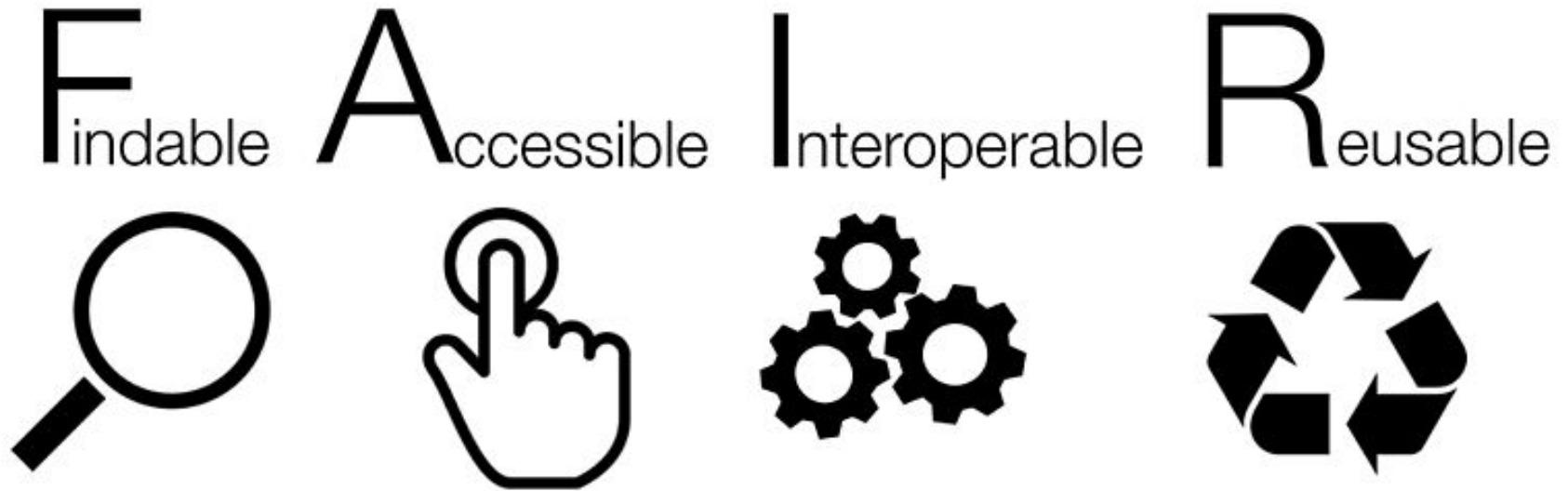
Data Stewardship

— aims to enable long term care and re-use of data

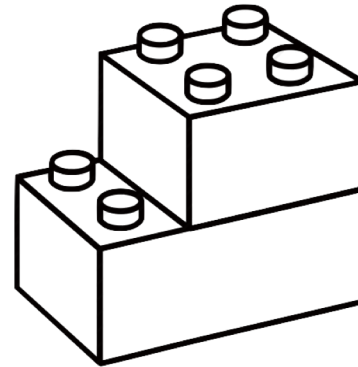
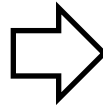


End goal

— FAIR principles for research data



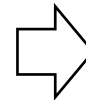
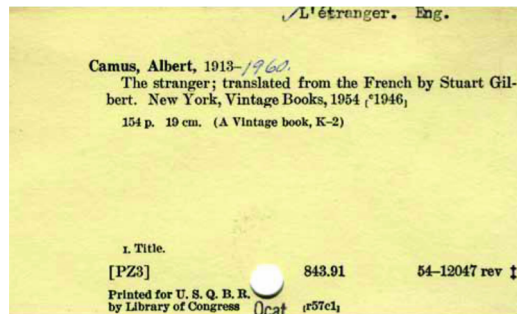
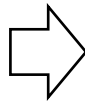
<https://www.youtube.com/watch?v=MSasvoa0dV4>



FAIR principles for research data

Findable

- (Meta)data
- Identifiers for (meta)data
- Indexed in a searchable resource



Rare gene deletions in genetic generalized and Rolandic epilepsies

Jabbari K, Bobbili DR, Lal D, Reinthaler EM, Schubert J, et al. (2018) Rare gene deletions in genetic generalized and Rolandic epilepsies. PLOS ONE 13(8): e0202022.

<https://doi.org/10.1371/journal.pone.0202022>

FAIR principles for research data

Findable

- Metadata
- Identifiers for (meta)data
- Indexed in a searchable resource

Hibsh, D., Schori, H., Efroni, S. & Shefi, O. *Figshare*
<http://dx.doi.org/10.6084/m9.figshare.1289242> (2015).

NCBI Sequence Read Archive [SRP059260](#) (2015).



Epi4K: Gene Discovery in 4,000 Epilepsy Genomes

dbGaP Study Accession: phs000653.v3.p1

Study Variables Documents Analyses Datasets Molecular Data

Dataset Name and Accession

Dataset Name: Epi4K_Subject_Phenotypes

Dataset Accession: pht008354.v1.p1

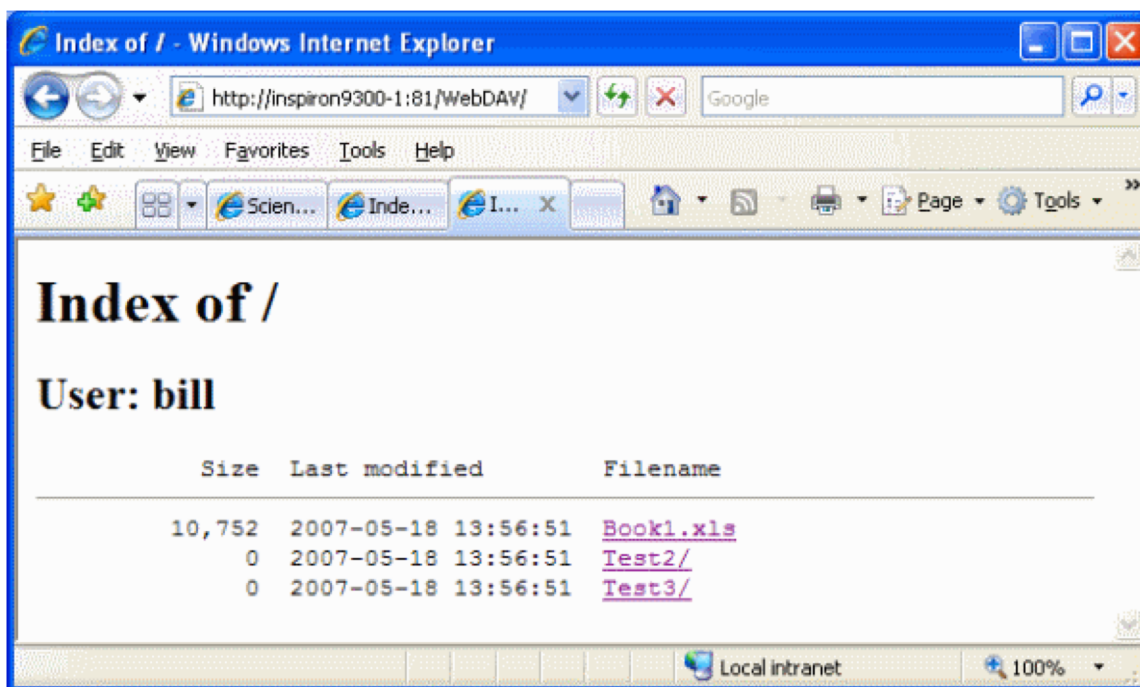
Dataset Description

Phenotype Data (All Sub-Studies of phs000653): Includes description of epilepsy category, e.g. infantile spasms, Lennox-Gastaut syndrome, generalized idiopathic epilepsy, focal epilepsies, etc., but also healthy parents, etc, plus gender and ethnicity data.

FAIR principles for research data

Not Findable

- Metadata
- Identifiers for (meta)data
- Indexed in a searchable resource



FAIR principles for research data

Accessible

- (Meta)data are retrievable by a protocol
- Open, free universally implementable
- Authentication/Authorization
- Metadata available even when data is not



Hibsh, D., Schori, H., Efroni, S. & Shefi, O. *Figshare*
<http://dx.doi.org/10.6084/m9.figshare.1289242> (2015).



HOME | HANDBOOK | FACTSHEETS | FAQs | RESOURCES | USERS | NEWS | MEMBERS AREA

Resolve a DOI Name

doi:

Go

A DOI is a unique persistent identifier for a published digital object

FAIR principles for research data

Accessible

- (Meta)data are retrievable by a protocol
- Open, free universally implementable
- Authentication/Authorization
- Metadata available even when data is not



10.1004/123456	URL	http://www.pub.com/
	URL	http://www.pub2.com/

Moved data

Data versions



The Y4 Seismic Network, 2014–2015

Network code: Y4
 Restricted: No
 Network KML file: [K](#)
 Seismic metadata: [fdsnws-station](#)
 Institution(s): GFZ
 Creator(s): Roessler, Dirk; Passarelli, Luigi; Govoni, Aladino; Bautz, Ralf; Dahm, Torsten; Maccaferri, Francesco; Rivalta, Eleonora; Schierjott, Jana; Woith, Heiko
 Description*: Extended Pollino Seismic Experiment, GFZ Potsdam (FEFI, CCMP-Pompei, NERA projects)
 Abstract: The temporary Extended Pollino; The temporary Extended Pollino Seismic Experiment (FDSN network code Y4) monitored the earthquake swarm in the Pollino Range region, Italy, between September 2014 and April 2015.

• • • • •

Unavailable data

the swarm sequence. Waveform data will be fully open after April 2017. [1] Pollino Seismic Experiment, 2012–2014, doi:10.14470/9N904936

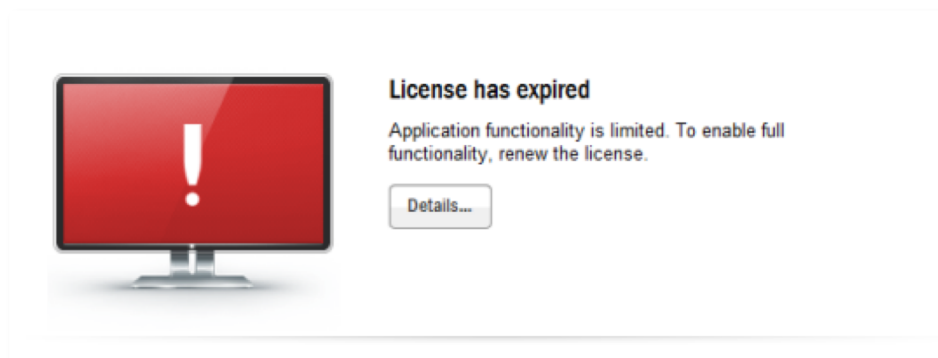
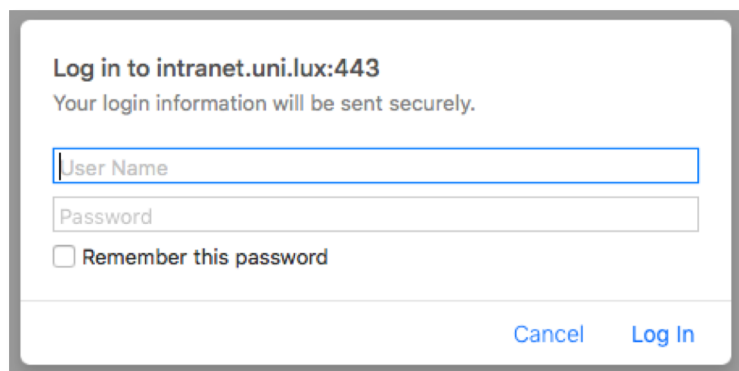
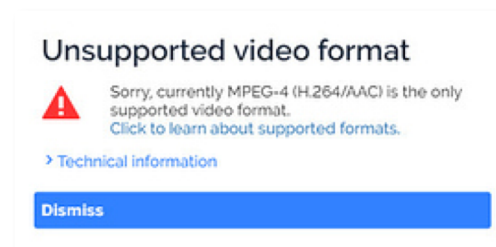
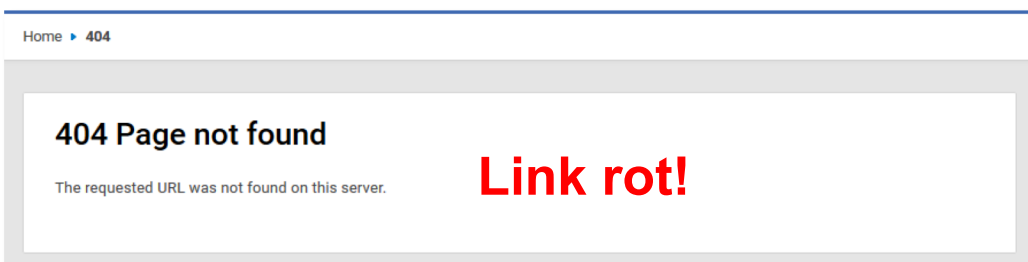
* Description is taken from seismic metadata, and may not match the preferred title for citations.

Sponsor(s): CCMP-POMPEI

FAIR principles for research data

Not Accessible

- (Meta)data are retrievable by a protocol
- Open, free universally implementable
- Authentication/Authorization
- Metadata available even when data is not



FAIR principles for research data



FAIR data overview - *Luiz Olavo Bonino da Silva ...*



SHARING DATA



I would like to exploit common genotype-phenotype relations between Alzheimer's Disease and Huntington's Disease...
I need to combine AD and HD data...



I can help with that!



FAIR principles for research data



FAIR data overview - Luiz Olavo Bonino da Silva ...



SHARING DATA



???

DOES NOT
COMPUTE



Here's my
data, have
fun!



米当局は、あなたの国籍に、在米日本領事代表にあなたが
逮捕又は拘束されたことを通報する必要があります。領事官
は通報を受けた後、あなたに電話を掛けたり、あるいはあなた
を訪問することができます。米当局は領事官の援助を求め
る必要はありませんが、あなたの家族との連絡、
書を取ってくれるかも知れない領事官に通報します。

Очи чёрные, очи живучие,
Очи страстные и прекрасные,
Как люблю я вас, как боюсь я вас,
Знать увидел вас я не в добрый час.

Очи чёрные, очи пламенные
И моят они в страны дальные,
Где царит любовь, где царит покой,
Где страдания нет, где вражды запрет.

Here's my
data, have
fun!



FAIR principles for research data

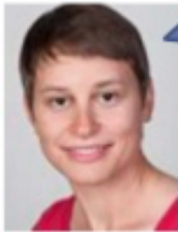


FAIR data overview - Luiz Olavo Bonino da Silva ...

DTL



SHARING LINKABLE DATA



I can go straight to answering my questions
with data from multiple data owners!
Patients will be so pleased with this speed-up!

Here's my
Linked Data,
have fun!



Here's my
Linked Data,
have fun!



FAIR principles for research data

Interoperable



What machine sees

What we expect to see in Data Integration/Analysis tool

- (Meta)data represented in formal, shared language
- Machine-actionable
- Controlled vocabularies
Tumour ≠ Tumor
- Community formats & standards
e.g. CDISC, HL7, ISA Tab...

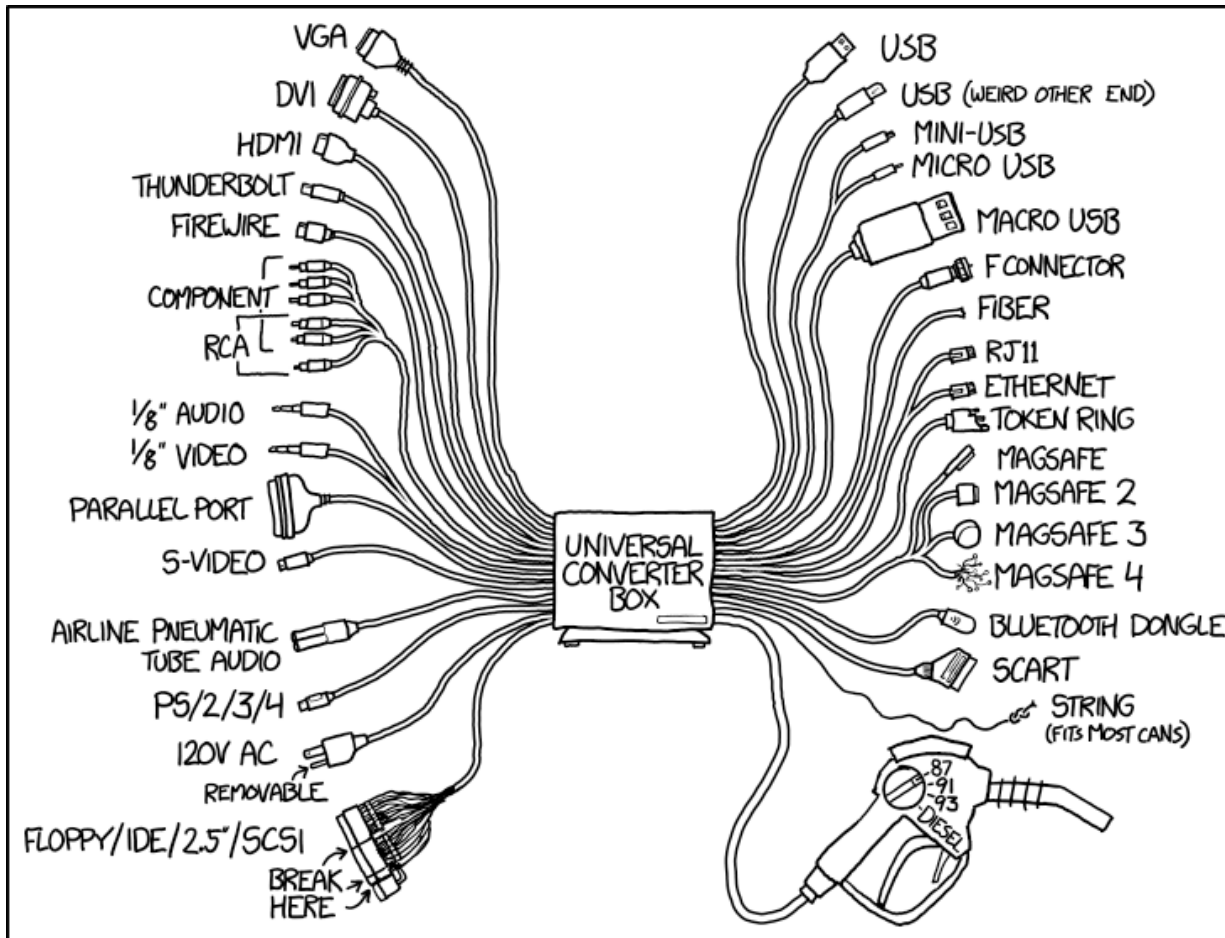


Source Name	Organism	Age	Unit	Sample Name	Protocol REF	Labeled Extract Name	...	Protocol REF	Data File
H1	H. Sapiens	35	Years	H1.sample1	Labeling	H1.sample1.labeled		Scanning	h1-s1.cel
H1	H. Sapiens	35	Years	H1.sample2				Scanning	h1-s2.cel
H2	H. Sapiens	33	Years	H2.sample1	Labeling	H2.sample1.labeled		Scanning	h2-s1.cel

FAIR principles for research data

Interoperable

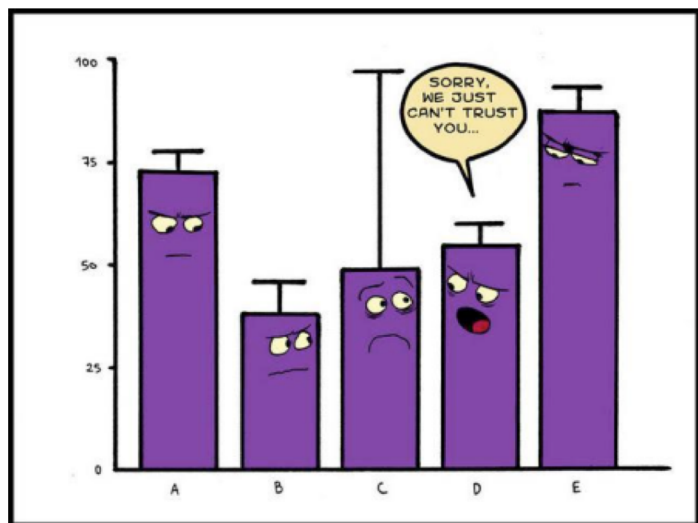
- use a standard that can be mapped to others



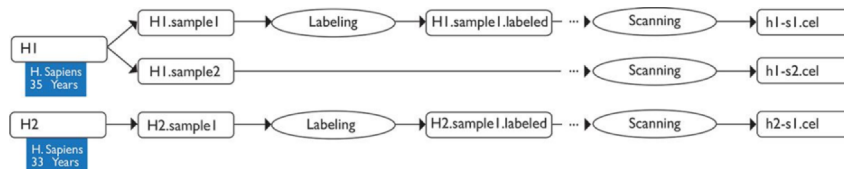
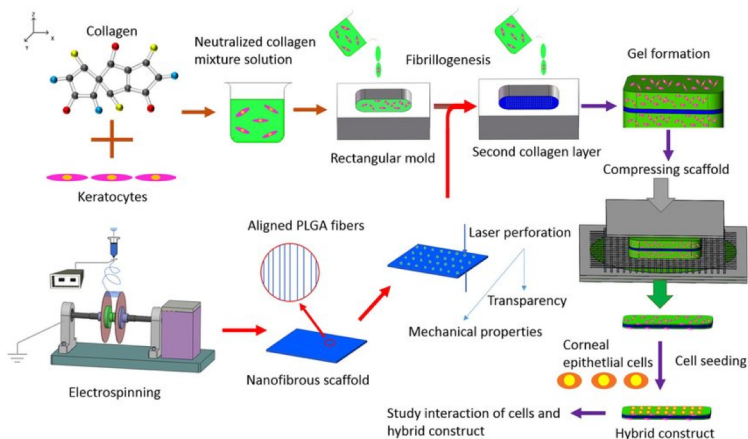
FAIR principles for research data

Reusable

- Multitude of metadata attributes
- Following community guideline
- Provenance



- cancer
- lung cancer
- lung cancer, 300 cases 200 controls
- lung cancer, 300 cases 200 controls, phenotyping, epigenetics, protocol A, platform B,



Ex-post-facto manual metadata collection = Pain

FAIR principles for research data

Reusable

- Descriptive Metadata, following community guideline
- Provenance of data

(GIGA)ⁿ_{DB}



European
genome-phenome
archive



THE MICHAEL J. FOX FOUNDATION
FOR PARKINSON'S RESEARCH



More metadata, more transparency

FAIR principles for research data

Reusable

- Accessible data use license

• Use Restrictions

Consent group	Is IRB required?	Data Access Committee	Number of participants
General Research Use ?	No	National Institute of Neurological Disorders and Stroke (ninds-dac@mail.nih.gov)	1617
Disease-Specific (Epilepsy and seizure disorders, GSO) ?	No	National Institute of Neurological Disorders and Stroke (ninds-dac@mail.nih.gov)	569

Publicly stored research

All publicly stored research outputs are stored under **Creative Commons Licenses**.

Why FAIR data is important

— Political pressure

- Increased push by public funders for maximum use of research results



H2020

3. Open access to research data (Extended Open Research Data Pilot)

What?

Beneficiaries of actions that participate in the Open Research Data Pilot must give **open, free-of-charge access** to the end-user to **digital research data** generated during the action (⚠ new in Horizon 2020).

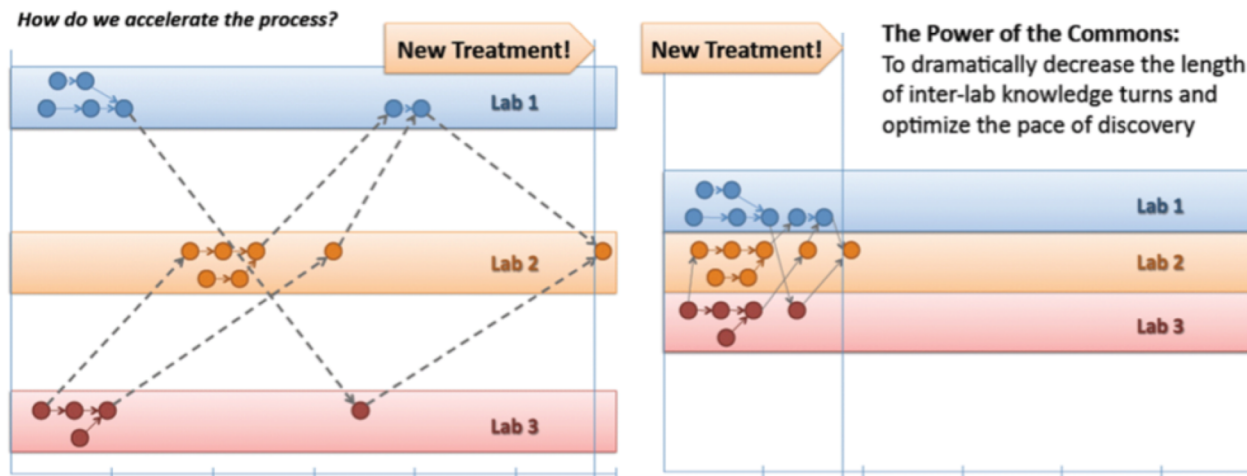
NIH

...

Why FAIR data is important

— Scientific value

- Pooling data, improved results, new questions
- Validation of models/methods over other data
- Accelerates “Knowledge Turns”

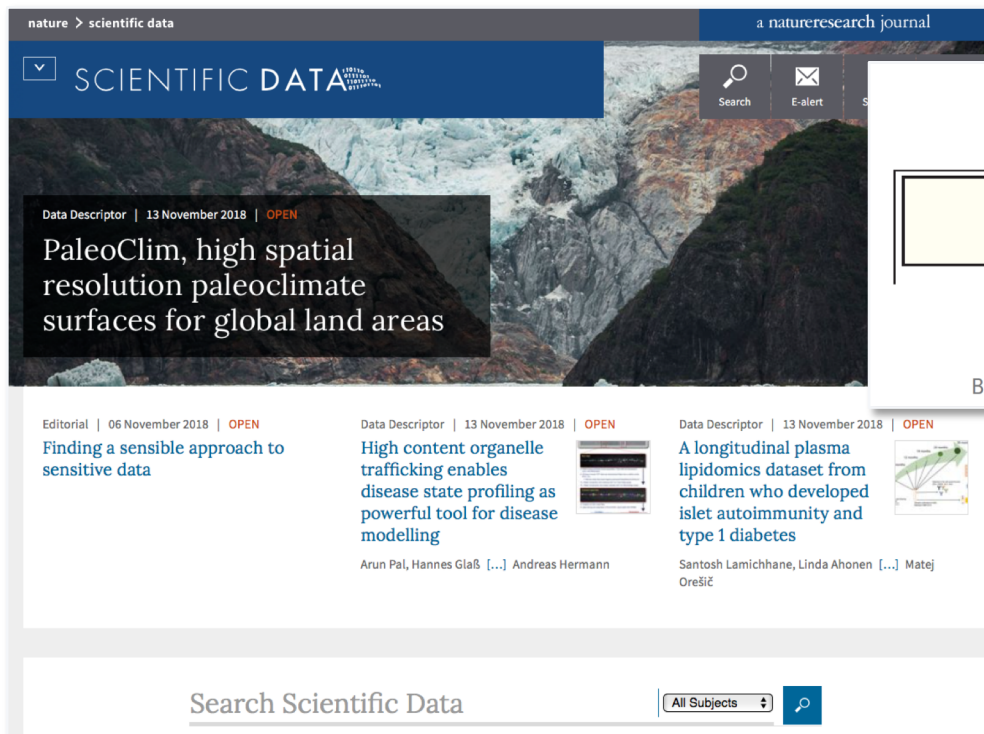


"Accelerating Scientists' Knowledge Turns" C A Goble, D De Roure, S Bechhofer

Why FAIR data is important

— New Incentives for scientists

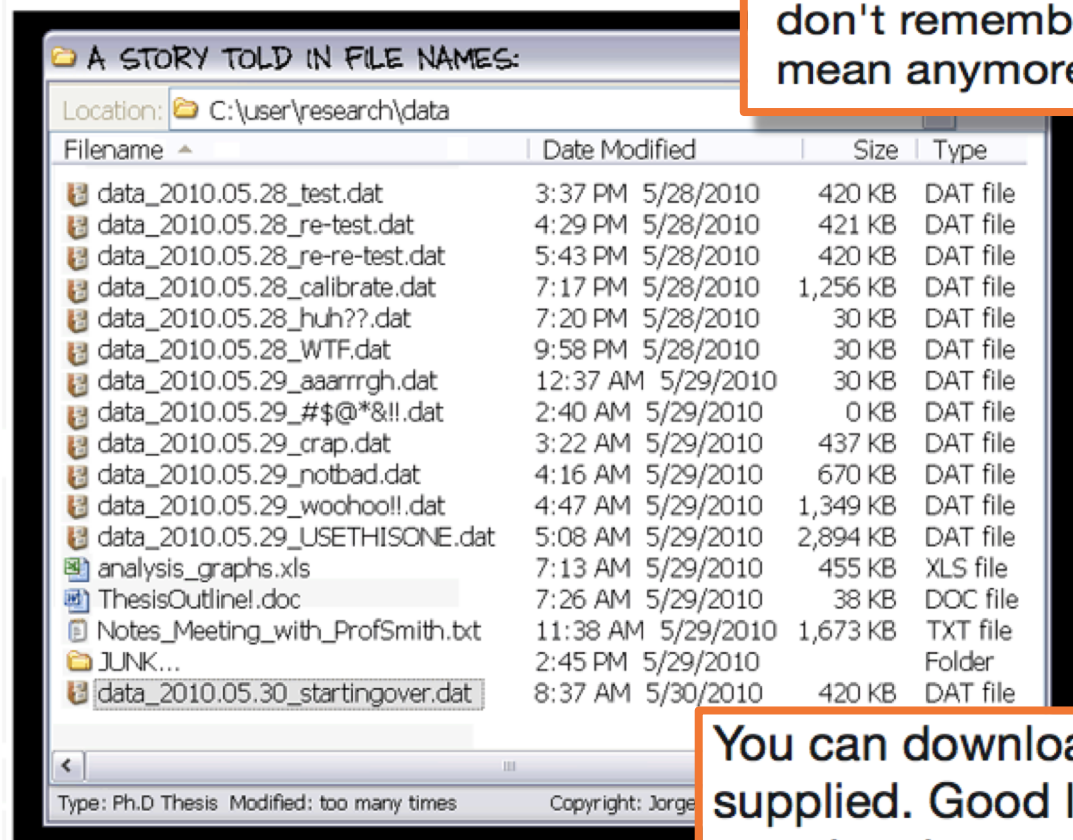
- Increased visibility, attracts new collaborations
- Data sharing increases research citation 9%
- FAIR data is being incorporated in the scholarly communication system (data paper, data citation)



Why FAIR data is important

— Improved research quality, reproducibility

- Data + code +documentation



I can't send you the original data because I don't remember what my excel file names mean anymore [#overlyhonestmethods](#)

You can download our code from the URL supplied. Good luck downloading the only postdoc that can get it to run, though [#OverlyHonestMethods](#)

Why FAIR data is important

— Improved research quality, reproducibility

- Data + code + documentation

ICLR 2018 Reproducibility Challenge

See the [2019 edition](#) of the challenge!

Background:

One of the challenges in machine learning research is to ensure that published results are reliable and reproducible. In support of this, the goal of this challenge is to investigate reproducibility of empirical results submitted to the [2018 International Conference on Learning Representations](#).

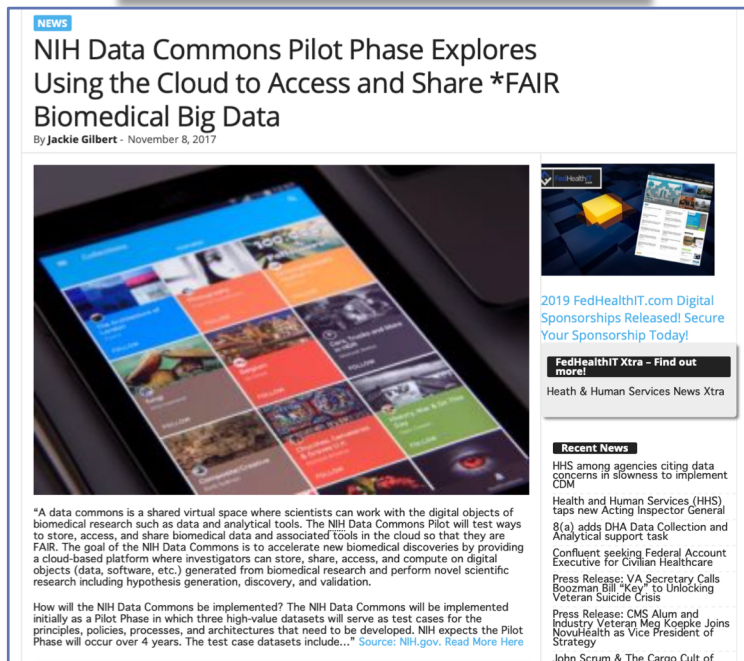
We are choosing ICLR for this challenge because the timing is right for course-based participants (see below), and because papers submitted to the conference are automatically made available publicly on [Open Review](#).

SIGMOD Most Reproducible Paper Award

SIGMOD recognizes the best papers in terms of reproducibility. The three most reproducible papers are picked every year. The criteria are as follows: (i) coverage (ideal: all results can be verified), (ii) ease of reproducibility (ideal: just works), (iii) flexibility (ideal: can change workloads, queries, data and get similar behavior with published results), and (iv) portability (ideal: linux, mac, windows).

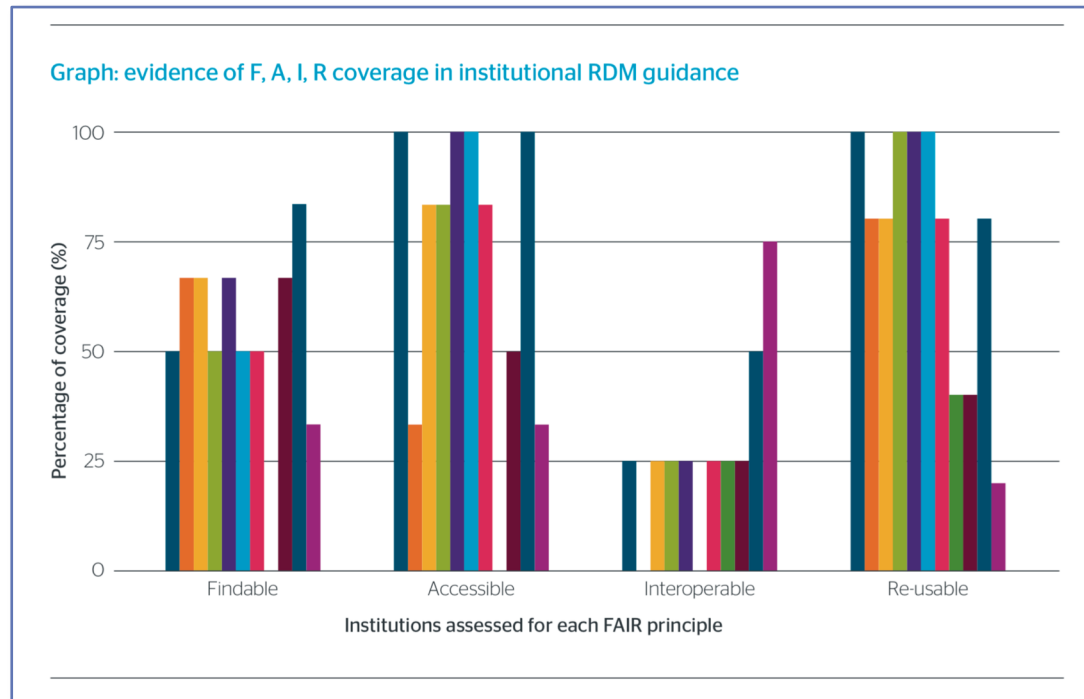


FAIR is spread across the lands...



FAIR is spread across the lands...but not necessarily all the peoples

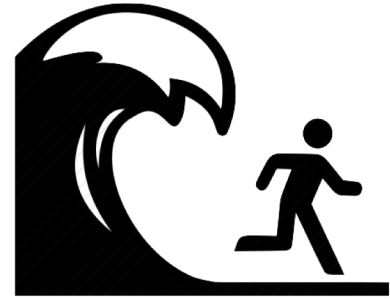
Stakeholder FAIR awareness



Government, Funder, Publisher, National & International Infrastructures...	
Institutional	
Researchers	

Achieving FAIR'ness

Data Management Planning



- Think of Data Management before data arrives!

Data Management Plan in a grant proposal is meant for this.

1. Identify

Your project's data

Legal and Ethical requirements

Your approach to:

Storage & Backup, Organization, Sharing,

Documentation, Preservation

Roles and Responsibilities

2. Implement!

Achieving FAIR'ness

- Assume your responsibility
- Primary responsible for RDM is the researcher
- Data Stewards and IT are in supporting roles
- Budget for DM in your proposals

Achieving FAIR'ness

- Use Virtual Research Environments, as well as FAIR tools and data, and standards
- Last mile for infrastructure providers
- First mile for researchers



Taverna



FAIRsharing.org
standards, databases, policies

Take home messages



- Data is a valuable research output
- You can optimize the value by performing RDM
- Managed data is FAIR
 - more of a (research) culture change and funding problem than technical
 - long term preservation and re-use of data is ultimate goal
 - planning is key initial step for good data management
 - primary responsible for RDM is the researcher!



Thank you