FAIR Principles for Research Data

Pinar Alper

Training on "Best practices in research data management and stewardship"

14 June 2021

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.
  - Publishing and sharing after the project completion e.g. depositing the data in a community repository.

Whyte, A., Tedds, J. (2011). 'Making the Case for Research Data Management'. DCC Briefing Papers.

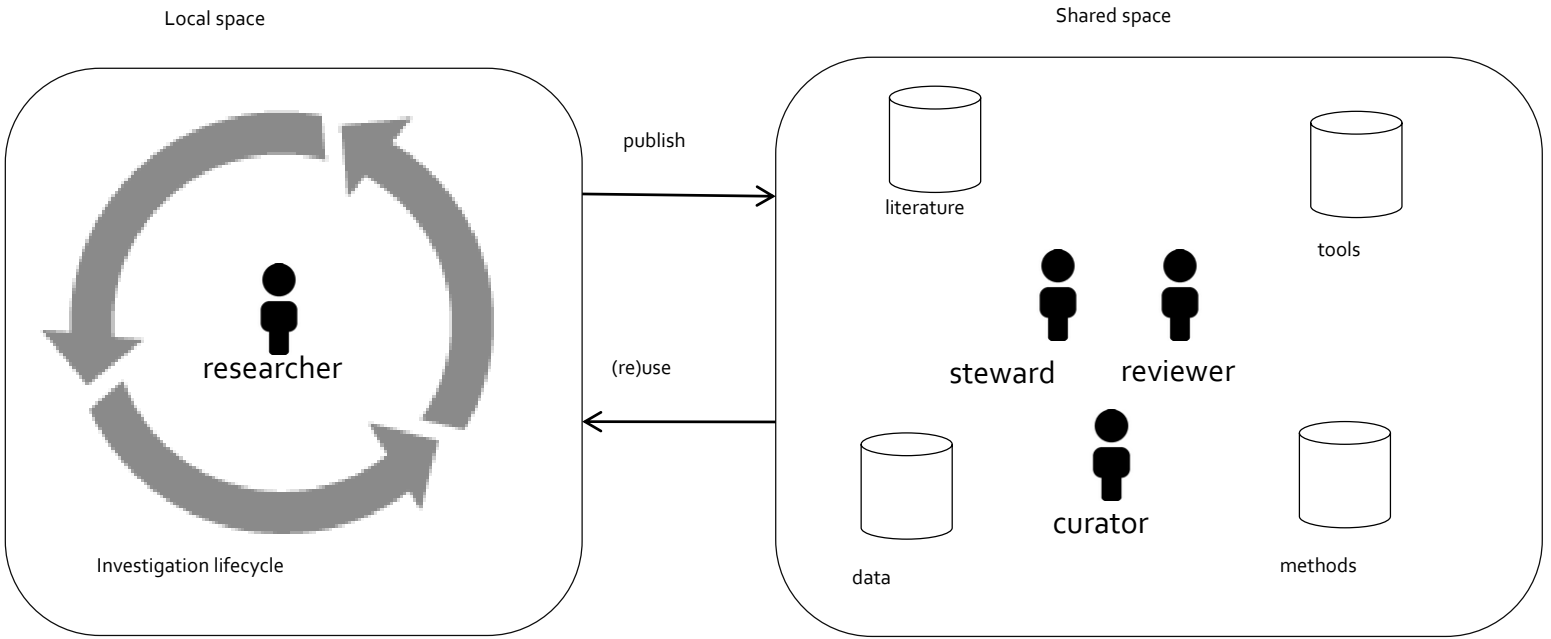
Data Stewardship = RDM ++

" Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data."

Wilkinson M, Dumontier M et al. Nature Scientific Data 2016. "The FAIR Guiding Principles for scientific data management and stewardship"

Data Stewardship

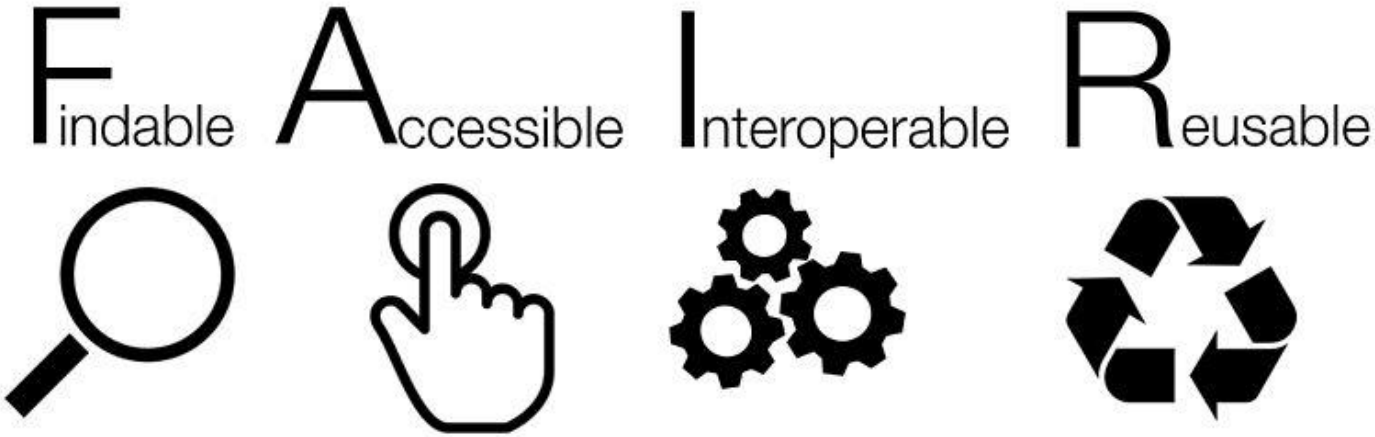
— aims to enable long term care and re-use of data



15 %

End goal

— FAIR principles for research data

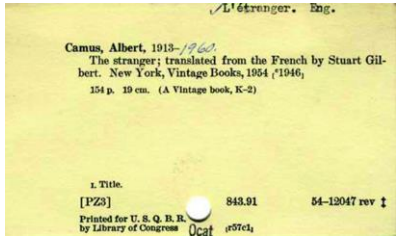


Wilkinson M, Dumontier M et al. Nature Scientific Data 2016. "The FAIR Guiding Principles for scientific data management and stewardship"

FAIR principles for research data

Findable

- (Meta)data
- Unique and eternal identifiers for (meta)data
- Indexed in a searchable resource
- Metadata contains data identifier



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

LCSB

elixir LUXEMBOURG

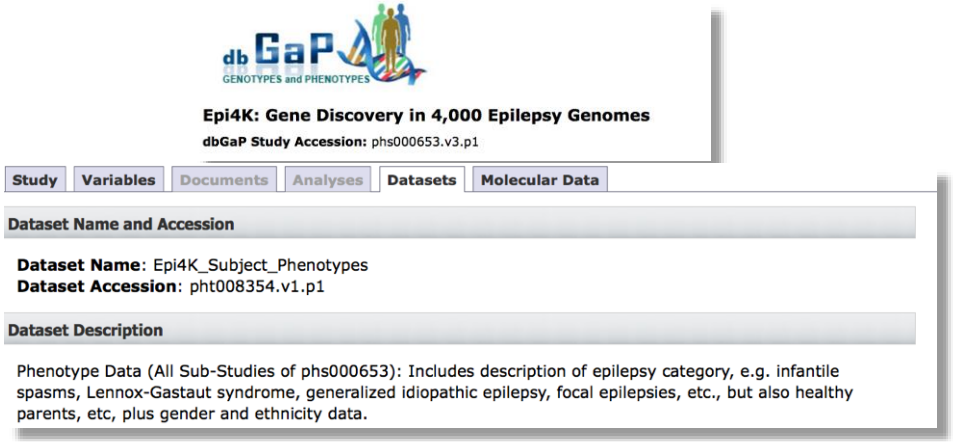
FAIR principles for research data

Findable

- Metadata
- Unique and external 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).



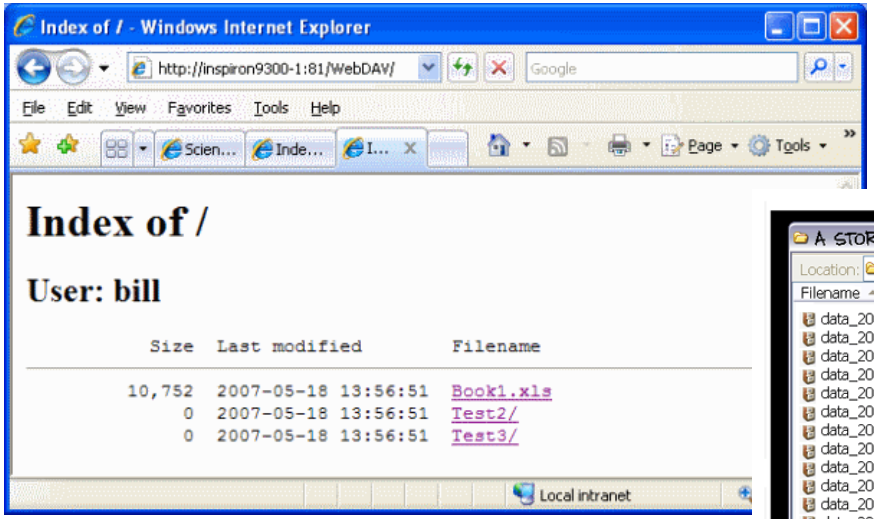
LCSB

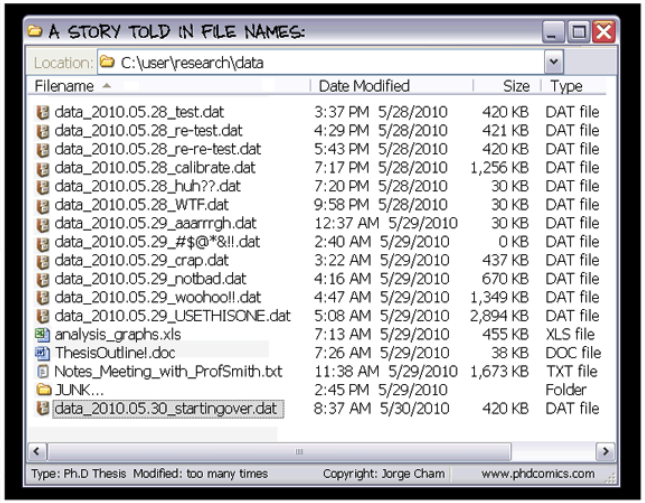
elixir LUXEMBOURG

FAIR principles for research data

Not (so)Findable

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



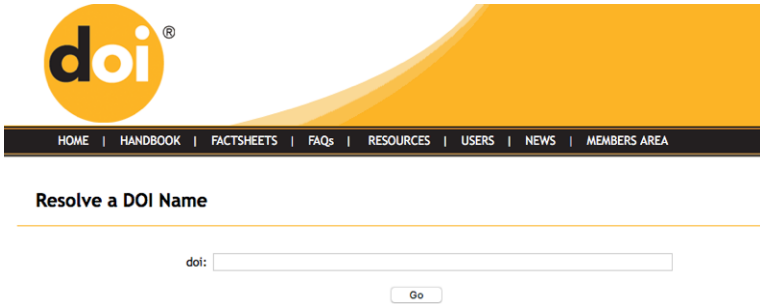


FAIR principles for research data

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).



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

FAIR principles for research data

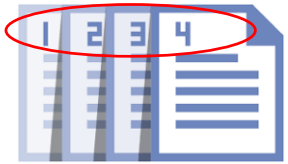
Accessible

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

Moved data

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

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*

Type: *Temporary*
Archived at: *GFZ*
Time Range: 2014–2015

Identifier: *doi:10.14470/L9180569*
(Citation information)
DataCite metadata: *HTML | JSON | XML | INSPIRE*

the swarm sequence. Waveform data will be fully open after April 2017. [1] Pollino Seismic Experiment, 2012-2014, doi:10.14470/9N904956
* Description is taken from seismic metadata, and may not match the preferred title for citations.
Sponsor(s): *CCMP-POMPEI*

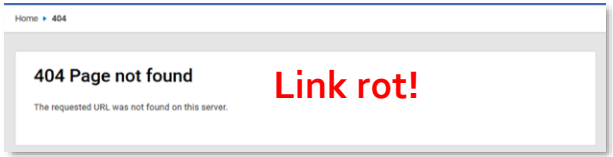
Embargoed data

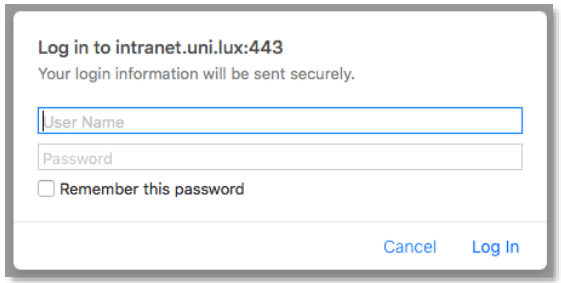
FAIR principles for research data

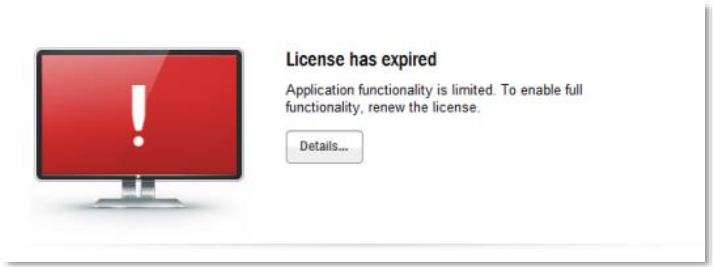
Not so accessible

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

Data are available on request due to privacy or other restrictions







LCSB

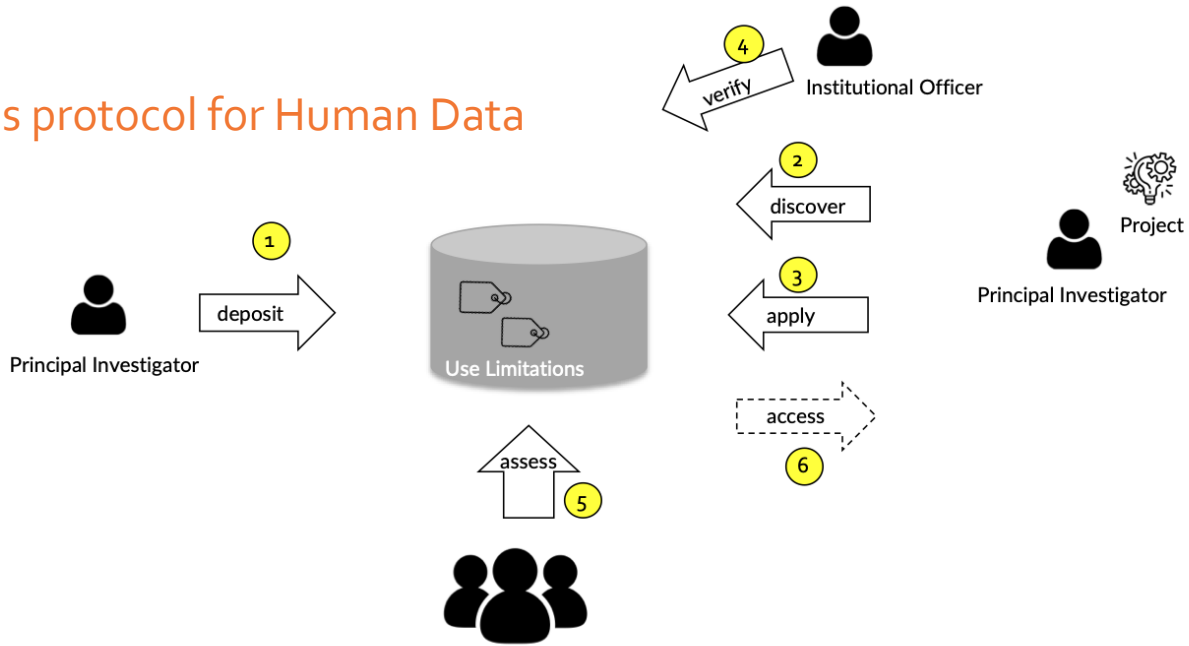
elixir LUXEMBOURG

FAIR principles for research data

Accessible ≠ Unrestricted for all

Accessible ≠ Free(OPEN ACCESS)

Access protocol for Human Data



39 %

FAIR principles for research data

Let's apply our technique for the deep learning of **gene mutation ← →drug** relations from literature to **epilepsy**. Here's data from a German cohort, which we can curate.

Good bye
Christmas break.

FAIR principles for research data

Let's apply our technique of deep learning of **gene mutation ← →drug** relations from literature to **epilepsy**. Here's a German cohort's data we can curate.

ID	Sample	Gender	Race	Vater	Mutter	Birthdate	Disease	Mutation1	Mutation2	Mutation3	Family ID		Comment
101	L1001	male	German	104	103	26/2/1992	Idiopathic Generalized Epilepsy	A37V	F73T	Y301V	K1	AID1029	Available
102	L102	m	German	104	103	29/2/1992	Unaffected	D40V	F73T	Y301V	K1	AID1381	Available
102	L1003	m	GErman	"	"	"		D124V		D124T	K1		
102	L1003	m	German	"	"	"		E345T	F78I	I98N	K1		
103	L1004	m	Caucasian			12/11/70	0	E12T	F73T	I98T	K1	AID2738	available: no
104	L1005	male	Caucasian			11-12-70	No	D45H			K1	AID2731	Available
104	L1103										K1	AID2735	
106	T1007	f	Asian	107	108	1975-12-10	IGE	C98K	F73T		F2	AID1291	Sample lost
108	T1008	f	Chinese			1972-10-01	Genetic Generalized, Father had GGE	K76V			F2	AID2389	
109	L2987	male	Chinese			5.4.1970	Unaffected	V98G	F73T	Y301V	L4	AID3849	Vater von Kevin
110	L4002	male	Russian	Horst	Inga	1/1/1990	Focal epilepsy		F73T		L4		yes
111	L1872	Female	East-German			31.4.1970	Temporal lobe epilepsy	C27G	F73T	Y301V	L4	AID8782	yes

Manual interoperation.

C

`mutation` table

id	gene name	acc no ncbi	nt seq	amplicon start	amplicon end	m1 name	m2 name	allele
[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]
274	HvSNAC1	JF796130	ATGGGG...	1	1132	7009	1	.n
275	HvSNAC1	JF796130	ATGGGG...	1	1132	1825	1	.o
[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]

homo hetero	nt before	nt position	nt after	localisation	mutation type	aa seq	aa before	aa position	aa after
[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]
homo	G	883	A	coding	silent	MGMPAA...	G	295	G
homo	G	965	A	coding	missense	MGMPAA...	G	276	S
[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]

LCSB

elixir LUXEMBOURG

FAIR principles for research data

Let's apply our technique of deep learning of **gene mutation ← →drug** relations from literature to **epilepsy**. Here's a German cohort's data we can curate.

ID	Sample	Gender	Race	Vater	Mutter	Birthdate	Disease	Mutation1	Mutation2	Mutation3	Family ID		Comment
101	L1001	male	German	104	103	26/2/1992	Idiopath…eralized Epilepsy	A37V		Y301V	K1	AID1029	Available
102	L102	m	German	104	103	29/2/1992	Unaffe…	D4…		Y301V	K1	AID1381	Available
102	L1003	m	GErman	"	"	"		[illegible]		D124T	K1		
102	L1003	m	German	"	"	"		E345T		I98N	K1		
103	L1004	m	Caucasian			12/11/70	0	E12T		I98T	K1	AID2738	available: no
104	L1005	male	Caucasian			11-12-70	No	D45H			K1	AID2731	Available
104	L1103										K1	AID2735	
106	T1007	f	Asian	107	108	1975-12-10	IGE	C98K	F73T		F2	AID1291	Sample lost
108	T1008	f	Chinese			1972-10-01	Genetic Generalized, Father had GGE	K76V			F2	AID2389	
109	L2987	male	Chinese			5.4.1970	Unaffected	V98G	F73T	Y301V	L4	AID3849	Vater von Kevin
110	L4002	male	Russian	Horst	Inga	1/1/1990	Focal epilepsy		F73T		L4		yes
111	L1872	Female	East-German			31.4.1970	Temporal lobe epilepsy	C27G	F73T	Y301V	L4	AID8782	yes

Automated interoperation.

C `mutation` table

id	gen…	acc no ncbi	nt seq	amplicon start	amplicon end	m1 name	…	allele
[…]	[…]	[…]	[…]	[…]	[…]	[…]	[…]	[…]
27…	HvSNA…1	JF796130	ATGGGG…	1	1132	[…]	[…]	.n
275	Hv…	JF796130	ATGGGG…	1	1132	1825	[…]	.o
[…]		[…]	[…]	[…]	[…]	[…]	[…]	[…]

homo hetero	nt before	nt position	nt after	localisation	mutation type	aa seq	aa before	aa position	aa after
[…]	[…]	[…]	[…]	[…]	[…]	[…]	[…]	[…]	[…]
homo	G	883	A	coding	silent	MGMPAA…	G	295	G
homo	G	965	A	coding	missense	MGMPAA…	G	276	S
[…]	[…]	[…]	[…]	[…]	[…]	[…]	[…]	[…]	[…]

FAIR principles for research data

Interoperable

What machine sees

- (Meta)data represented in formal, shared language
- Machine-actionable
- Controlled vocabularies

  Tumour ≠ Tumor
- Community formats & standards
  e.g. CDISC, HL7, ISA Tab…

What we expect to see in Data Integration/Analysis tool

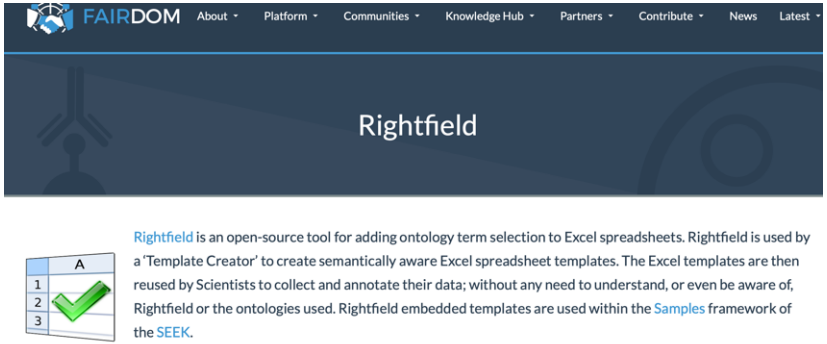
Source Name	Organism	Age	Unit	Sample Name	Protocol REF	Labeled Extract Name	…	Protocol REF	Data File
H1	H. Sapiens	35	Years	H1.sample1	Labeling	H1.sample1.labeled		Scanning	h1-s1.cel
H1	H. Sapiens	35	Years	H1.sample2				Scanning	h1-s2.cel
H2	H. Sapiens	33	Years	H2.sample1	Labeling	H2.sample1.labeled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

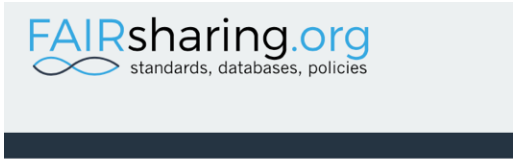
https://bioportal.bioontology.org

FAIR principles for research data

Interoperable

- resources towards achieving interoperable data





1412 Standards

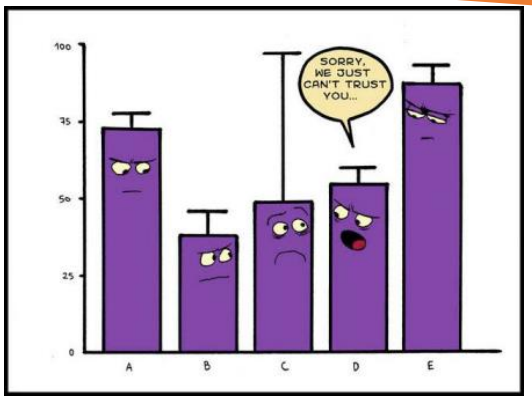
Terminology Artifact	784
Model/Format	414
Reporting Guideline	169
Identifier Schema	15
FAIR metrics	30

View all

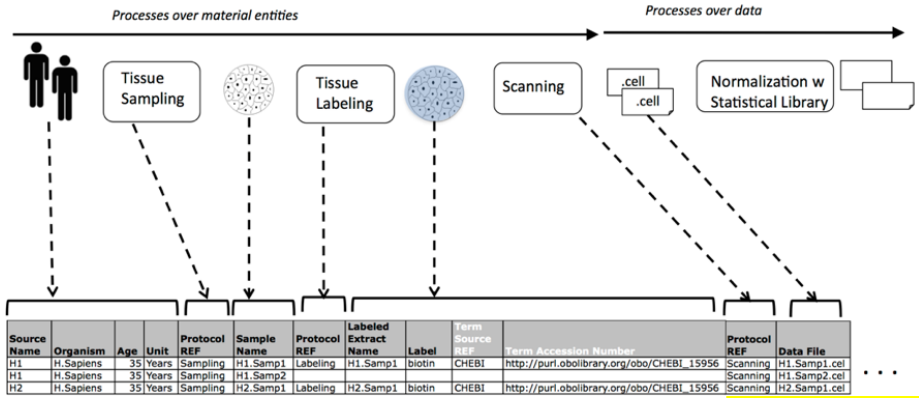
FAIR principles for research data

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, ...



Source Name	Organism	Age	Unit	Protocol REF	Sample Name	Protocol REF	Labeled Extract Name	Label	Term Source REF	Term Accession Number	Protocol REF	Data File
H1	H.Sapiens	35	Years	Sampling	H1.Samp1	Labeling	H1.Samp1	biotin	CHEBI	http://purl.obolibrary.org/obo/CHEBI_15956	Scanning	H1.Samp1.cel
H1	H.Sapiens	35	Years	Sampling	H1.Samp2						Scanning	H1.Samp2.cel
H2	H.Sapiens	35	Years	Sampling	H2.Samp1	Labeling	H2.Samp1	biotin	CHEBI	http://purl.obolibrary.org/obo/CHEBI_15956	Scanning	H2.Samp1.cel

Ex-post-facto provenance collection = pain!

https://fairsharing.org/standards/

FAIR principles for research data

Reusable

- Descriptive metadata, following community guideline
- Provenance of data
- Clear and accessible data use license

"generalist repositories"

"domain repositories"

More metadata, more transparency, more likelihood of re-use

Why FAIR data?

— 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).

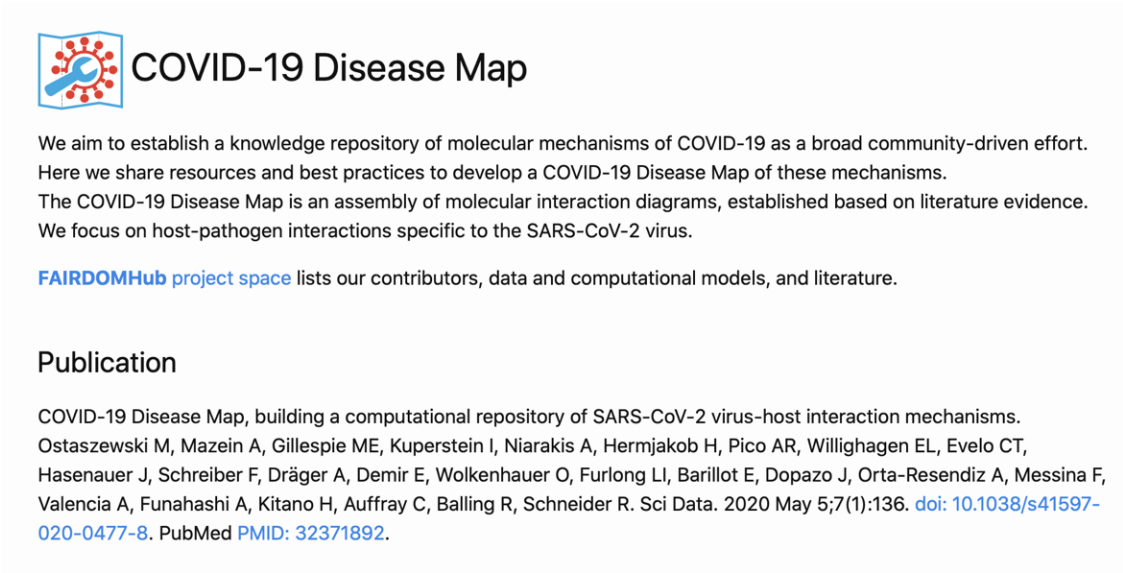
NIH

...

Why FAIR data?

— Scientific value

- Pooling results, for improving results or new questions
- Validation of models/methods over other data
- Accelerated inter-lab exchange of knowledge



COVID-19 Disease Map

We aim to establish a knowledge repository of molecular mechanisms of COVID-19 as a broad community-driven effort.
Here we share resources and best practices to develop a COVID-19 Disease Map of these mechanisms.
The COVID-19 Disease Map is an assembly of molecular interaction diagrams, established based on literature evidence.
We focus on host-pathogen interactions specific to the SARS-CoV-2 virus.

FAIRDOMHub project space lists our contributors, data and computational models, and literature.

Publication

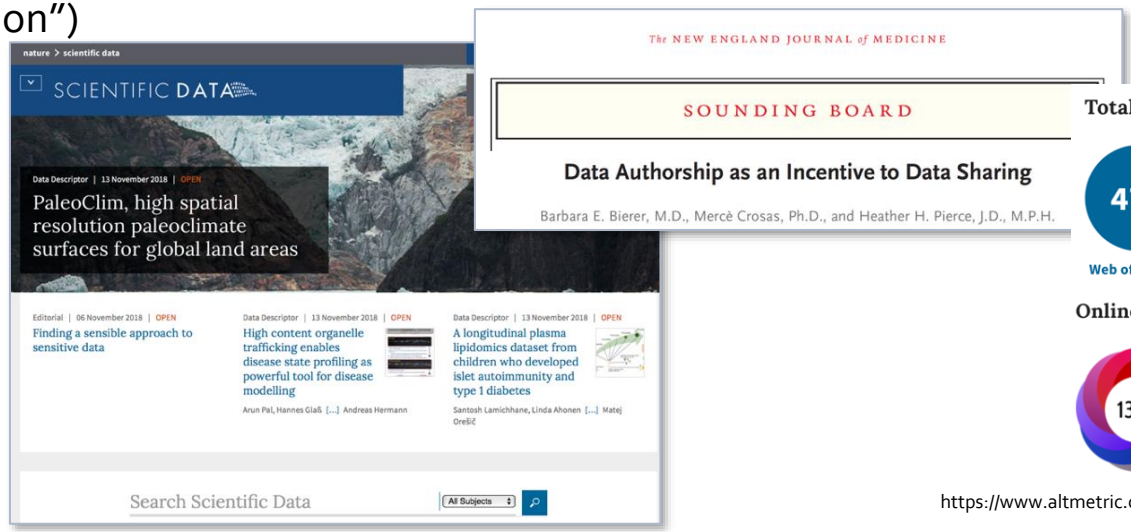
COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. Ostaszewski M, Mazein A, Gillespie ME, Kuperstein I, Niarakis A, Hermjakob H, Pico AR, Willighagen EL, Evelo CT, Hasenauer J, Schreiber F, Dräger A, Demir E, Wolkenhauer O, Furlong LI, Barillot E, Dopazo J, Orta-Resendiz A, Messina F, Valencia A, Funahashi A, Kitano H, Auffray C, Balling R, Schneider R. Sci Data. 2020 May 5;7(1):136. doi: 10.1038/s41597-020-0477-8. PubMed PMID: 32371892.

https://fairdomhub.org/projects/190#models

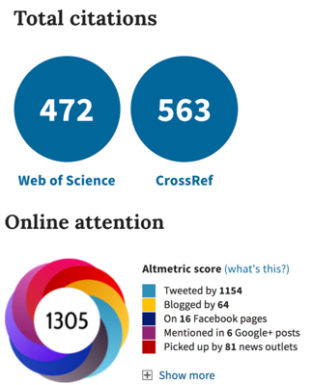
Why FAIR data?

— 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")



https://www.nature.com/sdata/



https://www.altmetric.com/products/altmetric-badges/

Piwowar,H A. and Vision,T J. et al. Data reuse and the open data citation advantage . PeerJ 2013 Volume 1

Why FAIR data?

— 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

FAIR is spread across the lands…

EUROPEAN COMMISSION
Directorate-General for Research & Innovation

H2020 Programme

Guidelines on

FAIR Data Management in Horizon 2020

NEWS

NIH Data Commons Pilot Phase Explores Using the Cloud to Access and Share *FAIR Biomedical Big Data

By **Jackie Gilbert** - November 8, 2017

2019 FedHealthIT.com Digital Sponsorships Released! Secure Your Sponsorship Today!

FedHealthIT Xtra – Find out more!

Heath & Human Services News Xtra

Recent News

HHS among agencies citing data concerns in slowness to implement CDM

Health and Human Services (HHS) taps new Acting Inspector General

8(a) adds DHA Data Collection and Analytical support task

Confluent seeking Federal Account Executive for Civilian Healthcare

Press Release: VA Secretary Calls Boozman Bill "Key" to Unlocking Veteran Suicide Cross

Press Release: CMS Alum and Industry Veteran Meg Koepke Joins NovuHealth as Vice President of Strategy

Johns Scrum & The Cargo Cult of

"A data commons is a shared virtual space where scientists can work with the digital objects of biomedical research such as data and analytical tools. The NIH Data Commons Pilot will test ways to store, access, and share biomedical data and associated tools in the cloud so that they are FAIR. The goal of the NIH Data Commons is to accelerate new biomedical discoveries by providing a cloud-based platform where investigators can store, share, access, and compute on digital objects (data, software, etc.) generated from biomedical research and perform novel scientific research including hypothesis generation, discovery, and validation.

How will the NIH Data Commons be implemented? The NIH Data Commons will be implemented initially as a Pilot Phase in which three high-value datasets will serve as test cases for the principles, policies, processes, and architectures that need to be developed. NIH expects the Pilot Phase will occur over 4 years. The test case datasets include..." Source: NIH.gov. Read More Here

EUROPEAN COMMISSION
DIRECTORATE-GENERAL FOR RESEARCH & INNOVATION

The Director-General

Brussels, 10 July 2017

EOSC Declaration

RECOGNISING the challenges of data driven research in pursuing excellent science;

GRANTING that the vision of European Open Science is that of a research data commons, widely inclusive of all disciplines and Member States, sustainable in the long-term,

CONFIRMING that the implementation of the EOSC is a process, not a project, by its nature iterative and based on constant learning and mutual alignment;

UPHOLDING that the EOSC Summit marked the beginning and not the end of this process, one based on continuous engagement with scientific stakeholders, the European Commission,

PROPOSES that all EOSC stakeholders consider sharing the following intents and will actively support their implementation in the respective capacities:

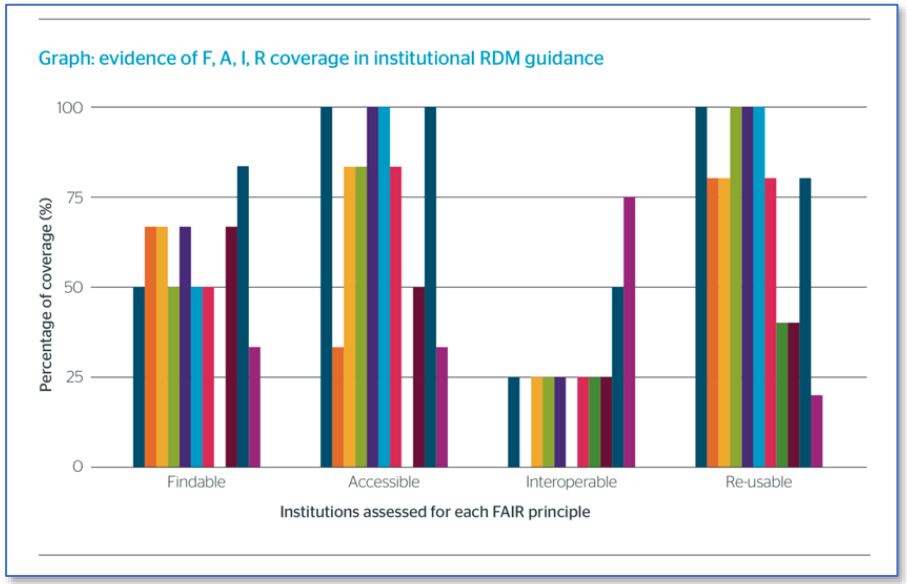
Data culture and FAIR data

- [**Data culture**] European science must be grounded in a common culture of data stewardship, so that research data is recognised as a significant output of research and is appropriately curated throughout and after the period conducting the research. Only a considerable cultural change will enable long-term reuse for science and for innovation of data created by research activities: no disciplines, institutions or countries must be left behind.

Adapted from "FAIRy stories for Christmas" Carole A Goble Keynote SWAT4HCLS Keynote 2017

…but not necessarily across all the peoples

Stakeholder FAIR awareness

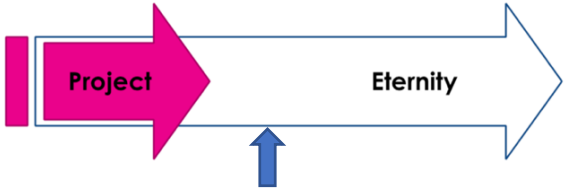


……has evolved into "FAIR fatigue" before "FAIR adoption"

Adapted from "FAIRy stories for Christmas" Carole A Goble Keynote SWAT4HCLS Keynote 2017

Achieving FAIR'ness

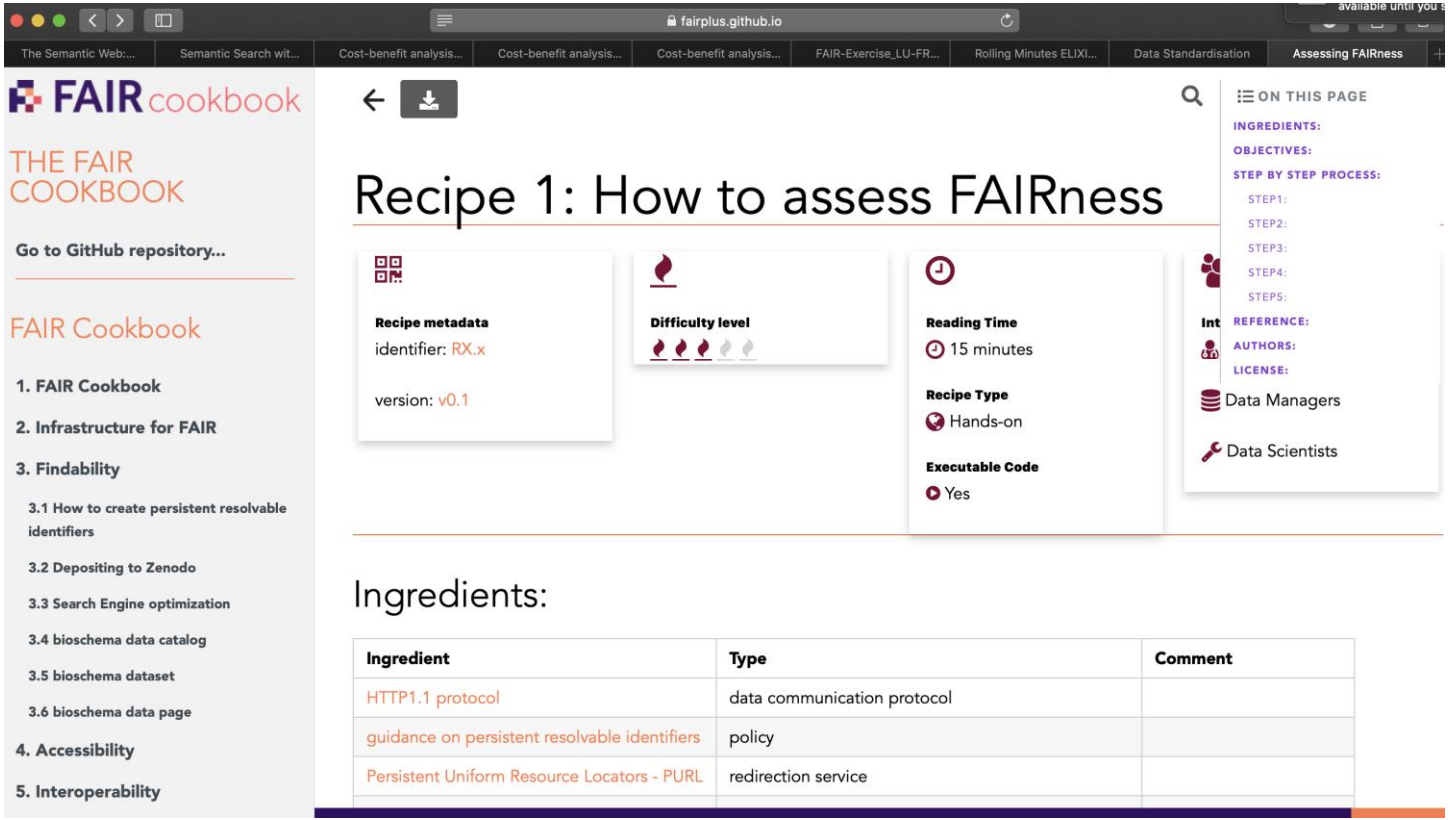
— Posthoc "FAIRification"



- Applied for datasets of higher value/re-usability potential
- Costly process
- Requires FAIRification experts
- **Assumption**: a percentage of research data will inevitably be non-FAIR at project end.

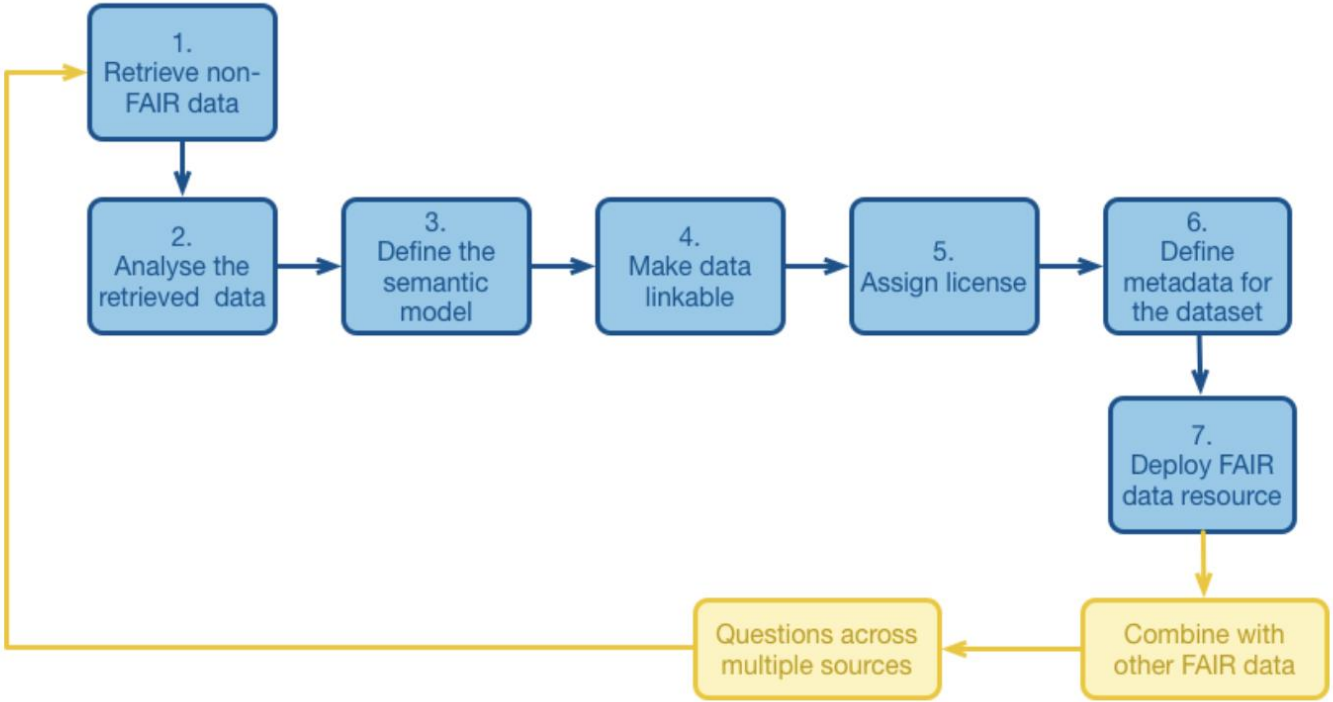
FAIRification is an expertise

— FAIRPlus



FAIRification is an expertise

— GO FAIR



https://www.go-fair.org/go-fair-initiative/

Achieving FAIR'ness

— Data that born FAIR

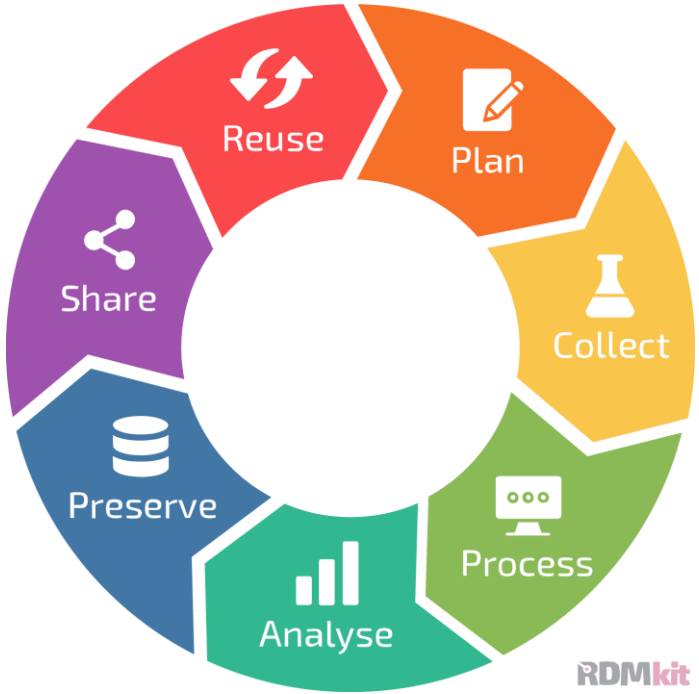
- Use research infrastructures, FAIR tools, standards and practices from Day 1
- **Assumption**: FAIR can only be achieved at scale by good data management practice.

LCSB

elixir LUXEMBOURG

Good RDM practice during the entire project lifecycle

ELIXIR RDMkit



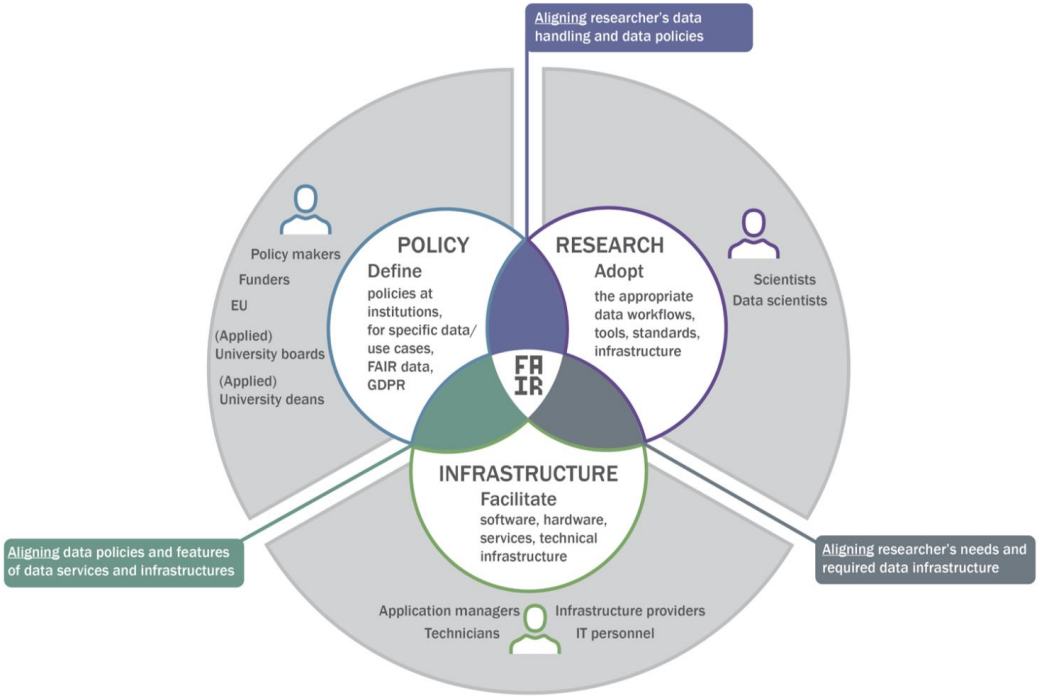
https://rdmkit.elixir-europe.org

LCSB

elixir LUXEMBOURG

Professionalizing RDM Support

— "Professionalising data stewardship" in the Netherlands.



"Professionalising data stewardship in the Netherlands. Competences, training and education. Dutch roadmap towards national implementation of FAIR data stewardship" M Jetten et al 10.5281/zenodo.4320504.

Conclusions

- Data is a valuable research output
- You can optimize its value by performing RDM
- Managed data is FAIR!

Thank you

elixir LUXEMBOURG

LCSB