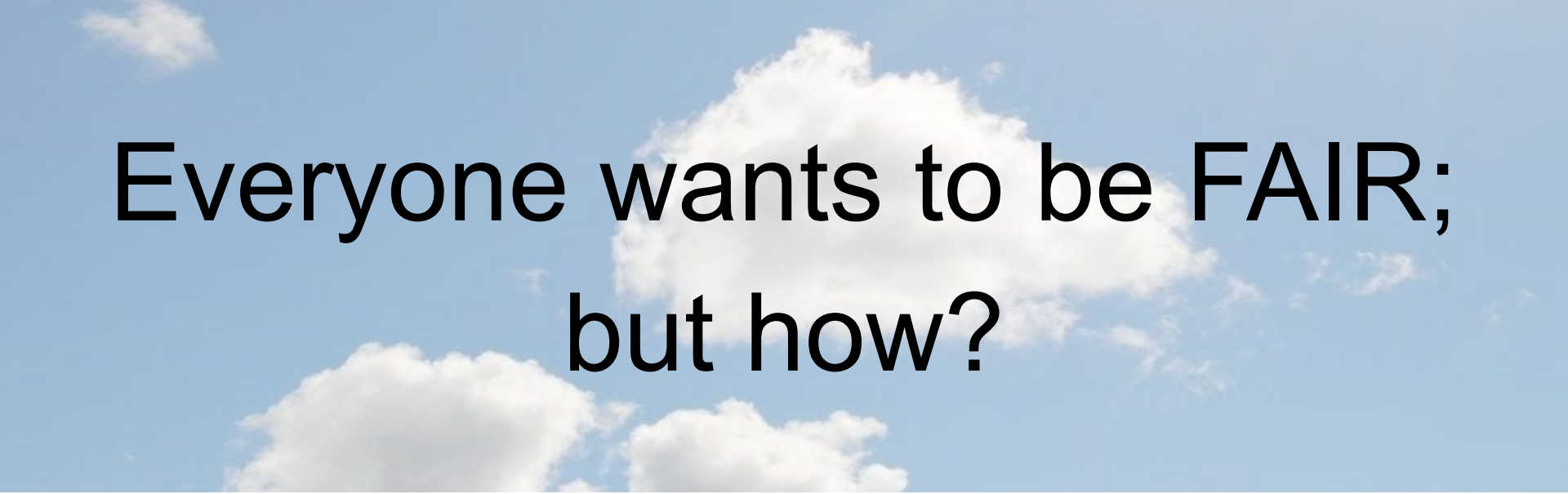
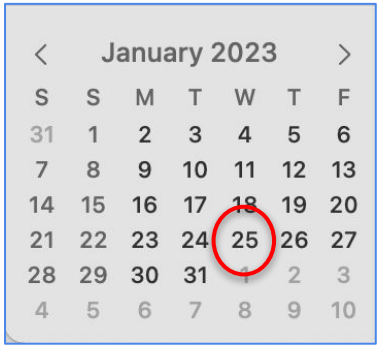


Everyone wants to be FAIR; but how?

Maryann E. Martone, Ph. D.
Professor Emerita, UCSD
Founder, SciCrunch.com

*A tech startup out of UCSD that provides services around Research Resource Identifiers



January 25, 2023

- US National Institutes of Health new data sharing policy goes into effect
- Most data should be shared
- Data should be FAIR
- “As open as possible; as closed as necessary”
- ***All data must be managed***

Final NIH Policy for Data Management and Sharing

Notice Number:
NOT-OD-21-013

Key Dates

Release Date: October 29, 2020
Effective Date: January 25, 2023

Related Announcements

NOT-HG-21-023 - Notice Announcing NHGRI Guidance for Third-Party Involvement in Extramural Research

NOT-HG-21-022 - Notice Announcing the National Human Genome Research Institute's Expectation for Sharing Quality Metadata and Phenotypic Data

NOT-OD-21-014 – Supplemental Information to the NIH Policy for Data Management and Sharing: Elements of an NIH Data Management and Sharing Plan

NOT-OD-21-015 – Supplemental Information to the NIH Policy for Data Management and Sharing: Allowable Costs for Data Management and Sharing

NOT-OD-21-016 – Supplemental Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research

NOT-OD-20-013 - Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Issued by

Office of The Director, National Institutes of Health (OD)

Purpose

Summary

The National Institutes of Health (NIH) is issuing this final NIH Policy for Data Management and Sharing (DMS Policy) to promote the management and sharing of scientific data generated from NIH-funded or conducted research. This Policy establishes the requirements of submission of Data Management and Sharing Plans (hereinafter Plans) and compliance with NIH Institute, Center, or Office (ICO)-approved Plans. It also emphasizes the importance of good data management practices and establishes the expectation for maximizing the appropriate sharing of scientific data generated from NIH-funded or conducted research, with justified limitations or exceptions. This Policy applies to research funded or conducted by NIH that

The National Academies of SCIENCES ENGINEERING MEDICINE

HEALTH AND MEDICINE DIVISION

BOARD ON HEALTH SCIENCES POLICY

BOARD ON HEALTH CARE SERVICES

Changing the Culture of Data Management and Sharing — A Workshop

The tenor of conversation around data sharing has changed: there was a desire to move beyond simple compliance. Data sharing should be impactful and a benefit to all stakeholders. Stakeholders should actively participate in defining DMS.

April 28-29, 2021

#DataSharing

https://www.nationalacademies.org/event/04-29-2021/changing-the-culture-of-data-management-and-sharing-a-workshop

Data management and sharing is good for...

- **Science**
  - Transparency
  - Reproducibility
  - Reduced waste
  - Driving discovery
- **Future me**
  - One most likely to benefit from good data management and sharing through stable archives
  - No one ever regretted annotating and documenting too much
- **Me**
  - Answer to the underpowered study
  - Data sharing and good data management are closely aligned
  - Compliance with mandates: journal and funder
- **My colleagues**
  - Easy to engage with colleagues over well annotated data and associated code
  - PIs: What happens when the graduate student or post doc leaves?

Research Data Management Guide for Researchers

	Ad Hoc	One-Time	Active and Informative	Optimized for Re-Use
Planning your project	When it comes to my data, I have a "way of doing things" but no standard or documented plans.	I create some formal plans about how I will manage my data at the start of a project, but I generally don't refer back to them.	I develop detailed plans about how I will manage my data that I actively revisit and revise over the course of a project.	I have created plans for managing my data that are designed to streamline its future use by myself or others.
Organizing your data	I don't follow a consistent approach for keeping my data organized, so it often takes time to find things.	I have an approach for organizing my data, but I only put it into action after my project is complete.	I have an approach for organizing my data that I implement prospectively, but it not necessarily standardized.	I organize my data so that others can navigate, understand, and use it without me being present.
Saving and backing up your data	I decide what data is important while I am working on it and typically save it in a single location.	I know what data needs to be saved and I back it up after I'm done working on it to reduce the risk of loss.	I have a system for regularly saving important data while I am working on it. I have multiple backups.	I save my data in a manner and location designed maximize opportunities for re-use by myself and others.

Borghi J, Abrams S, Lowenberg D, Simms S, Chodacki J (2018) Support Your Data: A Research Data Management Guide for Researchers. Research Ideas and Outcomes 4: e26439. https://doi.org/10.3897/rio.4.e26439

Research Data Management Guide for Researchers

	Ad Hoc	One-Time	Active and Informative	Optimized for Re-Use
Planning your project	When it comes to my data, I have a "way of doing things" but no standard or documented plans.	I create some formal plans about how I will manage my data at the start of a project, but I generally don't refer back to them.	I develop detailed plans about how I will manage my data that I actively revisit and revise over the course of a project.	I have created plans for managing my data that are designed to streamline its future use by myself or others.
Organizing your data	I don't follow a consistent approach for keeping my data organized, so it often takes time to find things.	I have an approach for organizing my data, but I only put it into action after my project is complete.	I have an approach for organizing my data that I implement prospectively, but it not necessarily standardized.	I organize my data so that others can navigate, understand, and use it without me being present.
Saving and backing up your data	I decide what data is important while I am working on it and typically save it in a single location.	I know what data needs to be saved and I back it up after I'm done working on it to reduce the risk of loss.	I have a system for regularly saving important data while I am working on it. I have multiple backups.	I save my data in a manner and location designed maximize opportunities for re-use by myself and others.

If you can share with yourself and your colleagues, you will be much better positioned to share with the community at large

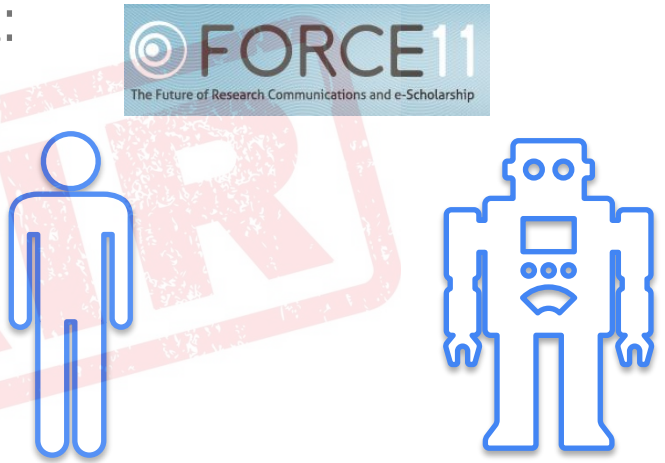
Borghi J, Abrams S, Lowenberg D, Simms S, Chodacki J (2018) Support Your Data: A Research Data Management Guide for Researchers. Research Ideas and Outcomes 4: e26439. https://doi.org/10.3897/rio.4.e26439

But how?

The FAIR Guiding Principles for scientific data management and stewardship

High level principles to make data:

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**e-usable



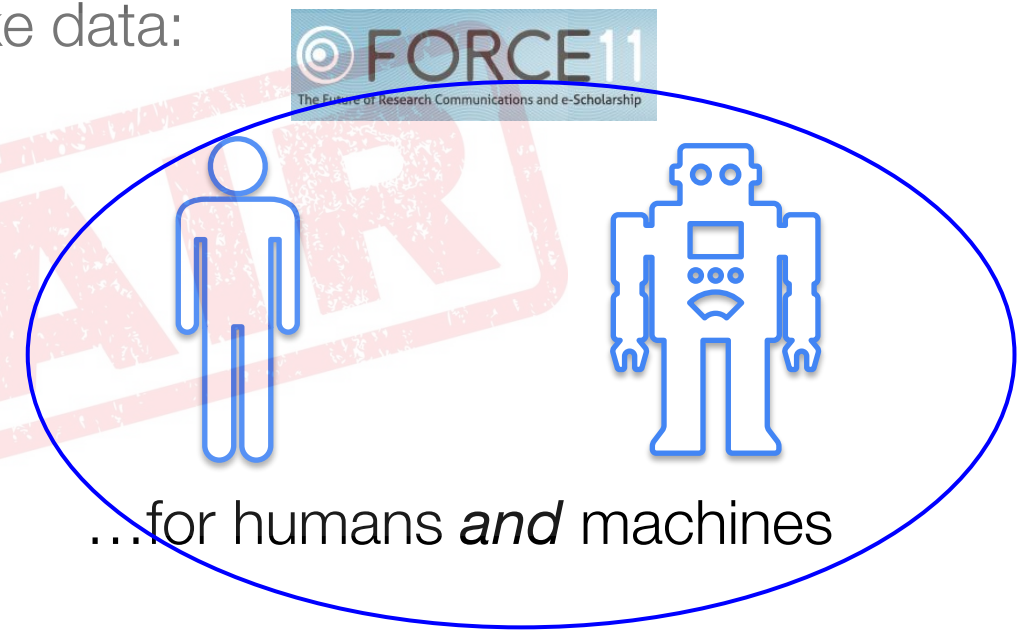
…for humans *and* machines

Mark D. Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* (2016). DOI: 10.1038/sdata.2016.18

The FAIR Guiding Principles for scientific data management and stewardship

High level principles to make data:

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**e-usable



Mark D. Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* (2016). DOI: 10.1038/sdata.2016.18

Findable

- F1. (meta)data are assigned a *globally unique and persistent* identifier
- F2. data are described with rich metadata
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

Accessible

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
- A1.1 the protocol is open, free, and universally implementable
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. *metadata are accessible, even when the data are no longer available*

Interoperable

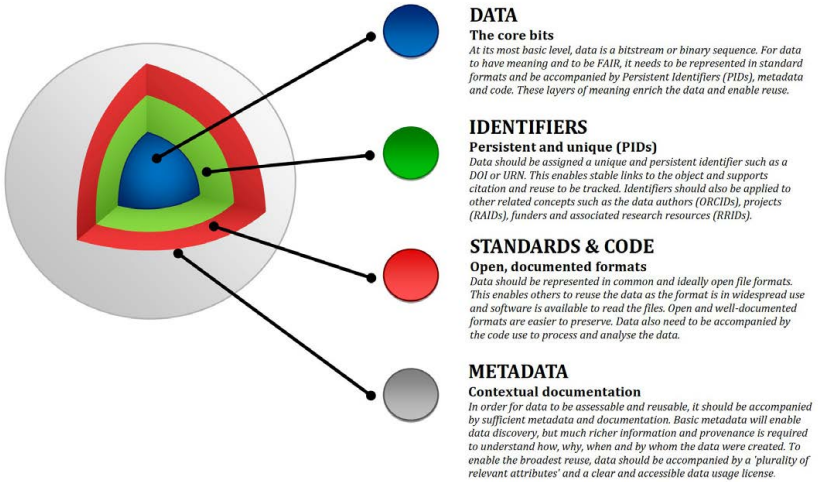
- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

Re-usable

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1. (meta)data are released with a clear and accessible data usage license
- R1.2. (meta)data are associated with detailed provenance
- R1.3. (meta)data meet domain-relevant community standards

FAIR in a nutshell

1. Well documented data are easier to find, understand and use
2. Properly formatted data are easier to use in a variety of software
3. Data that are linked to other research objects like related data, code, publications, protocols are easier to use
4. Data that are shared in established, trustworthy repositories are easier to find and use.



Adapted from: White EP, Baldridge E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR. *Nine simple ways to make it easier to (re)use your data Ideas in Ecology and Evolution.* 2013:6.

Turning FAIR data into reality
Interim report from the European Commission Expert Group on FAIR data

FAIR Partnership

incf enabling open and FAIR neuroscience

Review, endorse and organize standards

- **Good data management**
- **Rich metadata**
- **Prepare to share**
- **Well documented**
- **Open formats**
- **Adopt/align to standards**
- **Submit to repository**

- Persistent identifier
- Machine based access
- Clear license
- Support for open, domain specific standards
- Machine readable metadata
- Future friendly formats
- Persistent metadata
- Bidirectional links
- Data citation

- Index
- Effective Search
- Persistent metadata

F2: Data are described with **rich** metadata; R1. meta(data) are richly described with a **plurality** of accurate and relevant attributes

- What exactly is rich metadata and a plurality?
- “I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description, and perhaps I could never succeed in intelligibly doing so. But *I know it when I see it*...”- Justice Stewart Potter

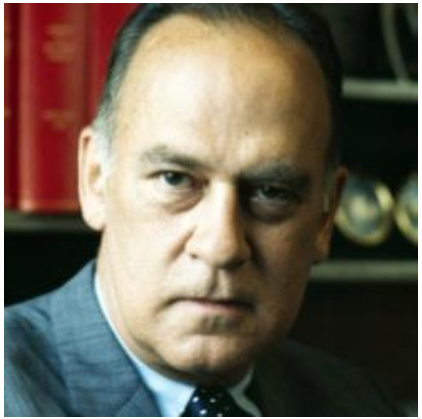
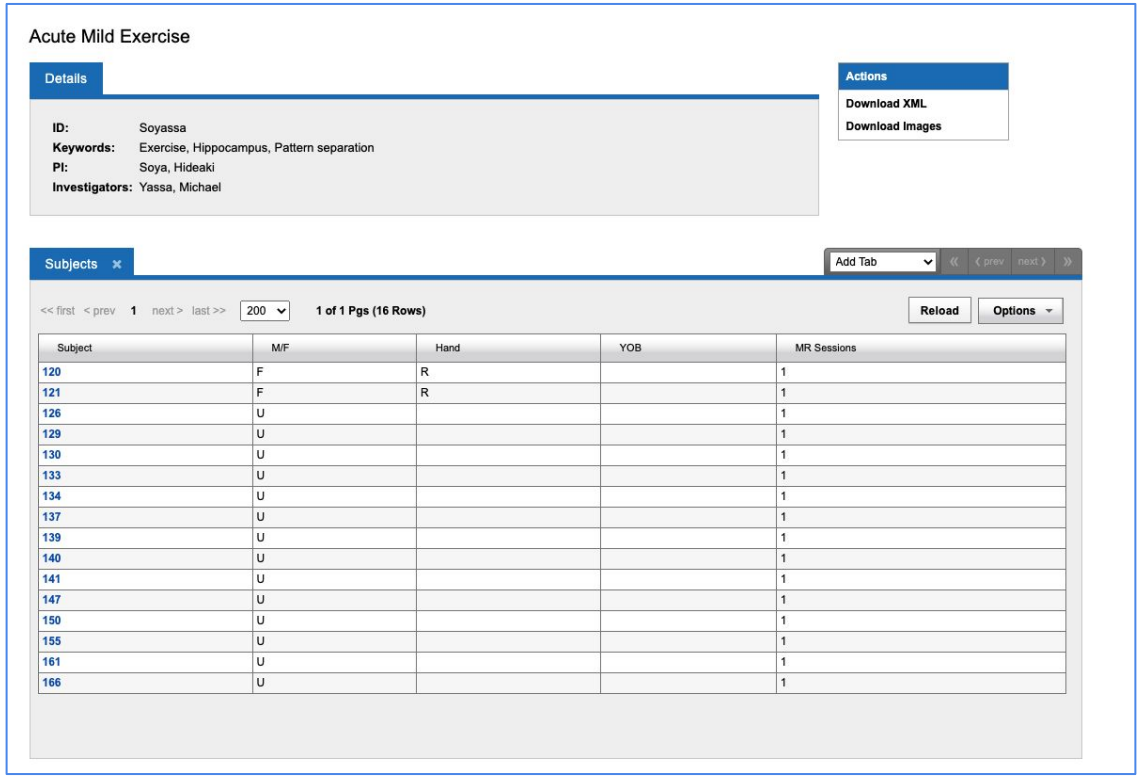


Exhibit A



Acute Mild Exercise

Details

ID: Soyassa
Keywords: Exercise, Hippocampus, Pattern separation
PI: Soya, Hideaki
Investigators: Yassa, Michael

Actions
Download XML
Download Images

Subjects

Add Tab

<< first < prev 1 next > last >> 200 1 of 1 Pgs (16 Rows)

Reload Options

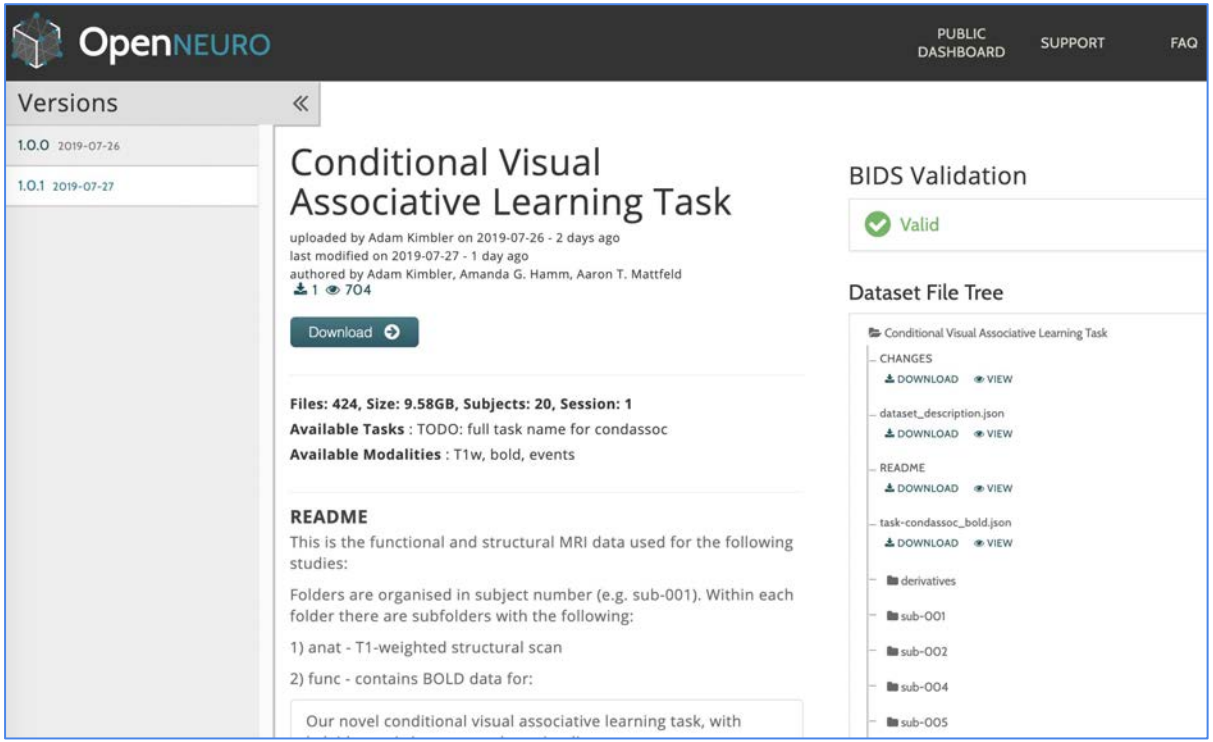
Subject	M/F	Hand	YOB	MR Sessions
120	F	R		1
121	F	R		1
126	U			1
129	U			1
130	U			1
133	U			1
134	U			1
137	U			1
139	U			1
140	U			1
141	U			1
147	U			1
150	U			1
155	U			1
161	U			1
166	U			1

If you can share with yourself and your colleagues, you will be much better positioned to share with the community at large

https://central.xnat.org/data/projects/Soyassa

Exhibit B

- Meaningful title and description
- Study purpose
- Technique
- Authors and contributors
- Citation
- Instructions on use
- Versions
- Access rights
- Subjects and other study attributes
- Links to other useful research objects



https://openneuro.org/datasets/ds002078/versions/1.0.1

"Rich metadata" recommendation

R1.2. (meta)data are associated with detailed **provenance**; I3. (meta)data include **qualified references** to other (meta)data

- Who produced the data?
- Where was it produced?
- Why was it produced?
- What produced it?
- How was it produced?

Create and manage FAIR protocols in Protocols.io; code in GitHub

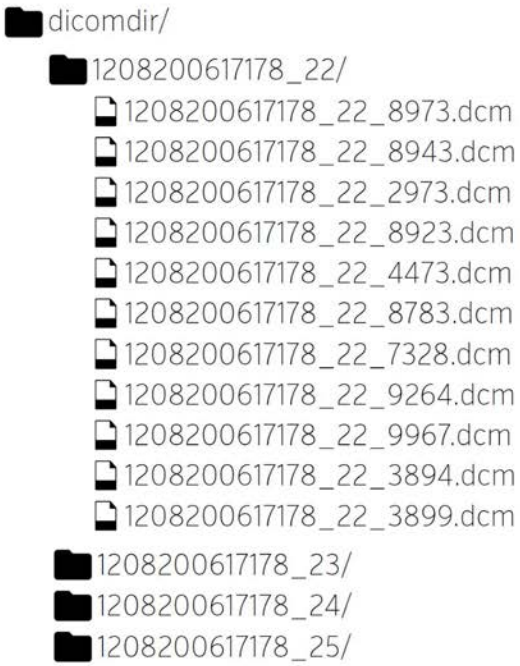
R1.3. (meta)data meet domain-relevant **community standards**

INCF endorsement: Standard has to be well documented and present evidence of use by the target audience beyond the group that developed it.

https://www.incf.org/resources/sbps

From DICOM to BIDS: File organization and naming

DICOM dataset



BIDS dataset

BIDS was designed for and with significant input from the researchers who were expected to use it!

BIDS has a formal extension process; 9 approved, > 20 in process

Slide courtesy of Russ Poldrack

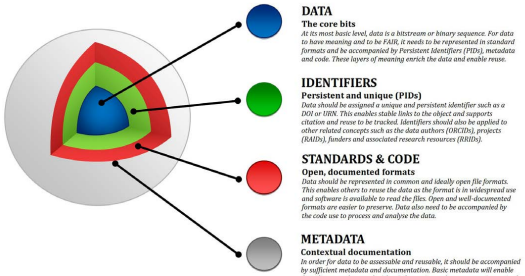
Benefits of using a standard

- It will be easy for **another researcher** to work on your data...This is especially important if you are **running your own lab** and anticipate more than one person working on the same data over time. By using BIDS you will save time trying to understand and reuse data acquired by a **graduate student or postdoc that has already left the lab**.
- There are a growing number of data **processing pipelines and data analysis software packages** that can understand data organized according to BIDS.
- Databases such as OpenNeuro.org, LORIS, COINS, XNAT, SciTran, and others will accept and export datasets organized according to BIDS. **If you ever plan to share your data publicly (nowadays some funders and journals require this) you can speed up the curation process by using BIDS.**
- There are **validation tools** (also available online) that can check your dataset integrity and let you easily spot missing values.

https://bids.neuroimaging.io/benefits

FAIR in a nutshell

1. Well documented data are easier to understand
    a. *Authoring and description: Who, where, why*
    b. *Experimental details (how)*
    c. *Data dictionaries, Read Me (what)*
2. Properly formatted data are easier to use in a variety of software
    a. *Standards*
    b. *Open formats*
    c. *File naming conventions*
3. Data that are linked to other research objects like code, publications, protocols are easier to reuse
    a. *Never pass up an opportunity to use a PID!*
4. Data that are shared in established, trustworthy repositories are easier to find and use.
    a. *PIDs*
    b. *Data stewardship: Local shared resources and global repositories*



Turning FAIR data into reality
Interim report from the European Commission Expert Group on FAIR data

Adapted from: White EP, Baldridge E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR. *Nine simple ways to make it easier to (re)use your data Ideas in Ecology and Evolution.* 2013:6.

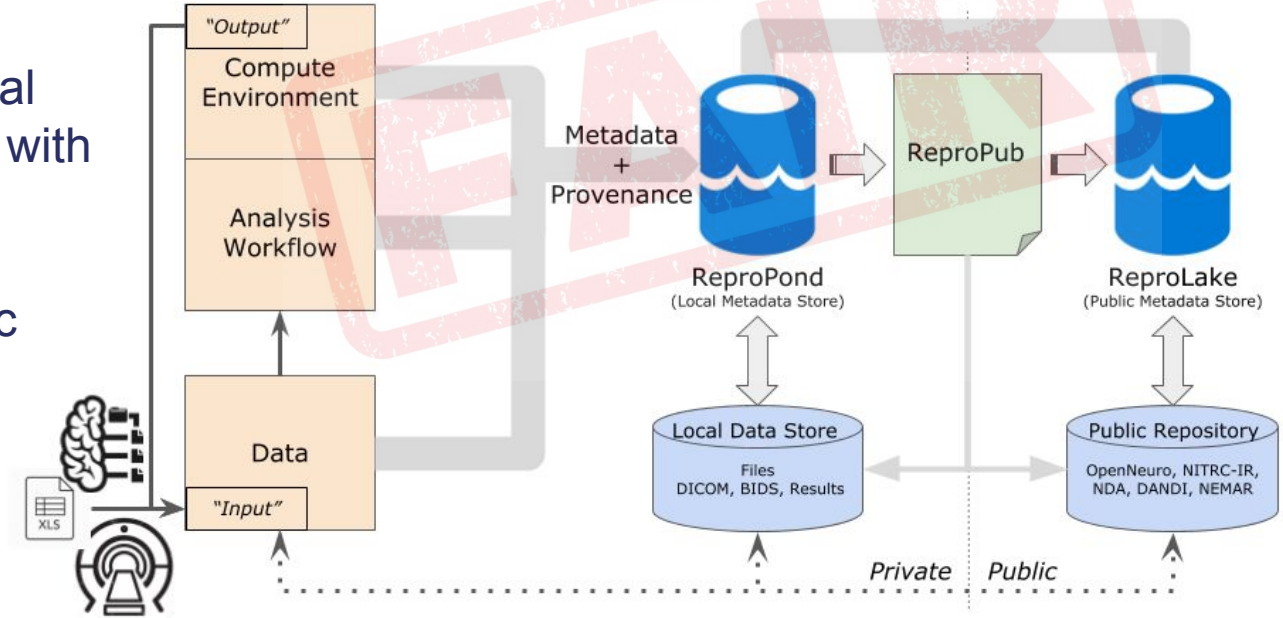
Resources

- dkNET.org: Data management and FAIR data resources
- INCF Standards Portfolio: Reviewed standards for neuroscience, linked to tools and training materials
- FAIRSharing.org: Data and metadata standards linked to data policies and databases
- Brain Imaging Data Structure: BIDS specifications, training, tools, governance and extensions
- Research Data Management Toolkit: Elixir; an online guide containing good data management practices applicable to research projects from the beginning to the end.
- Check your local library for data management training!

The ReproNIM data lakes

ABCD-ReproNim

- ReproPond - Local store of data and metadata
- ReproLake - Global store of metadata with pointers to MR data/images (e.g. contained in public repositories)

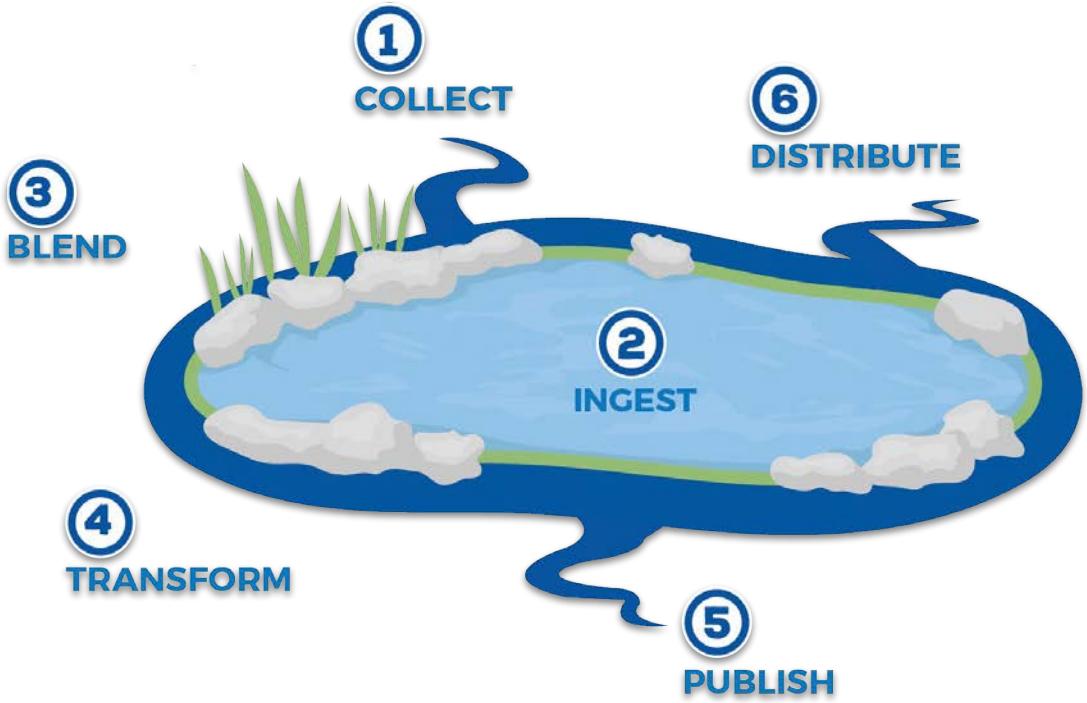


FAIR and you

- FAIR is defining best practices and community norms for publishing data for reuse
- NIH data policies are being revised and routine data sharing will become the norm
- FAIR takes ***effort, time and resources***
- The one most likely to benefit from FAIR data practices is you, your PI and "future you"
  - Well documented
  - Standard vocabularies and formats
  - Rich metadata

Data Lake

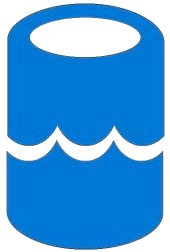
A data lake is usually a single store of all enterprise data including raw copies of source system data and transformed data used for tasks such as reporting, visualization, advanced analytics and machine learning. A data lake can include structured data from relational databases (rows and columns), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs) and binary data (images, audio, video).



https://en.wikipedia.org/wiki/Data_lake

The ReproNim ReproLake Infrastructure

Suite of ReproNim discovery tools that provides for search across the ReproLake and ReproPonds via query services and APIs

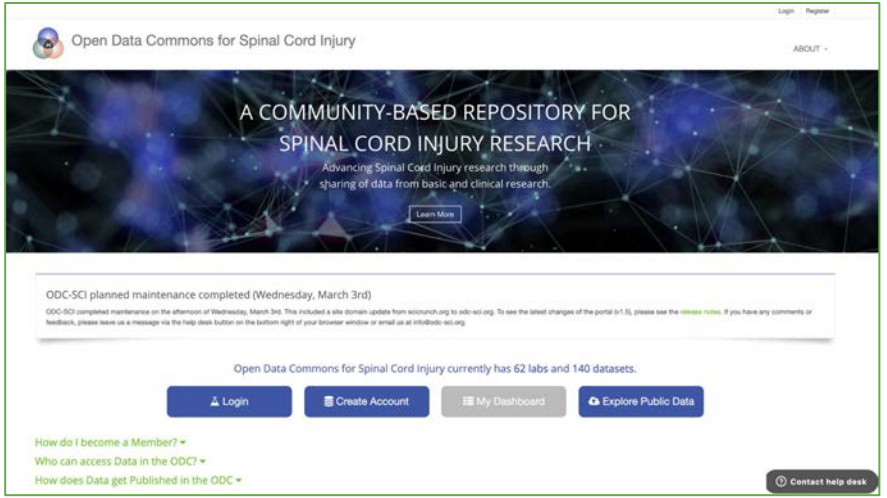


ReproPond
(Local Metadata Store)

ReproLake
(Global Metadata Store)
(Public Metadata Store)

Community repositories

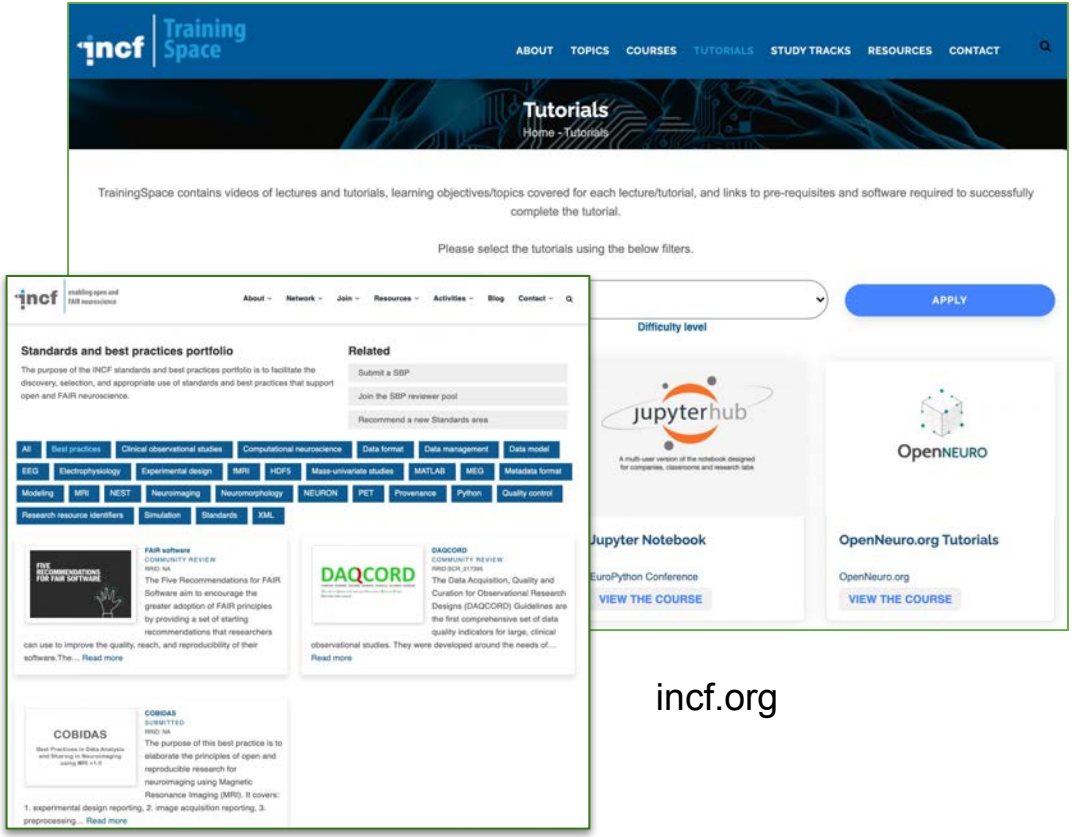
- Best practice: submit your data to a recognized community repository
- Next best: submit your data to an institutional or generalist repository
- Where can I find a repository?
  - NLM listing
  - Journal listings
  - dkNET listing
  - NIF search
  - FAIRsharing and re3data



Adam Ferguson, Karim Fouad and ODC-SCI Steering Committee

Training: Data management and data sharing

- Check your library!
- INCF Training Space
- Data management tools
  - California Digital Library
  - Portage DMP assistant
- Data management course
- INCF Standards and Best Practices Portfolio

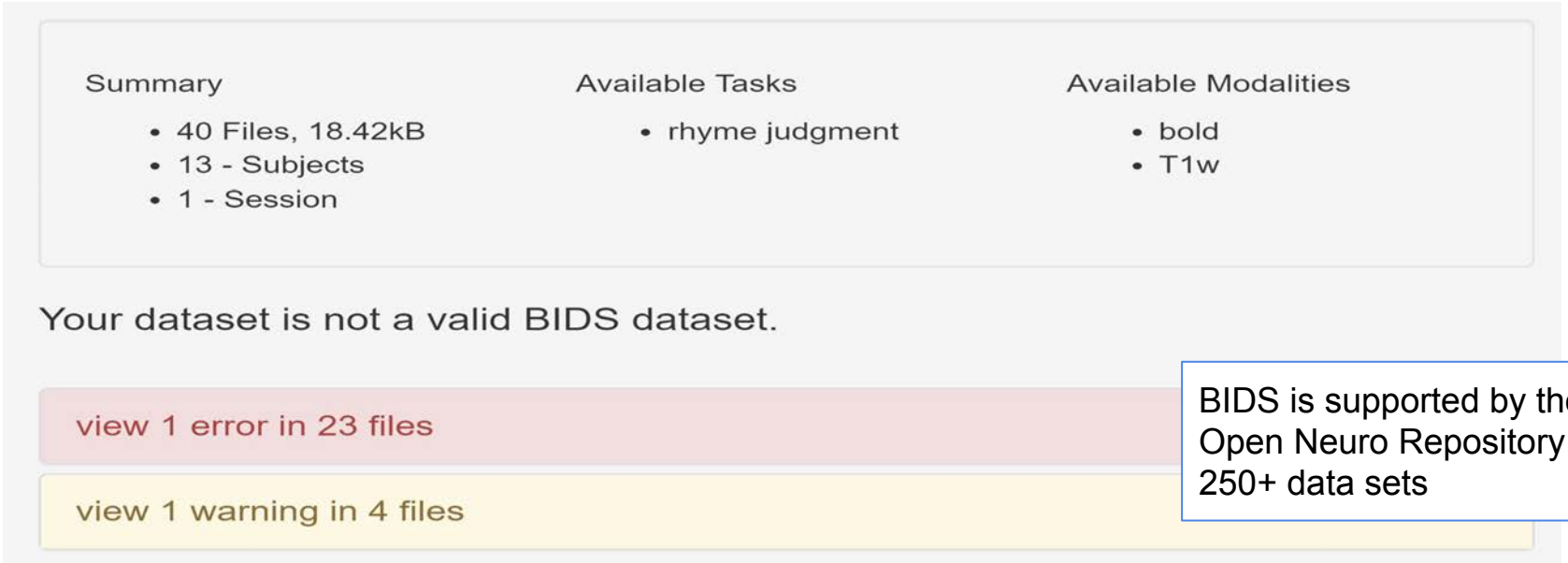


incf.org

The importance of automated validation

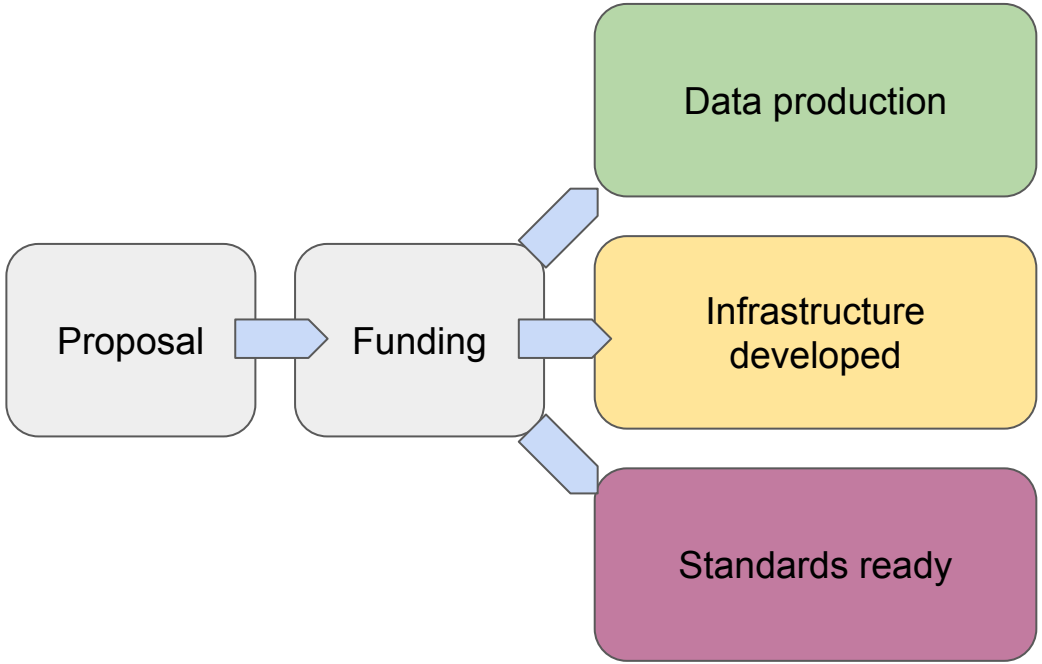
- A standard is only useful if a user can easily tell whether their data are compliant
- Validation is crucial for automated data ingestion

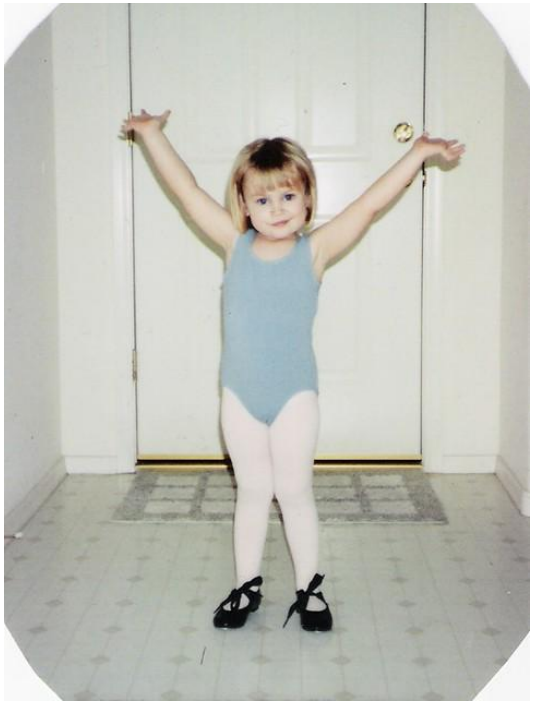
https://bids-standard.github.io/bids-validator/



BIDS is supported by the Open Neuro Repository: 250+ data sets

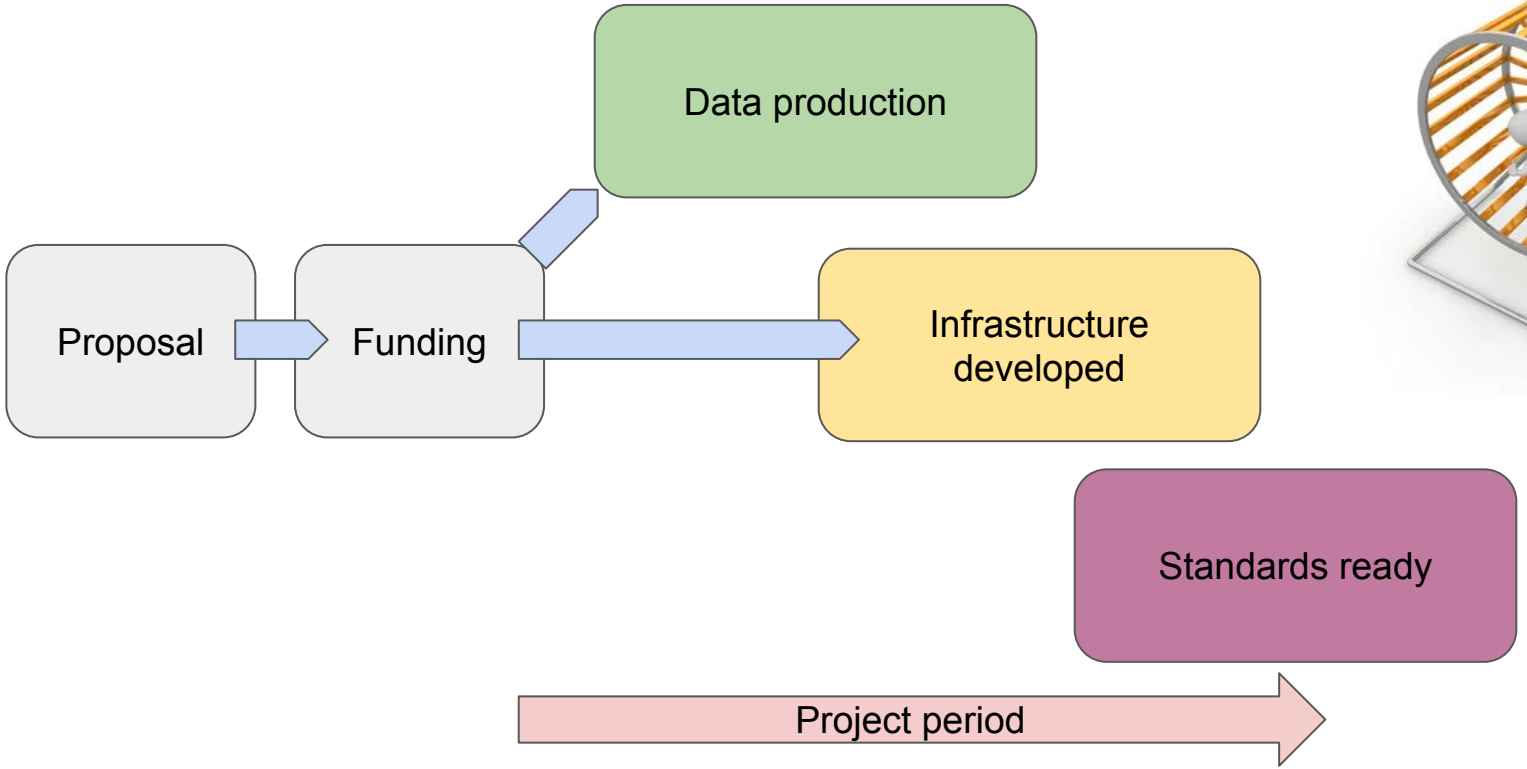
A typical biomedical project





We don’t want to reinvent the wheel so we’ll use existing standards.

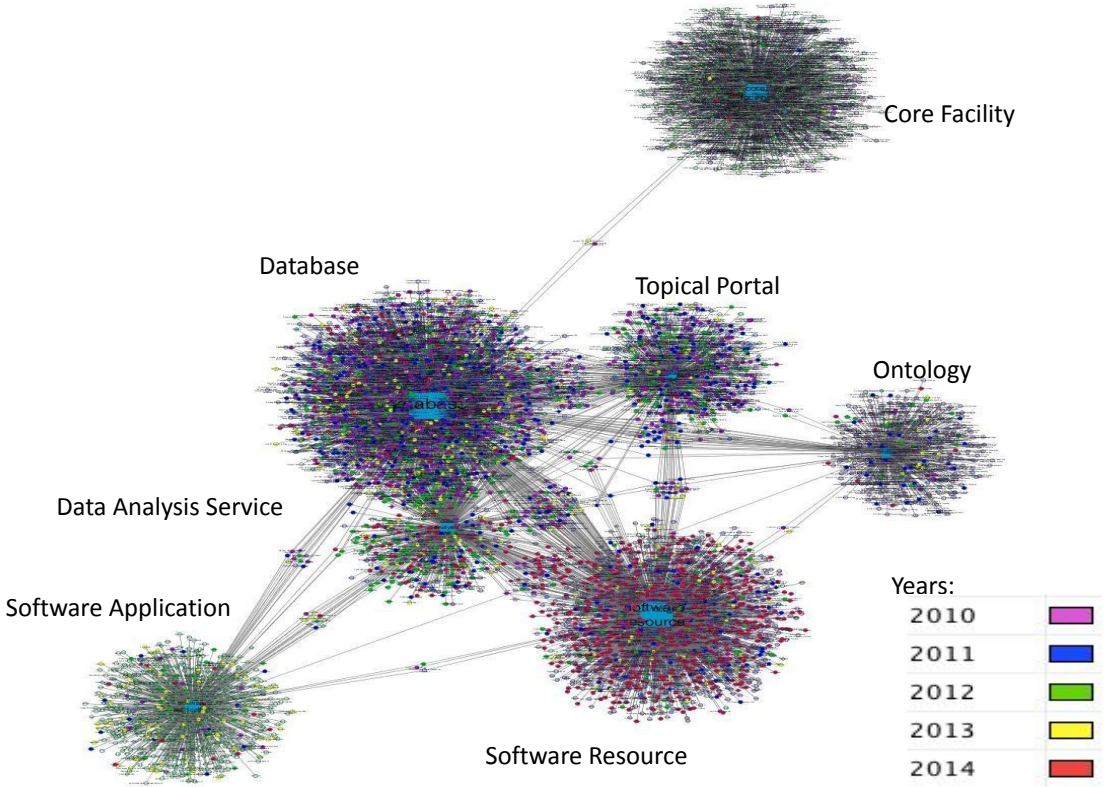
But what do I do now?



“We’re developing a standard! Great, when will it be ready? Three years”.

NIF

NEUROSCIENCE INFORMATION FRAMEWORK



NIF is an initiative of the NIH Blueprint consortium of institutes

- NIF has been tracking and cataloging the biomedical resource landscape since 2006
- > 15,000 resources in the Registry
  - 800M records from 264 sources in federation
  - NIF ontologies for neuroscience

http://neuinfo.org

FORCE11

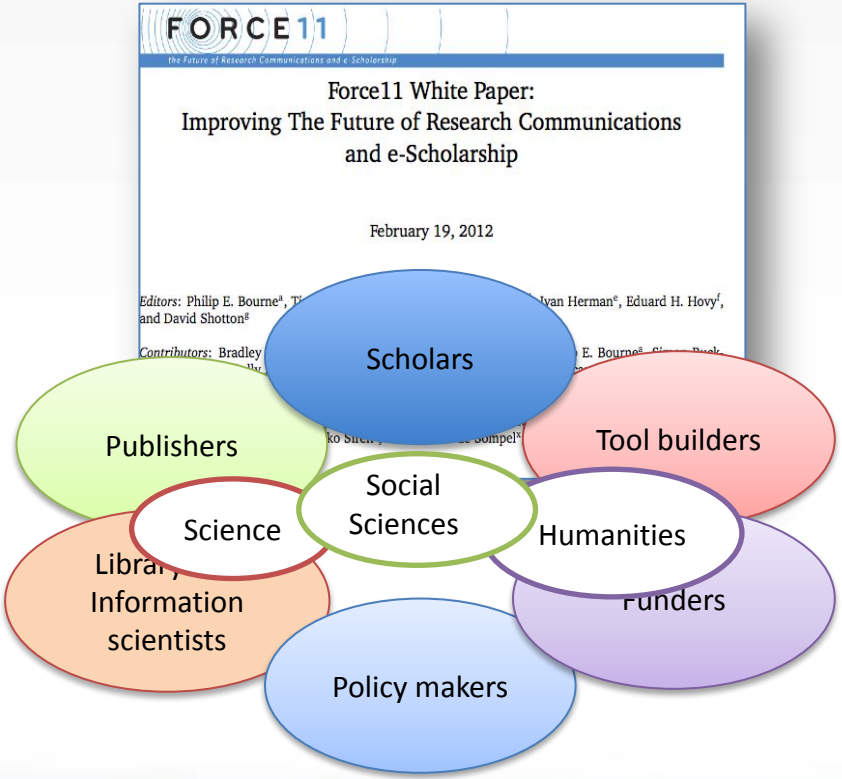
Future of Research Communications and E-Scholarship:

A grass roots effort to accelerate the pace and change the nature of scholarly communications and e-scholarship through technology, education and community

Why 11? We were born in 2011 in Dagstuhl, Germany

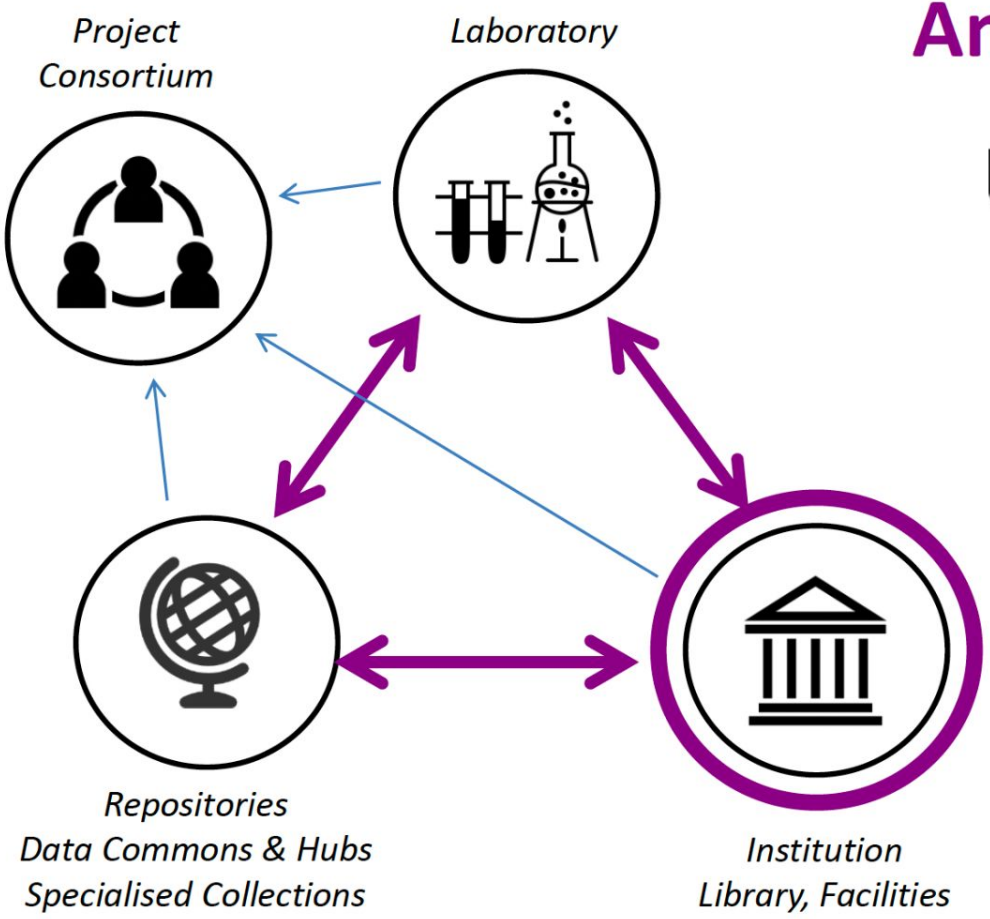
Principles laid out in the FORCE11 Manifesto

FORCE11 launched in July 2012



A data sharing tragedy in 3 acts

https://www.youtube.com/results?search_query=data+sharing+tragedy+in+3+acts



An interlinked triangle

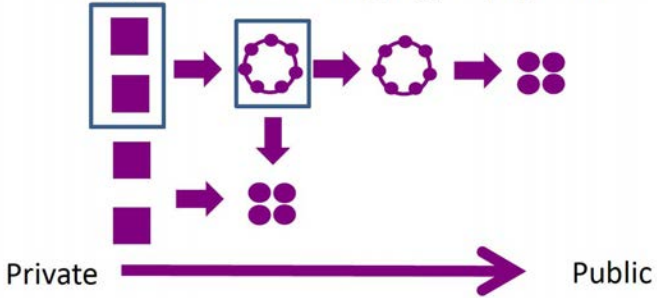
Trusted Research Environments

Access to analysis platforms

Support and encourage **incremental staged sharing** & permission control

- To get trust in data sharing and support team sharing

Support Data Staging & pipelines



Adapted from Carol Goble, NASM Workshop

Attitudes towards data sharing

Oh, alright

- Funders are making me
- The journal is making me

I don’t want to

- Fear of being scooped
- Fear of mistakes
- Too much time and effort relative to rewards
- Show me the money

I don’t know how

- What should I share?
- How should I share it?
- How will I get credit?

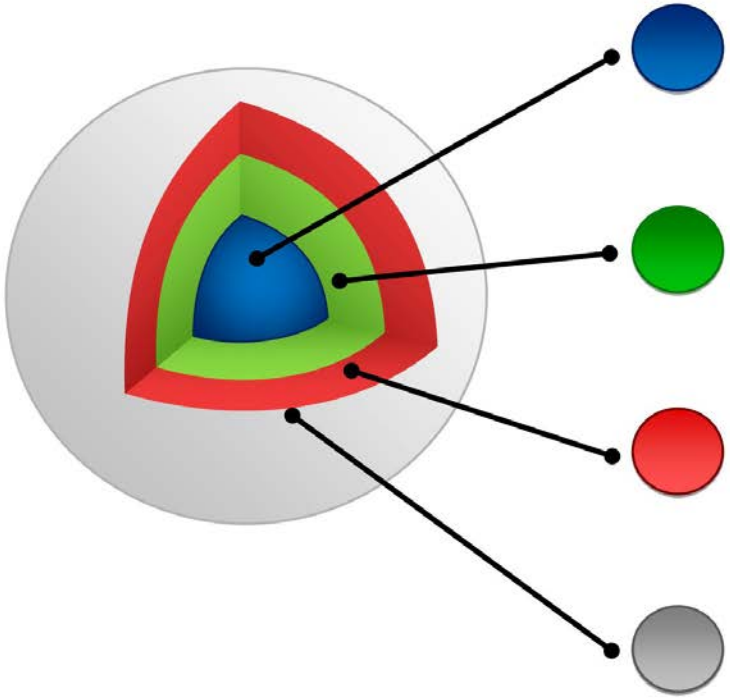
Data management as the gateway

- Data management plan
- Good data management practices
- Good data stewardship
- Machine actionable data
- Documentation
- Annotation

Findable

- F1. (meta)data are assigned a *globally unique and persistent* identifier
- F2. data are described with rich metadata
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

FAIR in a nutshell



DATA

The core bits

At its most basic level, data is a bitstream or binary sequence. For data to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and code. These layers of meaning enrich the data and enable reuse.

IDENTIFIERS

Persistent and unique (PIDs)

Data should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE

Open, documented formats

Data should be represented in common and ideally open file formats. This enables others to reuse the data as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

METADATA

Contextual documentation

In order for data to be assessable and reusable, it should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the data were created. To enable the broadest reuse, data should be accompanied by a 'plurality of relevant attributes' and a clear and accessible data usage license.

Fig 6, Turning FAIR data into reality
Interim report from the European Commission Expert Group on FAIR data

Brain Imaging Data Structure (BIDS)

- What is BIDS?
  - A file organization standard
  - A metadata standard

Slides courtesy of Russ Poldrack

Research Data Management Guide for Researchers

	Ad Hoc	One-Time	Active and Informative	Optimized for Re-Use
Planning your project	When it comes to my data, I have a "way of doing things" but no standard or documented plans.	I create some formal plans about how I will manage my data at the start of a project but I generally do not refer back to	I develop detailed plans about how I will manage my data and actively revisit and revise over the course of a project	I have created plans for managing my data that are designed to streamline its future use by myself or others.
Organizing your data	I don't follow a consistent approach for keeping my data organized, so it often takes time to find things.	I have an approach for organizing my data, but I only put it into action after my project is complete.	I have an approach for organizing my data that is implemented respectively, but it not necessarily standardized.	I organize my data so that others can navigate, understand, and use it without me being present.
Saving and backing up your data	I decide what data is important while I am working on it and typically save it in a single location.	I know what data needs to be saved and I back it up after I'm done working on it to reduce the risk of loss.	I have a system for saving data while I am working on it. I have multiple backups.	I save data in a location designed to maximize opportunities for re-use by myself and others.

Borghi J, Abrams S, Lowenberg D, Simms S, Chodacki J (2018) Support Your Data: A Research Data Management Guide for Researchers. Research Ideas and Outcomes 4: e26439. https://doi.org/10.3897/rio.4.e26439

BIDS and NIDM

BIDS – Raw

- Layout-based representation of data
- Focused on capturing the most commonly used metadata necessary for sharing datasets and running common imaging workflows
- Metadata stored in human readable json and tsv file formats and encoded in file / directory names
- Imaging data stored as NifTI, with format for other modalities still under discussion

NIDM-E

- Layout-agnostic structured information about data
- Specifications for constructing unambiguous metadata descriptions for finding and combining data across studies, modalities, and sources
- Metadata stored as graphs, serialized to any Resource Description Framework (RDF) format
- Agnostic to imaging data format
- Rich query support via semantic-web tools (e.g. SPARQL)

BIDS – Derivatives

- Layout-based representation for derived data
- Focused (currently) on common neuroimaging-based workflows
- Derivative metadata encoded in file/directory names with fixed vocabularies

NIDM-R and NIDM-W

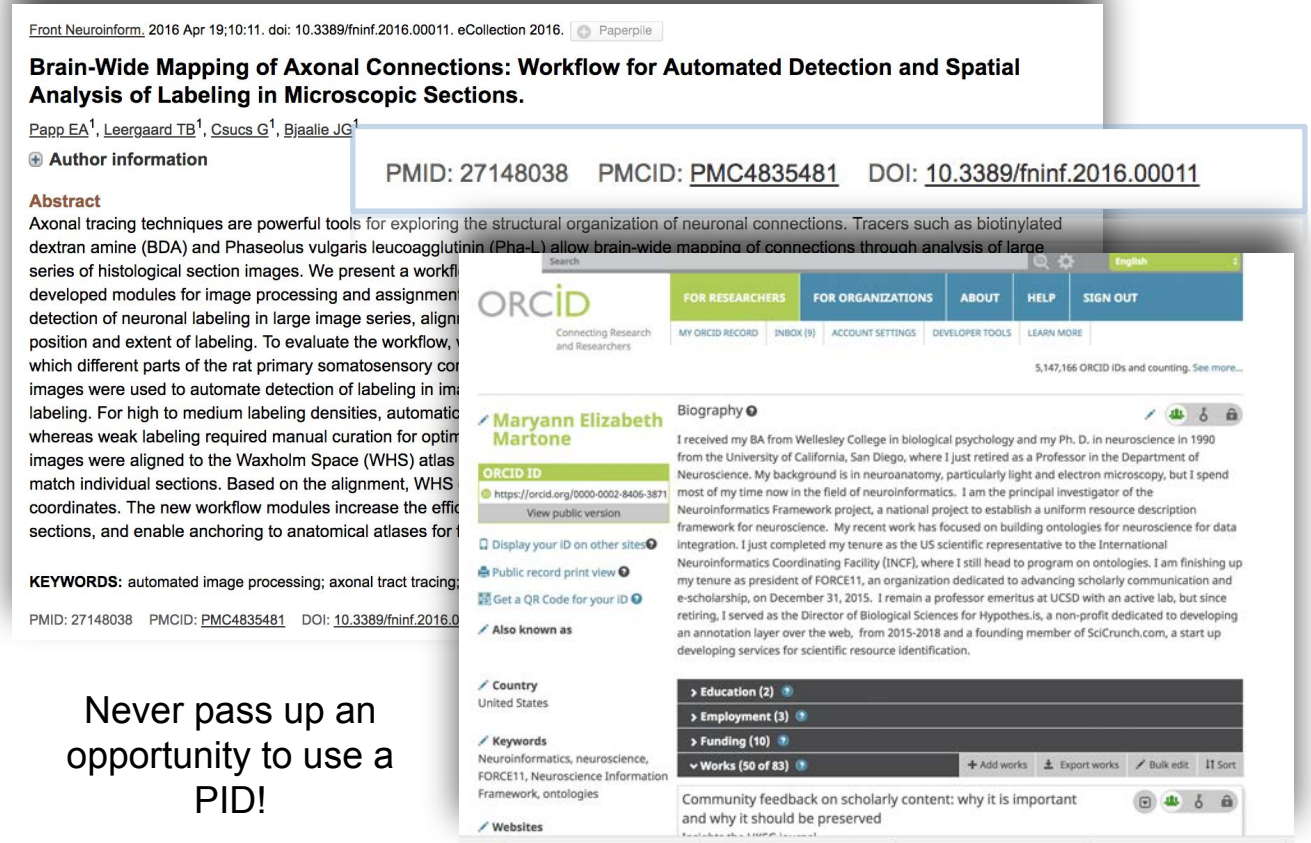
- NIDM-R: Layout-agnostic structured information about derived results from workflows
- NIDM-W: Precise descriptions of workflow, parameters, etc. used to generate derived result

BIDS and NIDM are synergistic!
BIDS helps you organize your data
NIDM helps you unambiguously describe your data

Data are assigned a persistent identifier. Huh?

- *“Principle F1 is arguably the most important because it will be hard to achieve other aspects of FAIR without globally unique and persistent identifiers. Hence, compliance with F1 will already take you a long way towards publishing FAIR data”*-GoFAIR
  - Unique in the world + Stable (persistent)
    - Identify only a single object for all time (never reused)
    - Only persistent and unique because organizations stand behind them***
  - Can be resolvable, i.e., you can plug it into a web browser and be taken to the object *independent of its location*
  - Allows a digital object to be reliably tied to its metadata

The world of unique and persistent identifiers



Never pass up an opportunity to use a PID!

- DOI: Digital object identifier
- ORCID: Researcher identifier
- RRID: Resource Identifier
- Globally unique: identifies one thing only
- Unlike URL's or catalog numbers, may *NOT* be re-used
- **Issued by registries who track and identify unique entities**

Findable

- **F1. (meta)data are assigned a *globally unique and persistent* identifier**
- **F2. data are described with rich metadata**
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

Accessible

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
- A1.1 the protocol is open, free, and universally implementable
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. *metadata are accessible, even when the data are no longer available*

Interoperable

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

Re-usable

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1. (meta)data are released with a clear and accessible data usage license
- **R1.2. (meta)data are associated with detailed provenance**
- **R1.3. (meta)data meet domain-relevant community standards**