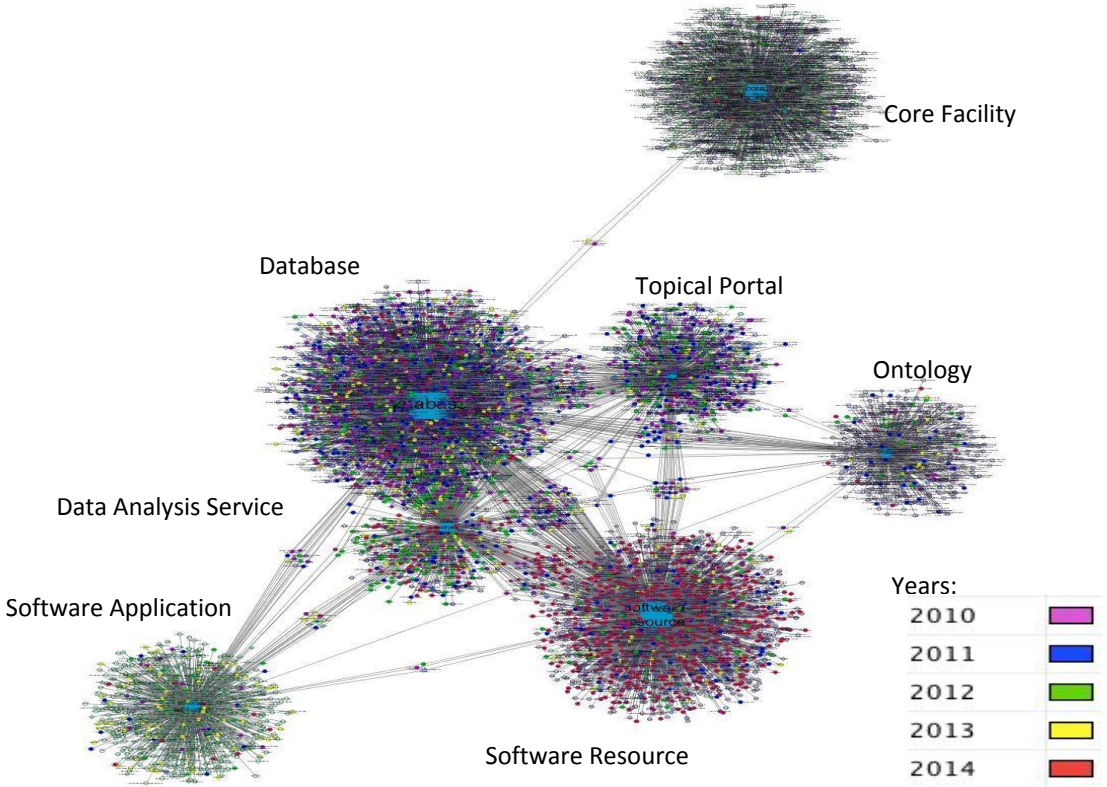
The FAIR Data Principles

Maryann E. Martone
Prof Emerita
University of California San Diego
CSO and Founder
SciCrunch Inc*

*develops tools and services around Research Resource Identifiers (RRIDs) to support reproducible science

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

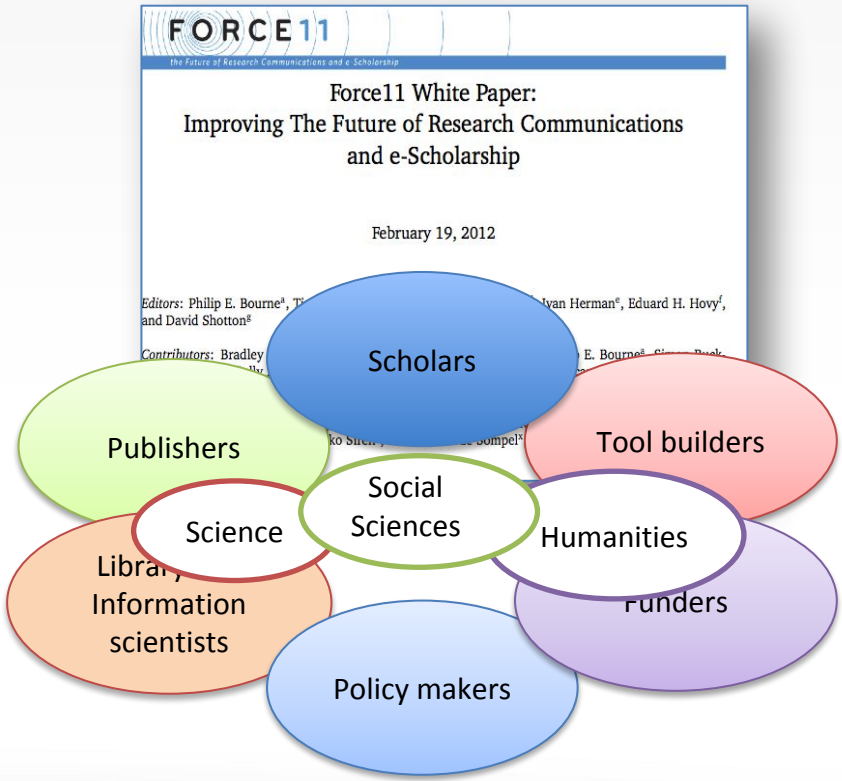
FORCE11

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? FORCE11 was born in 2011 in Dagstuhl, Germany

Principles laid out in the [FORCE11 Manifesto]



Data as a Research Product

Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record…"

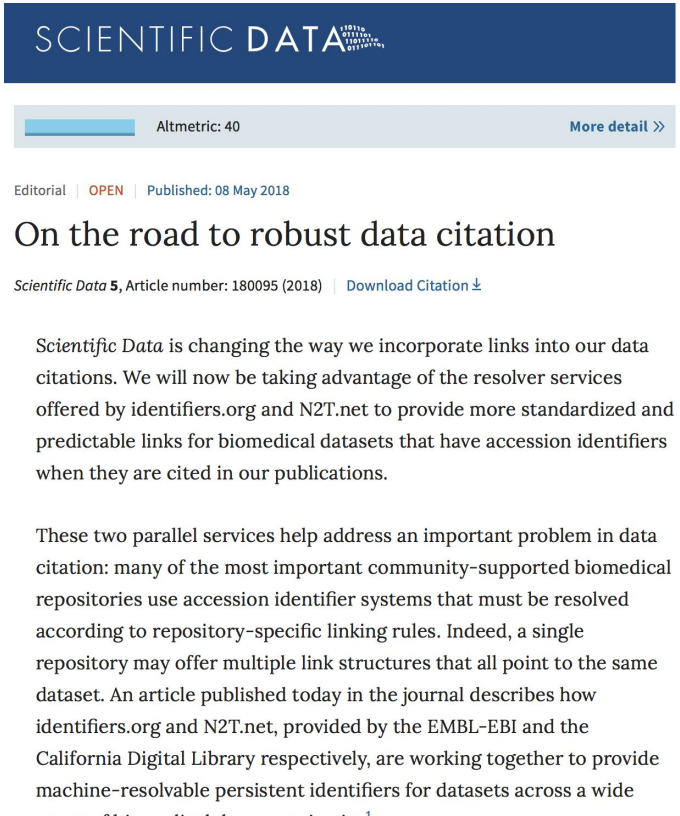
1. Data should be considered *legitimate, citable products of research*. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.
2. Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.
3. In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.

DC[1] Data Citation Principles

Joint Declaration of Data Citation Principles

https://www.force11.org/group/joint-declaration-data-citation-principles-final

An Ecosystem for Data Citation



SCIENTIFIC DATA

Altmetric: 40 More detail »

Editorial | OPEN | Published: 08 May 2018

On the road to robust data citation

Scientific Data **5**, Article number: 180095 (2018) | Download Citation

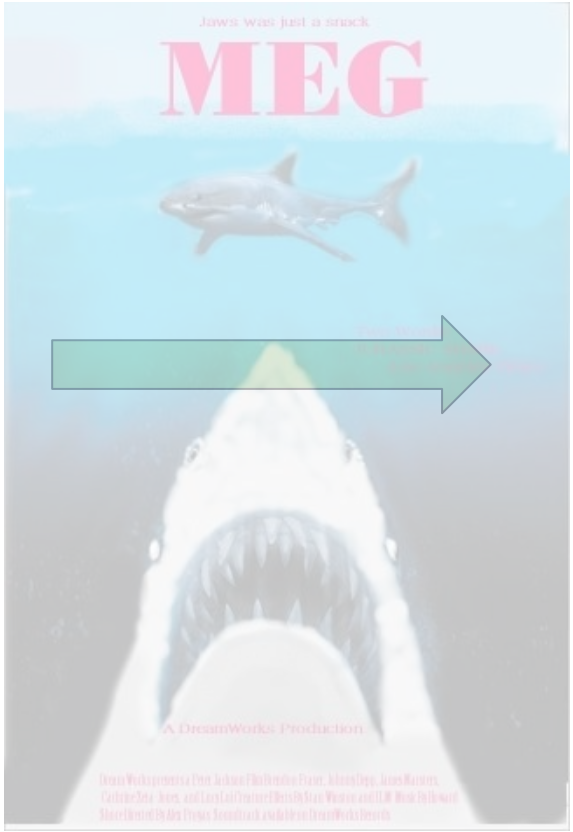
Scientific Data is changing the way we incorporate links into our data citations. We will now be taking advantage of the resolver services offered by identifiers.org and N2T.net to provide more standardized and predictable links for biomedical datasets that have accession identifiers when they are cited in our publications.

These two parallel services help address an important problem in data citation: many of the most important community-supported biomedical repositories use accession identifier systems that must be resolved according to repository-specific linking rules. Indeed, a single repository may offer multiple link structures that all point to the same dataset. An article published today in the journal describes how identifiers.org and N2T.net, provided by the EMBL-EBI and the California Digital Library respectively, are working together to provide machine-resolvable persistent identifiers for datasets across a wide range of biomedical data repositories[1].

Meta-analysis vs Mega-analysis

"Meta-analysis was created out of the need to extract useful information from the cryptic records of inferential data analyses in the abbreviated reports of research in journals and other printed sources…"

Glass, G. V. (2000) Meta-Analysis at 25.



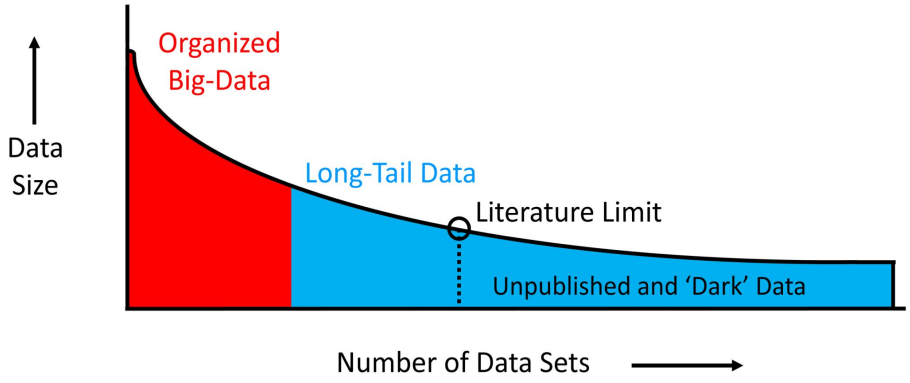
"Meta-analysis needs to be replaced by archives of raw data that permit ***the construction of complex data landscapes*** that depict the relationships among independent, dependent and mediating variables."

Glass, G. V. (2000) Meta-Analysis at 25.

Title courtesy of Drs. David Kennedy and Jean Frazier, Univ of Mass, Worcester

“Small data done right [are] big data”

- ...and it’s all small data in biomedicine
- Long-tail data: smaller data sets produced in the course of research
- How do we share data in a form that allows us *to use* data beyond the lifetime of the study
- How do we share data in a form that allows us *to more easily* combine data across studies



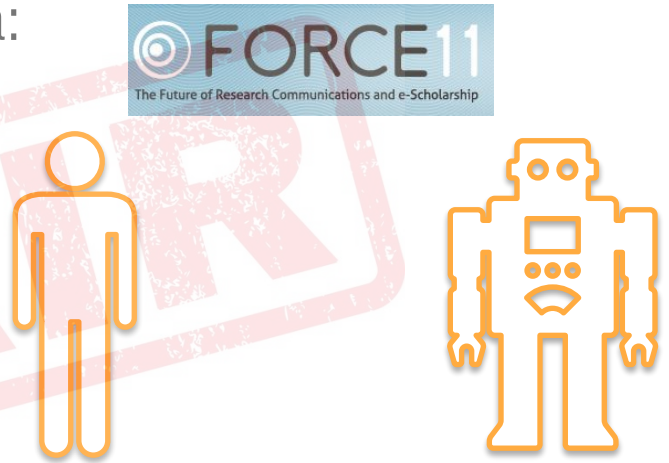
Quote: Neil McKenna, Baylor College

Ferguson et al., 2014, *Nature Neuroscience*

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

FAIR principles are

- "…characteristics that contemporary data resources, tools, vocabularies and infrastructures should exhibit to assist discovery and reuse by third-parties"-Wilkinson et al.,2016

- Recognized by major initiatives: EC-Elixir, H2020, US National Institutes of Health and G20

  - *"…we support appropriate efforts to promote open science and facilitate appropriate access to publicly funded research results* ***on findable, accessible, interoperable and reusable (FAIR) principles****"*-G20 Hangzhou Summit

FAIR principles are not…

- A standard
- Equal to RDF, Linked Data, Semantic Web or any particular implementation
- Equal to Open
- Just for life sciences

Good resource: https://www.dtls.nl/fair-data/

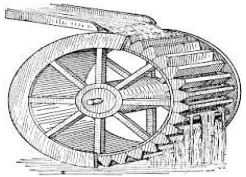
Mons et al., 2017

Why principles?

“we don't want to re-invent the wheel”



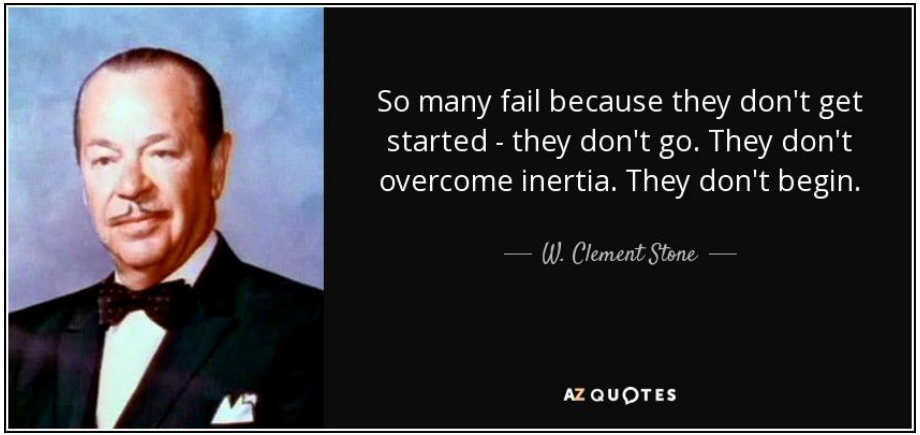







Principles provide aspirations and guidance while still respecting local needs and constraints

A closer look at FAIR



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

15 attributes of FAIR

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

Hacking culture needs to meet library culture

F1: (meta)data are assigned a *globally unique and persistent* identifier

- DOI: Digital object identifier
- ORCID: Researcher identifier
- Accession numbers
- Unlike URL's or catalog numbers, may ***NOT*** be re-used
- **Issued by registries who track and identify unique entities**

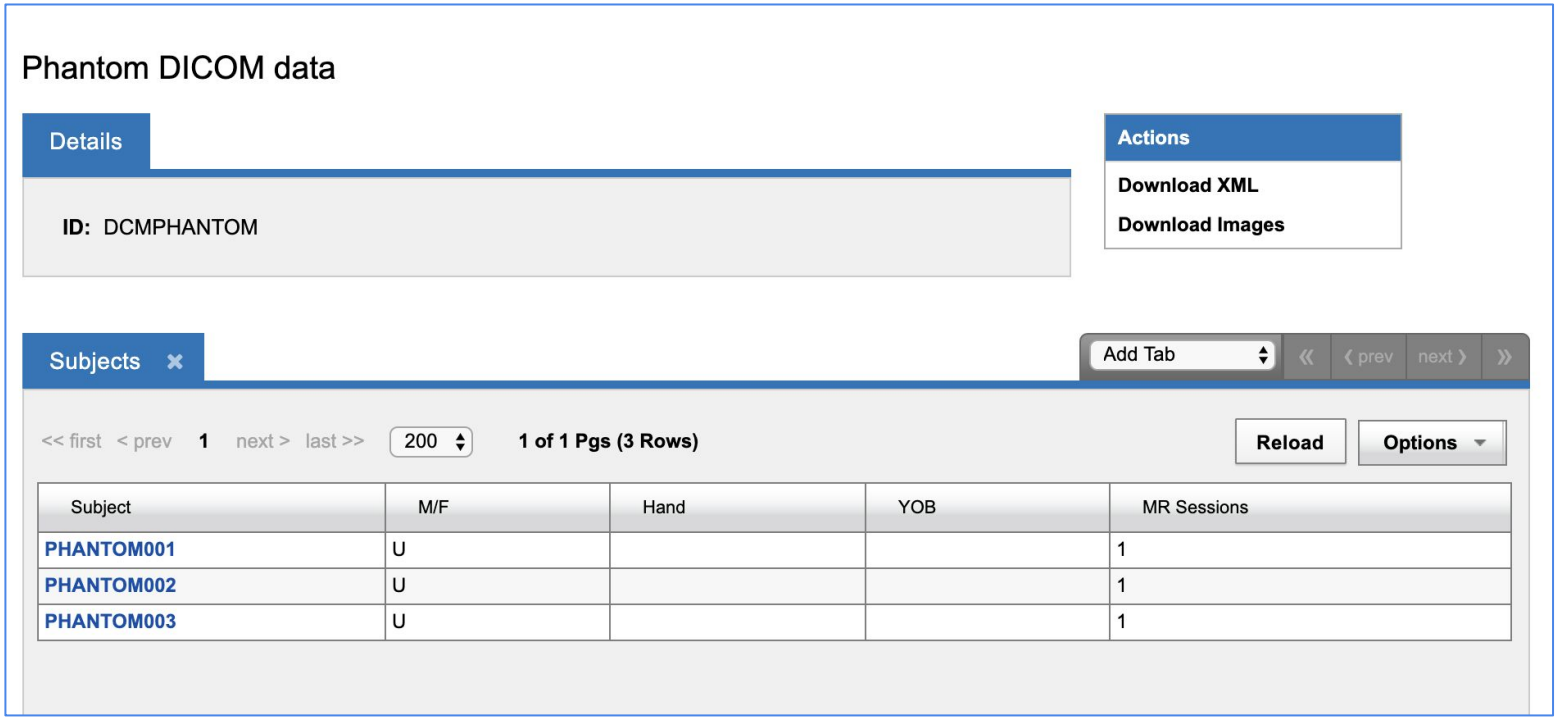
Why all this fuss about identifiers?

PIDapalooza
Reykjavik, Nov 2016

- *"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
- Stable address (persistent) + unique in the world
  - 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
- Allows an object to be reliably tied to its metadata

- PMID: 27151636 (non-actionable)
  - http://identifiers.org/pubmed/27151636
- DOI:10.1016/j.neuron.2016.04.030
  - http://dx.doi.org/10.1016/j.neuron.2016.04.030
- Globally unique vs locally unique

F2. data are (not) described with rich metadata

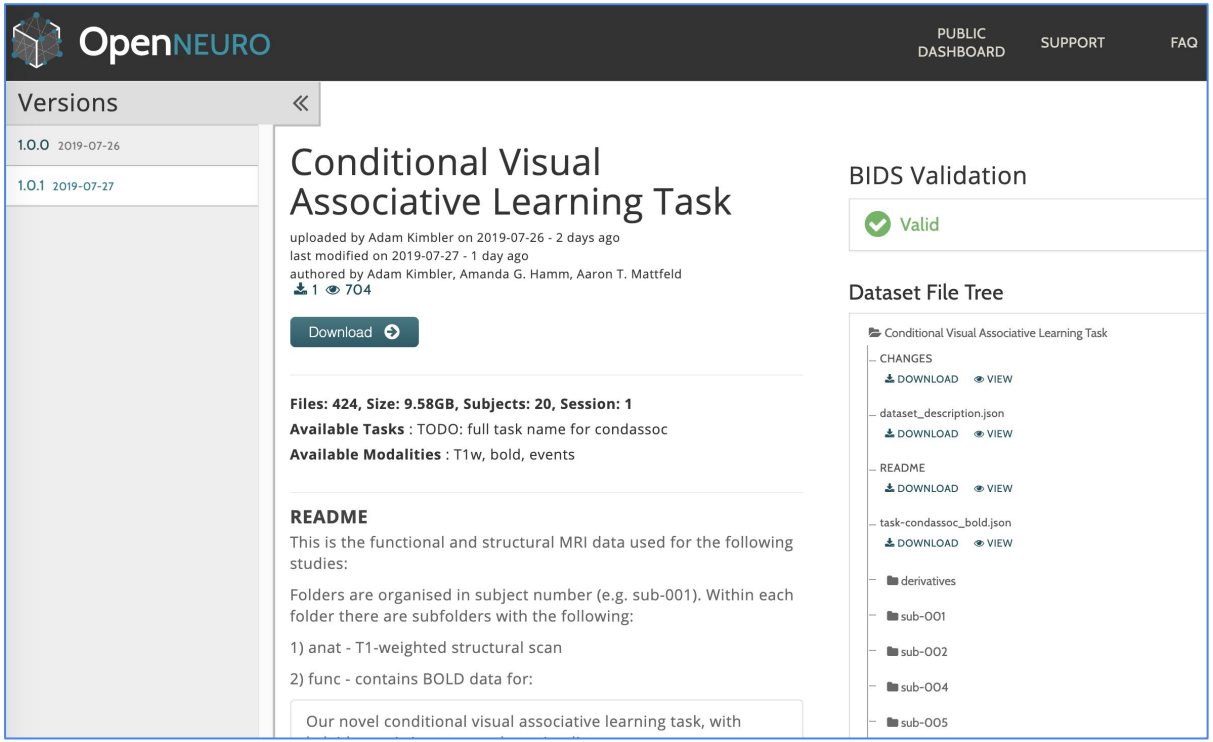


XNAT Central

Metadata: Attributes of data and information about them

F2. data are described with rich metadata

- Meaningful title and description
- Study purpose
- Technique
- Contributors
- Citation
- Instructions on use
- Versions
- Access rights
- Subjects and other study attributes



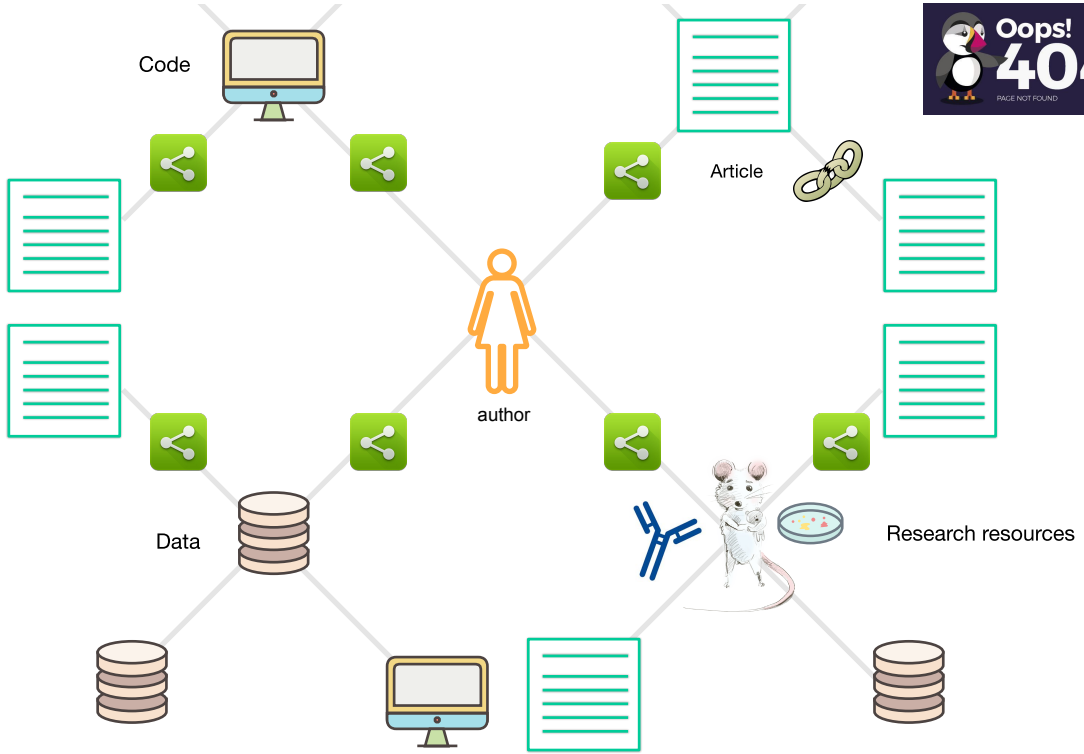
[“Rich metadata” recommendation]

Metadata: Attributes of data, e.g., file type, but also descriptive information about the data

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**

Stable PIDs provide the substrate for the scholarly web



Implementation: Landing page should continue to exist even beyond the lifespan of the data; provide information about the disposition of the data

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

I2: FAIR vocabularies

Anterior Cingulate

Anterior Cingulate Cortex

Anterior Cingulate Area

Anterior Cingulate Gyrus

ACA

ACC

BA24

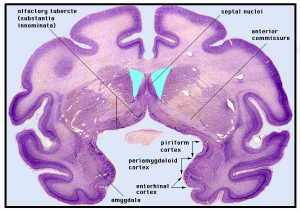
AnteriorCingulateCortex

Cingulate cortex (rostral)

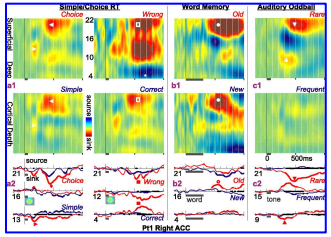
UBERON:0009835

Ensure that ***assertions*** about the meaning of data are unambiguous

Each concept has its own PID and other FAIR attributes



NeuroMaps

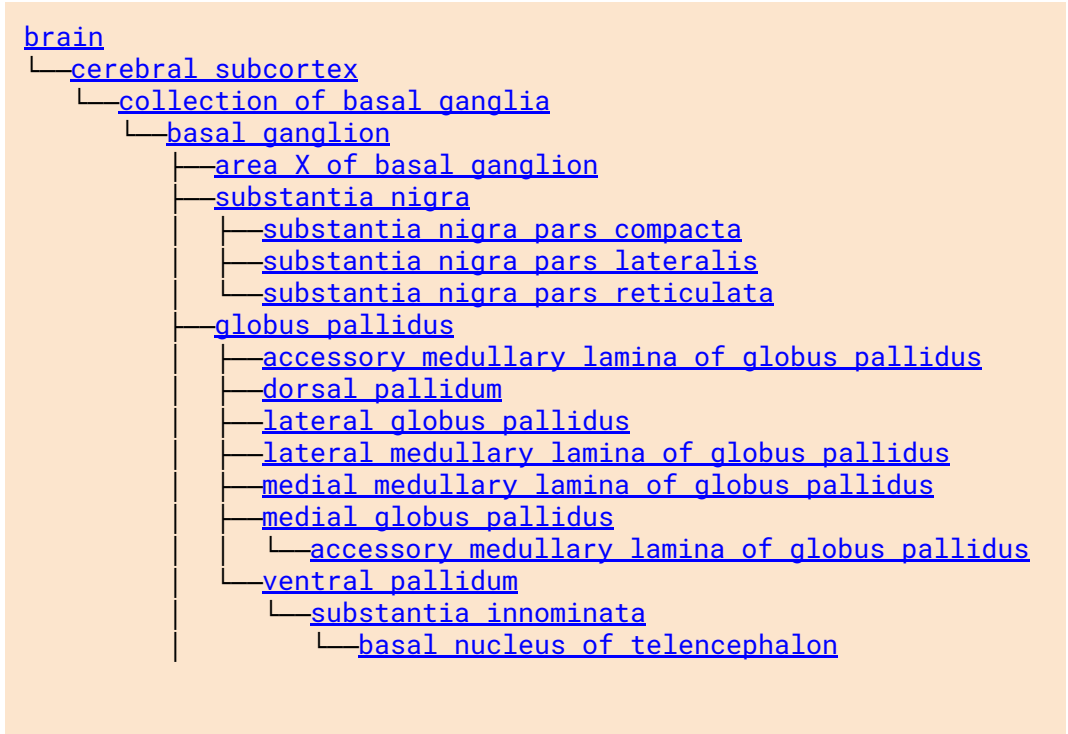


http://www.jneurosci.org/content/jneuro/25/3/604/F5.large.jpg

NIF

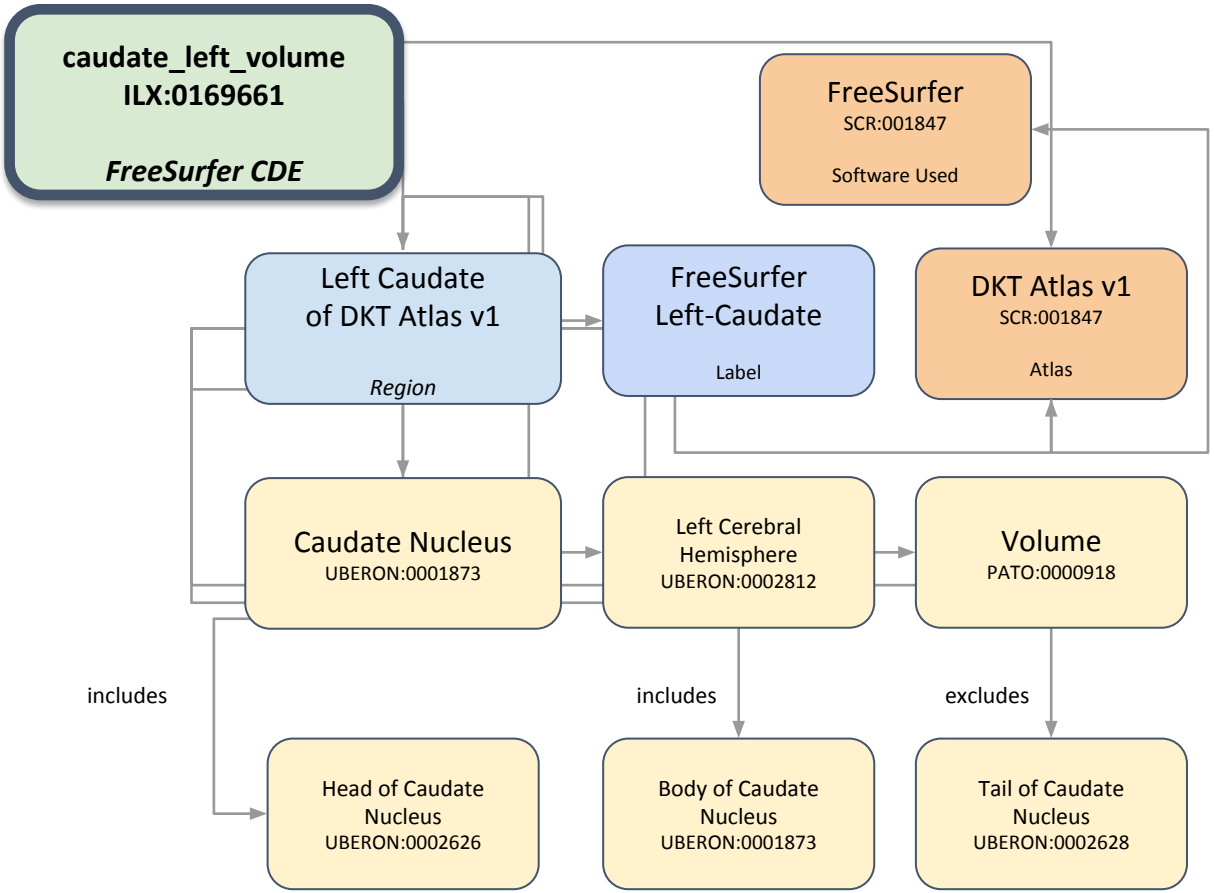
I2: FAIR vocabularies

- Ontologies and controlled vocabularies give us machine-readable ways of managing vocabularies, their relationships and their evolution over time
- Each concept has its own PID and other FAIR attributes
- Provide a richer context through qualified relationships
- Building and developing them for neuroscience has been a challenge
- NIF maintains full services for using, extending and building FAIR vocabularies



http://purl.obolibrary.org/obo/UBERON_0002477

FAIR vocabularies for neuroimaging



Interlex and SciGraph:
Platform for working with and building:

- Ontologies
- Common data elements
- Terminologies
  - Identifiers for concepts, CDE's and tools
  - Provenance
  - Qualified relationships to other data
  - Machine accessible

Jeff Grethe, Smruti Padhy,
Tom Gillespie, Satra Ghosh

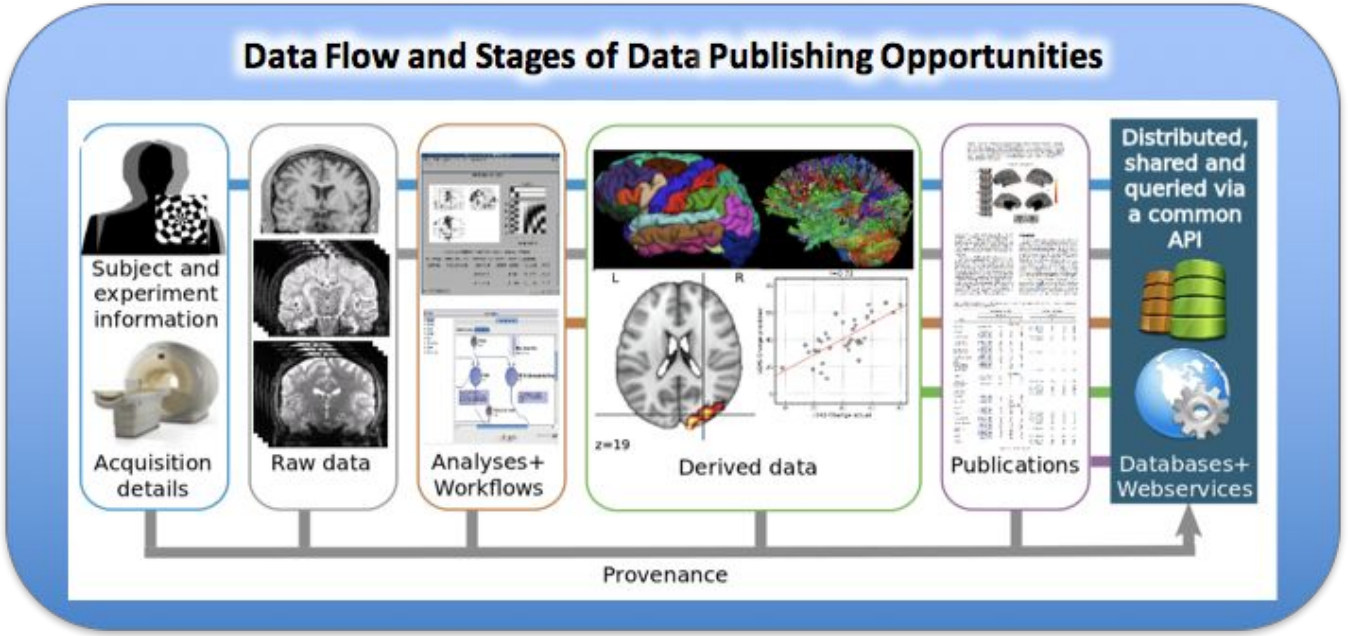
https://scicrunch.org/scicrunch/interlex/dashboar
d

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**

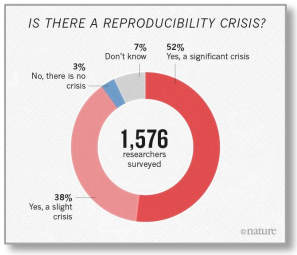
"…legal uncertainty interferes with the productive reuse of research data."-Carroll, 2015

General Neuroimaging Workflow



Metadata Descriptions

ReproNIM: Publishable workflows and tools for reproducible neuroimaging



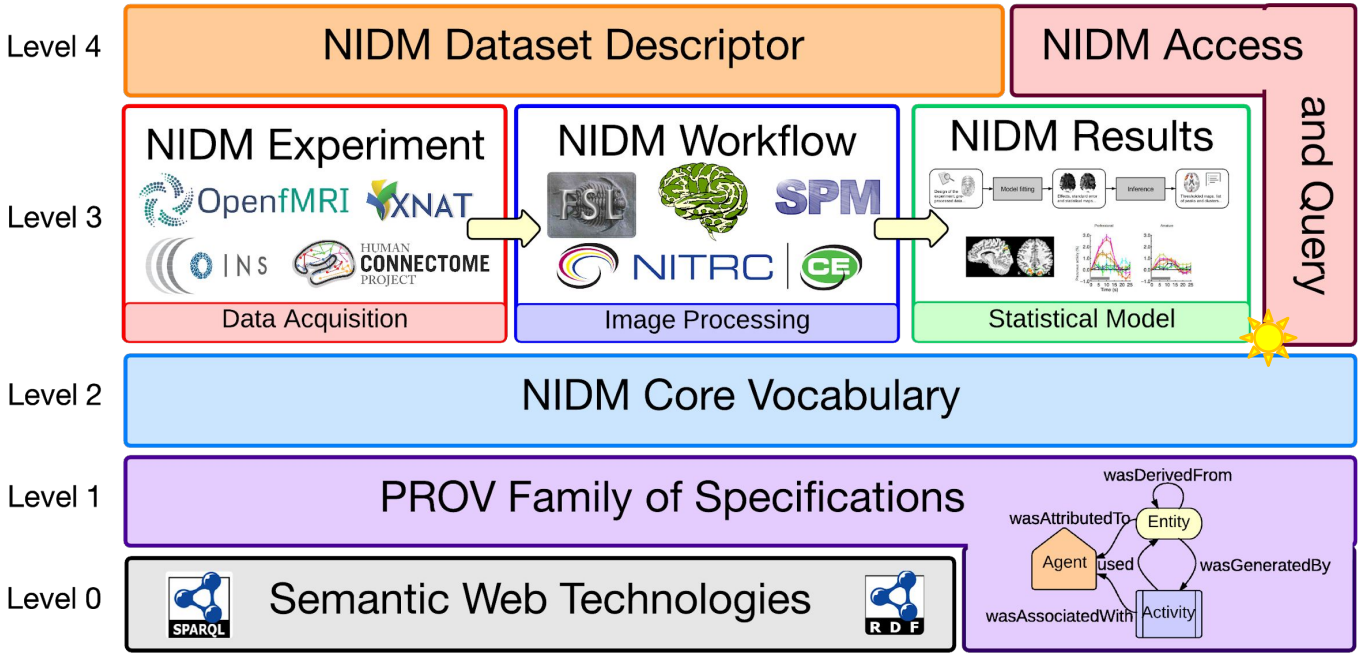
"Everything matters"
Kennedy et al. Front Neuroinform. 2019; 13: 1.doi: 10.3389/fninf.2019.00001

- Computational environments matter (Glatard et al., 2015);
- Tool selection matters (Tustison et al., 2014; Dickie et al., 2017);
- Tool version matters (Dickie et al., 2017);
- Statistical model matters (Tan et al., 2016);
- Study population characteristics matter.

ReproNIM.org

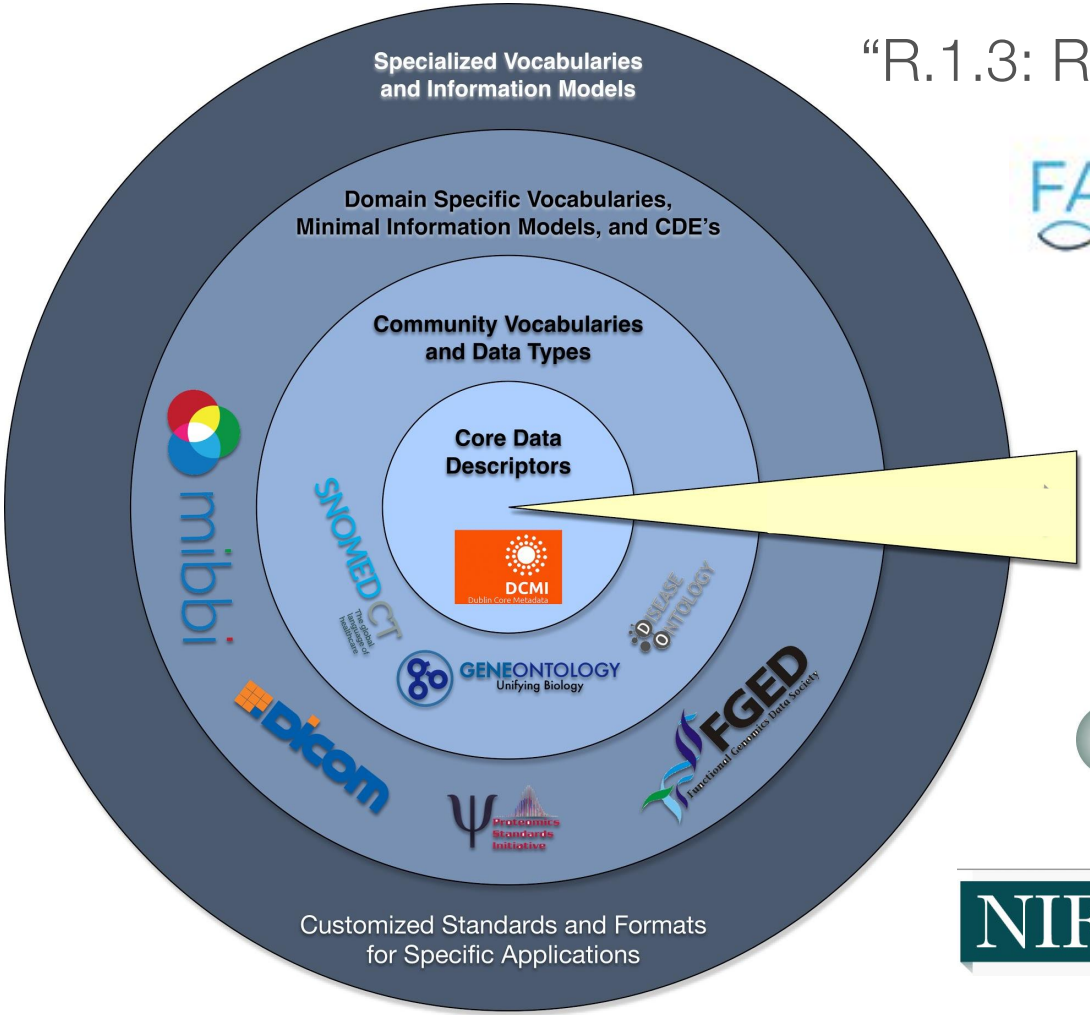
P41 Center led by David Kennedy

Neuroimaging Data Model



http://nidm.nidash.org/

“R.1.3: Relevant community standards”



FAIRsharing.org
standards, databases, policies

FORCE11
The Future of Research Communications and e-Scholarship

RDA
RESEARCH DATA ALLIANCE

incf

BETA

ReproNim

BRAIN IMAGING DATA STRUCTURE

NIF

NEURODATA WITHOUT BORDERS

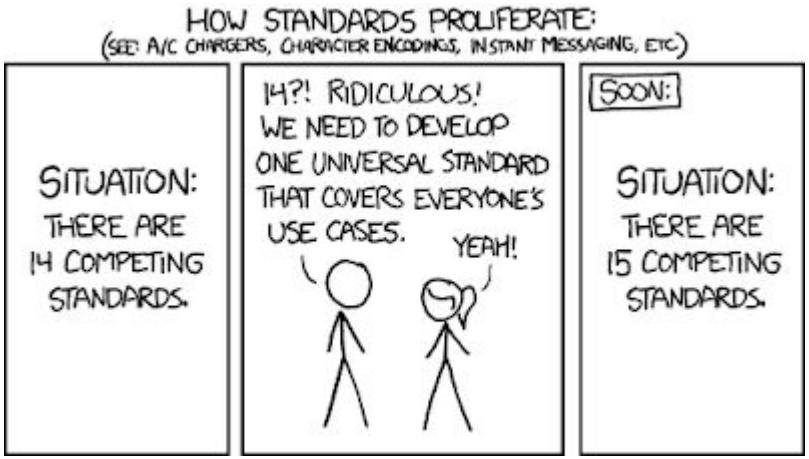
Motivation for use of Standards for Neuroimaging Data

- Imaging experiments are complicated and can be organized in many different ways
- Many files/file types per subject
- Members of the same lab may use different ways to arrange data
- Difficulty sharing data within and across large scale projects

"Think this is bad? You should see the inside of my head."

Slide courtesy of Jeffrey Grethe

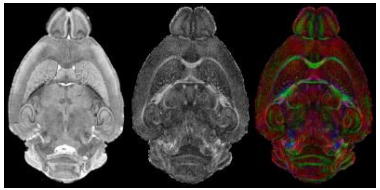
But which “community” standard?



- BIDS
- DICOM
- NIDM
- NIX
- NDA
- ...

A standards organization for neuroscience

- Dizzying number of proposed standards, many with very limited adoption
- International Neuroinformatics Facility (INCF): taking a leading role in coordinating standards and best practices for neuroscience data
- Independent evaluation and endorsement procedure, regardless of where they are developed
- Provides support for working groups who want to work through issues in interoperability



https://www.incf.org/activities/standards-and-best-practices

INCF endorsement criteria

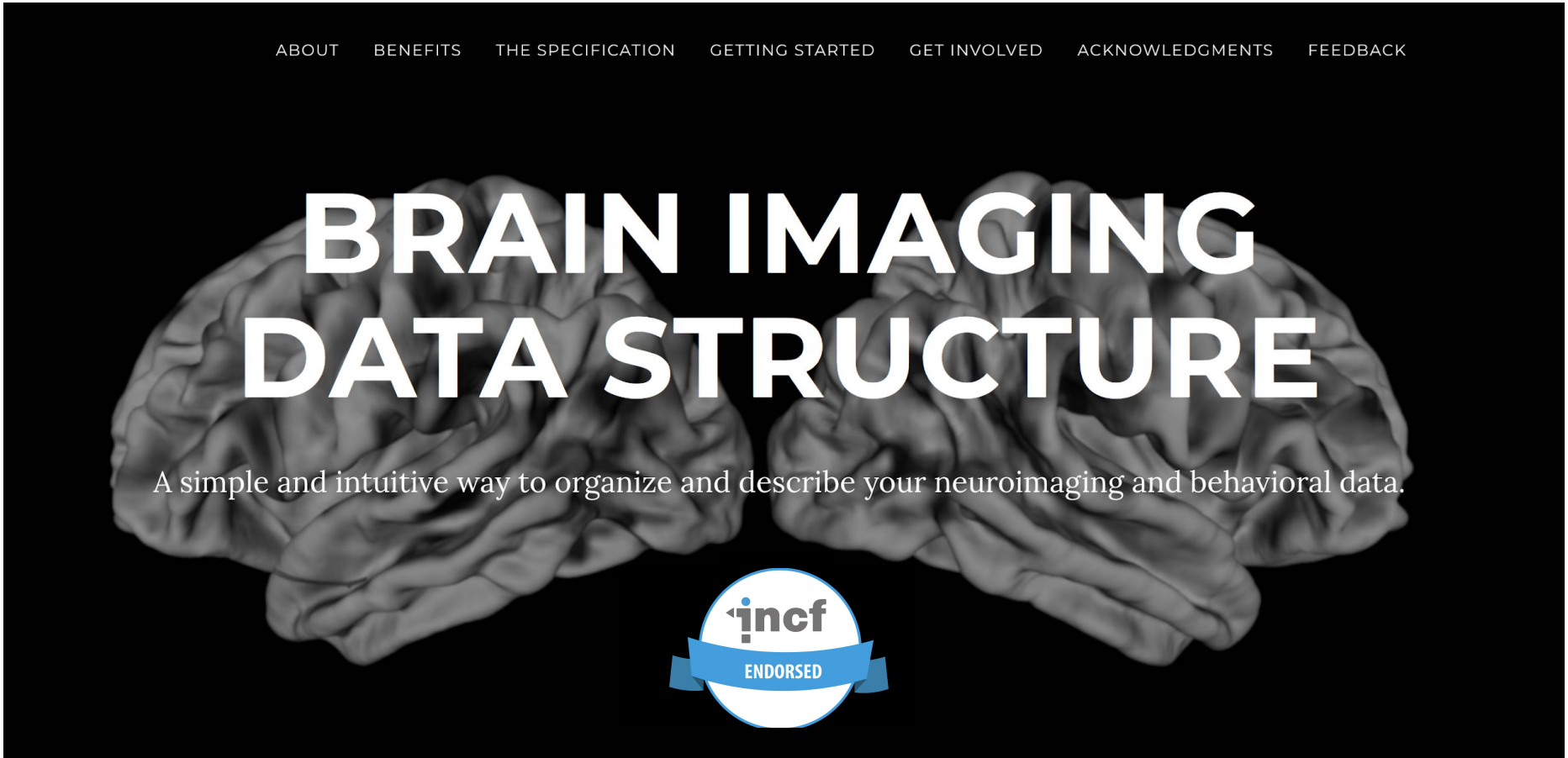
- **Open**
- **FAIR**
- **Citable**: Makes clear how to cite it to track usage and to acknowledge the contributions of those who develop and maintain it
- Has relevant tooling and implementations
- At least some associated tools must be open
- Governance
- Documentation and training materials
- Means for maintenance and long term sustainability
- Clear evidence of adoption and use ***outside*** of those who developed it

Abrams et al., OSF, 2018

Endorsement means...

- SBP will be listed as an INCF endorsed standard on the INCF website
- Author/steward can display the “Endorsed by INCF” badge
- INCF will include SBP in their training materials, courses and workshops
- INCF will promote its use and broad dissemination
- SBP development can be supported by INCF working groups and associated seed funding

BIDS was the first standard to go through the INCF endorsement process

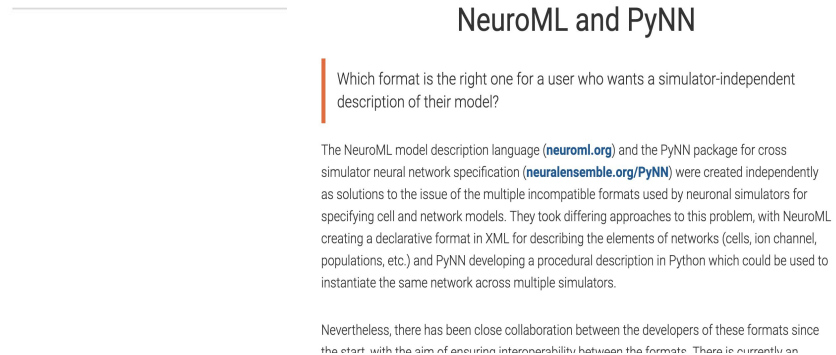


Goal: Harmonious, interoperable well documented and supported standards for neuroscience

- INCF reviews standards and best practices to avoid overlap and provide guidance on the use of complementary standards
- INCF provides substrate for groups to come together to harmonize standards for interoperability

Listed below are SBPs that have been endorsed by INCF:

Brain Imaging Data Structures (BIDS)	Neuroimaging
NeuroML	Computational neuroscience
PyNN	euroscience



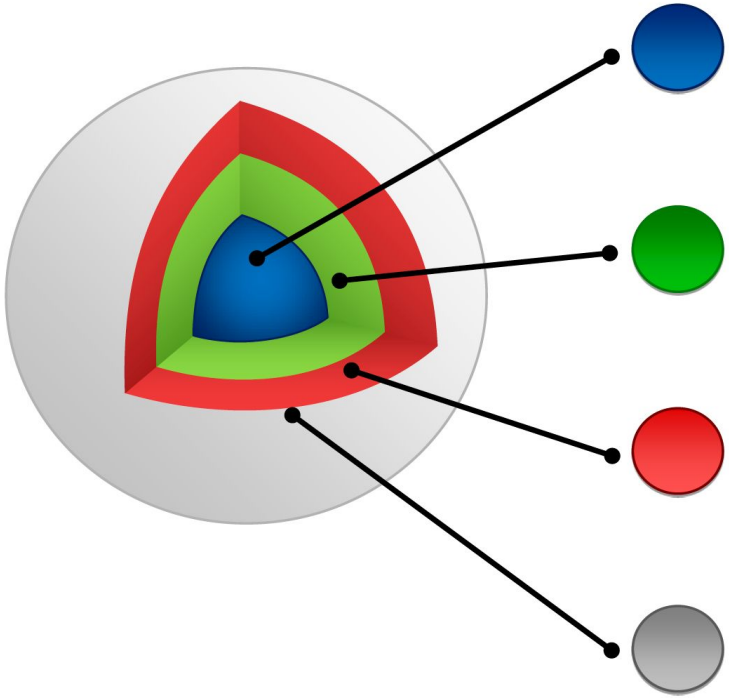
NeuroML and PyNN

Which format is the right one for a user who wants a simulator-independent description of their model?

The NeuroML model description language (**neuroml.org**) and the PyNN package for cross simulator neural network specification (**neuralensemble.org/PyNN**) were created independently as solutions to the issue of the multiple incompatible formats used by neuronal simulators for specifying cell and network models. They took differing approaches to this problem, with NeuroML creating a declarative format in XML for describing the elements of networks (cells, ion channel, populations, etc.) and PyNN developing a procedural description in Python which could be used to instantiate the same network across multiple simulators.

Nevertheless, there has been close collaboration between the developers of these formats since the start, with the aim of ensuring interoperability between the formats. There is currently an

Voila! FAIR data objects



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

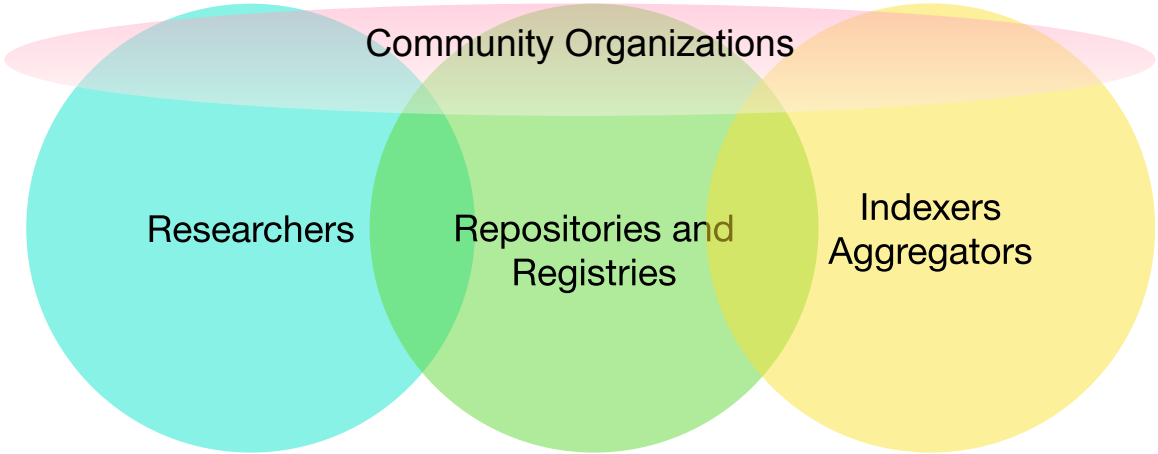
In order for data to be assessable and reusable, it should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the data were created. To enable the broadest 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

OVERWHELMED?

When you're waist-deep in tribbles, it's a bit difficult to remember that your original objective was to guard the quadrotriticale.

FAIR Partnership



- Good data management
- Rich metadata
- Prepare to share
- 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

Data repositories are publishers for data

- Open Neuro: free and open platform for sharing MRI, MEG, EEG, iEEG, and ECoG data
- NITRC-IR: Image repository for neuroimaging data
- NeuroVault: A place where researchers can publicly store and share unthresholded statistical maps, parcellations, and atlases produced by MRI and PET studies.
- NIMH Data Archive: provides infrastructure for sharing research data, tools, methods, and analyses
- Institutional repositories: University libraries taking the lead in providing data storage and data management assistance to researchers

And more via the Neuroscience Information Framework

Best practice: Deposit your data into a trusted repository for management and long term stewardship

Notice of Data Sharing Policy for the BRAIN Initiative

Notice Number: NOT-MH-19-010

You will put your data in a repository

Data archives that have been established include:

1) The Neuroscience Multi-omic Data Archive (https://nemoarchive.org/about.php, R24MH114788) to hold data from -omics experiments.

2) The Brain Image Library (http://www.brainimagelibrary.org/index.html, R24MH114793) to hold microscopy data.

3) Data Archive for the BRAIN Initiative (https://dabi.loni.usc.edu, R24MH114796) to hold data related to human electrophysiology experiments.

4) OpenNeuro (https://openneuro.org/, R24MH117179) to hold magnetic resonance imaging data.

5) Block and Object Storage Service (https://bossdb.org/, R24MH114785) to hold electron microscopy data.

You will use standards

All applications to BRAIN Initiative FOAs are required to include a Resource Sharing Plan as part of the app
21st Century Cures Act and these awards authorized under that Act, t he portion of that plan dealing with da

1) a summary of the data that will be shared

2) a description of the standard(s) that will be used to describe the data set

3) the data archive(s) that will house the data

4) the proposed timeline for submitting data to the archive and sharing data with the research community.

You will share your data

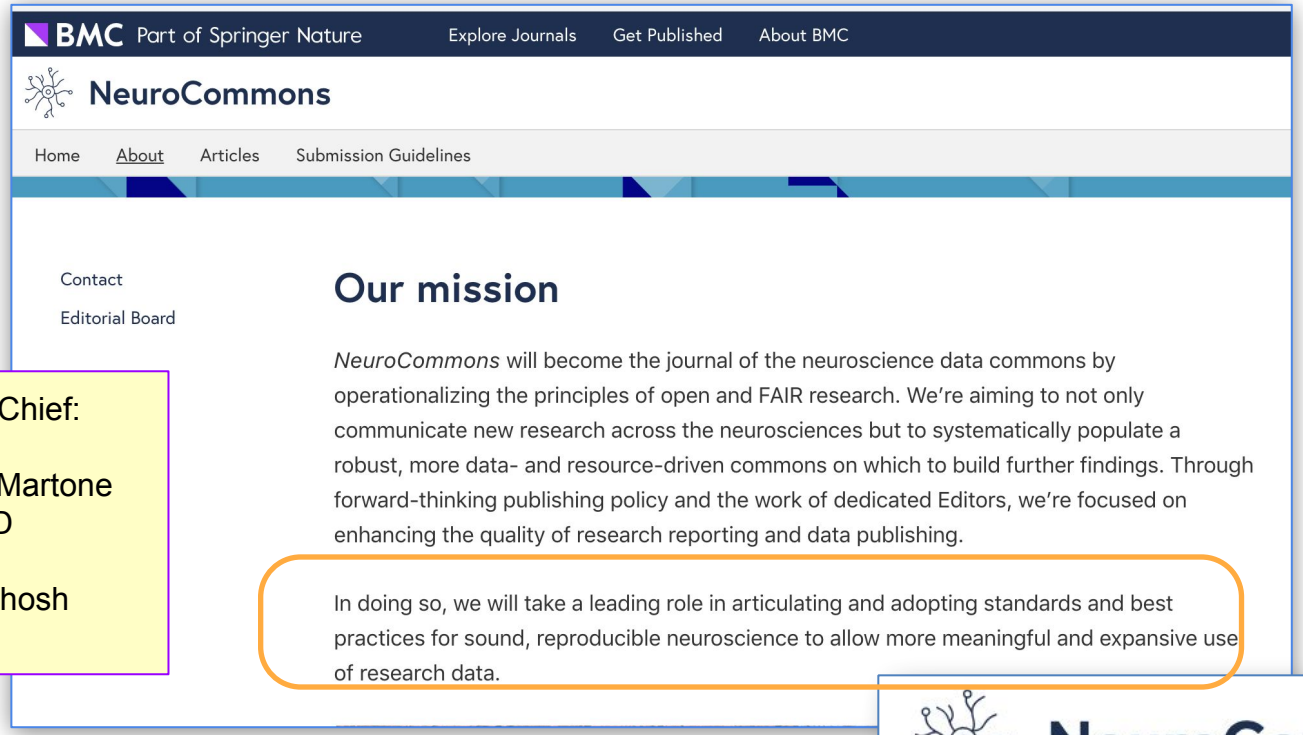
https://grants.nih.gov/grants/guide/notice-files/NOT-MH-19-010.html

FAIR Aspirations

*	**The basic core**: metadata, PID & access	F2. data are described with rich metadata F1. (meta)data are assigned a globally unique and persistent identifier A1. (meta)data are retrievable by their identifier using a standardized communications protocol
**	**Enhanced access**: catalogues for discovery, standard (controlled) access & licences	F4. (meta)data are registered or indexed in a searchable resource A1.1. the protocol is free, open and universally implementable A1.2. the protocol allows for an authentication and authorization procedure, where necessary R1.1. (meta)data are released with a clear and accessible data usage license
***	**Use of standards:** for metadata and data	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation R1.3. (meta)data meet domain relevant community standards F3. metadata clearly and explicitly include the identifier of the data it describes
****	**Rich, FAIR metadata**	R1. (meta)data are richly described with a plurality of accurate and relevant attributes I2. (meta)data uses vocabularies that follow FAIR principles
*****	**Provenance and additional context**	R1.2 (meta)data are associated with data provenance I3. (meta)data include qualified references to other (meta)data A2. metadata are accessible, even when the data are no longer available

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

Introducing NeuroCommons: Journal to support Open, FAIR and Citable Science



Editors in Chief:

Maryann E. Martone
UCSD

Satrajit Ghosh
MIT

NeuroCommons

https://neurocommons.biomedcentral.com/

Summary and Recommendations

- FAIR data principles give us guidance on how to prepare data for maximal utility to us and to others
  - Students may not think of “future you” but PI’s do
- FAIR requires technical and human infrastructure throughout the data lifecycle
  - FAIR begins with the researcher
  - Good practice: Add rich metadata to all research objects
  - Participate in the human infrastructure as well
- The more we practice FAIR, the easier it will get

Useful references

- The FAIR Guiding Principles for scientific data management and stewardship: Wilkinson et al. 2016
- FAIR Principles Explained: https://www.dtls.nl/fair-data/fair-principles-explained/
- ReproNIM FAIR Data Module: http://www.repronim.org/module-FAIR-data/
- GO-FAIR: Developing implementation networks and tools for FAIR: https://www.go-fair.org/
- FAIR Sharing.org: Site that provides information on scientific standards
- INCF Standards and Best Practices for FAIR neuroscience: https://www.incf.org/activities/standards-and-best-practices

Prepare to share FAIR: The importance of good data management

B. Reasons for not sharing data	
Reason	**Percent**
Data contains additional findings to publish	50.43
Data contains sensitive information	30.43
It would take too much time	25.22
Supervisor doesn't wish to share	16.52
Format of data make sharing difficult	15.62
Don't know how	14.78
Data is proprietary, subject to IP	1.74
Other	7.82
Can share, but require citation	29.57
Can share, but require authorship	9.57
	$N = 115$

Borghi and Van Gulick (2018) Data management and sharing in neuroimaging: Practices and perceptions of MRI researchers

Table 5: Reasons why data can and cannot be shared.

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 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 I implement consistently, 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 consistent 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