FAIR Data and You: Are you ready for 2023?

Learn how to succeed with your next NIH grant: New Data Mandates and Opportunities

Jeffrey S. Grethe

PI, NIDDK Information Network

Co-Director, FAIR Data Informatics Laboratory, UCSD

Topics for Today's Webinar

1. Open science and data sharing
2. Data as a research output
3. What are the FAIR principles?
4. Why were the FAIR data principles introduced?
5. A deeper dive into some of the FAIR principles
    a. Use of persistent identifiers
    b. Rich metadata
    c. Data licenses
    d. FAIR vocabularies
    e. Community standards
6. Introduction to data repositories
7. FAIR data and you: Importance of Data Management

About dkNET

- Research resource information portal for biomedical researchers
- Information network to connect DK researchers and NIH-funded resources and centers
- Funded by National Institute of Health (NIH) - National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)
- Developed and maintained by the FAIR Data Informatics Laboratory (fdilab.org) at UCSD (Supports major informatics projects in biomedicine)

FDI Lab
FAIR Data Informatics

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Get involved in the dkNET Community

dkNET Homepage: dkNET.org

Blog, Calendar

Check out dkNET Resour

Join Webinar

Sign up email list

Check Out or Post News and Funding Opportunities

Follow us @dknet_info

Open science

...the movement to make scientific research (including **publications, data, physical samples, and software**) and its dissemination accessible to all levels of an inquiring society, amateur or professional. Open science is transparent and accessible knowledge that is shared and developed through collaborative networks. It encompasses practices such as publishing open research, campaigning for open access, encouraging scientists to practice open notebook science, and generally making it easier to publish and communicate scientific knowledge. -adapted from Wikipedia

And then there's COVID...

OECD Policy Responses to Coronavirus (COVID-19)

Why open science is critical to combatting COVID-19

Updated 12 May 2020

Key messages

- In global emergencies like the coronavirus (COVID-19) pandemic, open science policies can remove obstacles to the free flow of research data and ideas, and thus accelerate the pace of research critical to combating the disease.
- ...aboration of research data has reached unprecedented levels, ...t least some of the data is relatively low, and outstanding ...cific standards, co-ordination and interoperability, as well as ...n.
- To strengthen the contribution of open science to the COVID-19 response, policy makers ...te data governance models, interoperabl... ...ts involving public sector, private sector ...able infrastructures, human and institu... ... to data across borders.

Perspective

The NEW ENGLAND JOURNAL of MEDICINE

Leveraging Open Science to Accelerate Research

Kushal T. Kadakia, M.Sc., Adam L. Beckman, B.S., Joseph S. Ross, M.D., M.H.S., and Harlan M. Krumholz, M.D.

#News #Coronavirus

Open-Access Publishing and the Coronavirus

Some publishers are moving toward an open model, prompted by COVID-19.

By Jack Grove for Times Higher Education // May 15, 2020 2 COMMENTS

The unrestricted sharing of scientific papers during the coronavirus pandemic may have hastened the shift toward more open-access publishing, scientists believe, as several leading journals move to make content publicly available.

Last month, Britain's Biochemical Society became the latest organization to make all of its published content free to view, citing the "extraordinary times with the current Covid-19 pandemic" as its reason for lifting paywalls on its Portland Press imprint until further notice.

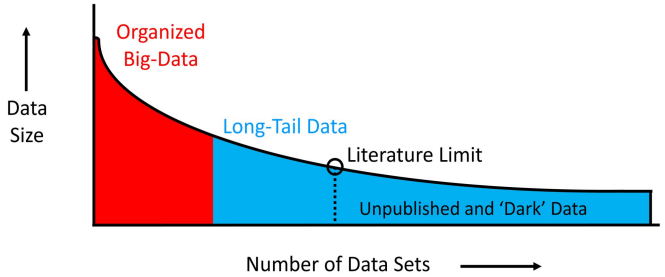
It follows the decision by Springer Nature, announced on April 8, to offer researchers a route to publishing open access in *Nature* and most *Nature*-

Motivations for open science

- Consequence of digital revolution: access and computation
- Two main arguments in biomedicine:
  - Research is largely funded by the public and the public should have access to the results
  - Open science = better science
    - Transparency
    - More eyes
    - Increasing computational capacity that can mine biomedical literature and data for new insights

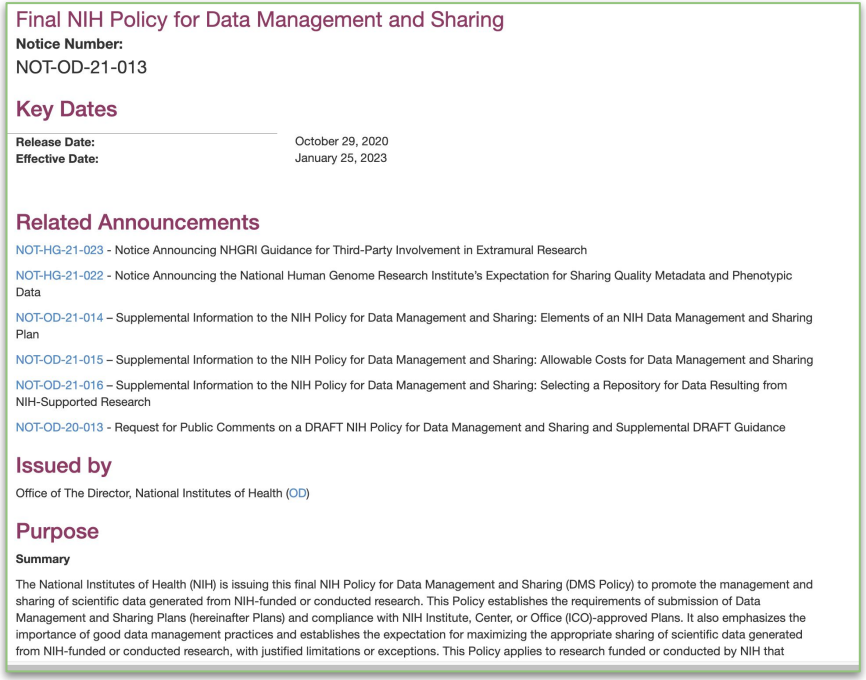
Trends in Open Data

- “Long tail data”
  - Make data underlying claims in a paper available
  - Publish data as a primary product of research, like you would a paper
    - Transparency
    - Reuse
    - Aggregation to increase analytic power
- “Big science” : generate large pools of freely available and more usable data “AI/ML Ready”
  - Spur data science in biomedical science
- Funders and journals are increasingly requiring data sharing



January 25, 2023

- US National Institutes of Health new data sharing policy goes into effect
- All data must be managed; most data should be shared
- “As open as possible; as closed as necessary”
- Mandates the ***inclusion, approval and execution*** of a Data Management and Sharing Plan
  - (DMP + S = DMS)



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

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

The Details...

The effective date of the DMS Policy is January 25, 2023, including for:

- Competing grant applications that are submitted to NIH for the January 25, 2023 and subsequent receipt dates;
- Proposals for contracts that are submitted to NIH on or after January 25, 2023;
- NIH Intramural Research Projects conducted on or after January 25, 2023; and
- Other funding agreements (e.g., Other Transactions) that are executed on or after January 25, 2023, unless otherwise stipulated by NIH.

More Details...

What

- Defines Scientific Data as: “The recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data **do not include** laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens.”
- Even those scientific data not used to support a publication are considered scientific data and within the final DMS Policy’s scope

When

- “[s]hared scientific data should be made accessible as soon as possible, and no later than the time of an associated publication, or the end of the award/support period, whichever comes first.”
- Researchers may share data underlying publication during the period of award but may share other data that have not yet led to a publication by the end of the award period.

Where

- Encourages the use of established repositories to the extent possible.

More Details...

How

- NIH encourages data management and data sharing practices consistent with the FAIR data principles

Funding

- Fees for long-term data preservation and sharing are allowable, but funds for these activities must be spent during the performance period, even for scientific data and metadata preserved and shared beyond the award period.

Repercussions

- After the end of the funding period, non-compliance with the NIH ICO-approved Plan may be taken into account by NIH for future funding decisions for the recipient institution

The DMS Policy applies to all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data. This includes research funded or conducted by extramural grants, contracts, Intramural Research Projects, or other funding agreements regardless of NIH funding level or funding mechanism.

Good data management is the gateway to data sharing

	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

Changing the culture around data management and sharing

- **Science and Society**
  - Transparency
  - Reproducibility
  - Reduced waste
  - ***Driving discovery***
- **Me**
  - Answer to the underpowered study
  - Data sharing and good data management are closely aligned
  - Compliance with mandates
  - Credit for the totality of my work
- **Future me**
  - One most likely to benefit from good data management and sharing through stable archives
  - No one ever regretted annotating too much
- **My colleagues (and PI)**
  - Easy to engage with colleagues over well annotated data and associated code
  - What happens when the post doc leaves?

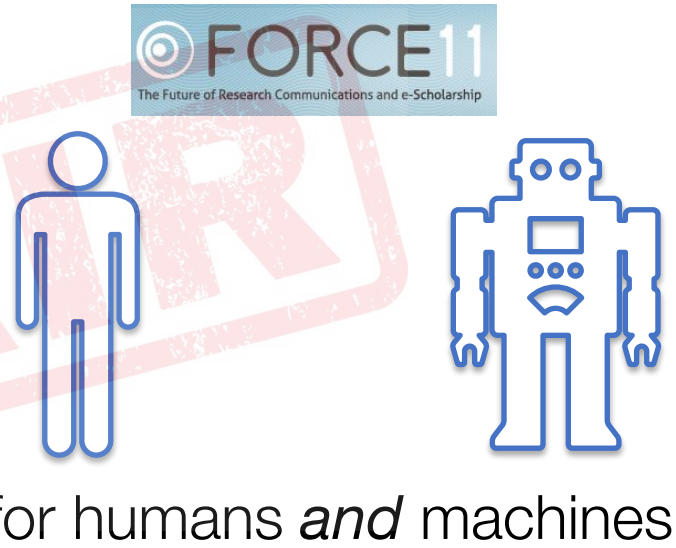
But how do I do that?

- “NIH encourages data management and data sharing practices consistent with the FAIR data principles.”
- “NIH strongly encourages the use of established repositories to the extent possible for preserving and sharing scientific data”

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

Definition: Machine readable

- "(meta)data in a format that can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost." ...

  There are two types of machine-readable data: human-readable data that is marked up so that it can also be read by machines (e.g. microformats, RDFa, HTML) and data file formats intended principally for processing by machines (CSV, RDF, XML, JSON).

  Again, these formats are only machine readable if the data contained within them is formally structured; exporting a CSV file from a badly structured spreadsheet does not make the data machine-readable. -adapted from Wikipedia

CSV file...or is it?

SubjectID	Subject Group	Species	Age	Unit	Sex	Strain
F001	1	Mouse	25	days	F	C57Bl/6
F002	1	Mouse	23	days	M	C57Bl/6
F003	1	Mouse	26	days	M	C57Bl/6
F004	1	Mouse	25	days	M	C57Bl/6
F005	1	Mouse	24	days	F	C57Bl/6

SubjectID	Subject Group	Species	Age	Unit	Sex	Strain
F001	1	Mouse	25	days	F	C57Bl/6
F002	1	Mouse	23	days	M	C57Bl/6
F003	1	Mouse	26	days	M	C57Bl/6
F004	1	Mouse	25	days	M	C57Bl/6
F005	1	Mouse	24	days	F	C57Bl/6

Good resource: https://schoolofdata.org/courses/

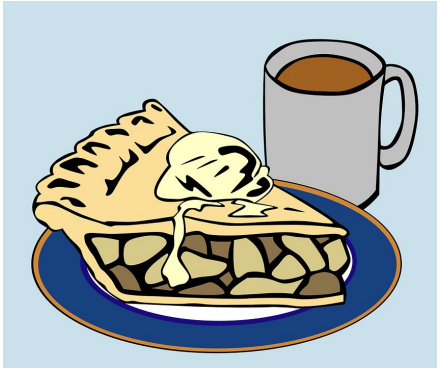
Metadata

- “Data about about data” e.g., file format, file size
- ***Descriptive metadata*** describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, keywords, study organism, condition studied etc

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

MOM

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

Findable

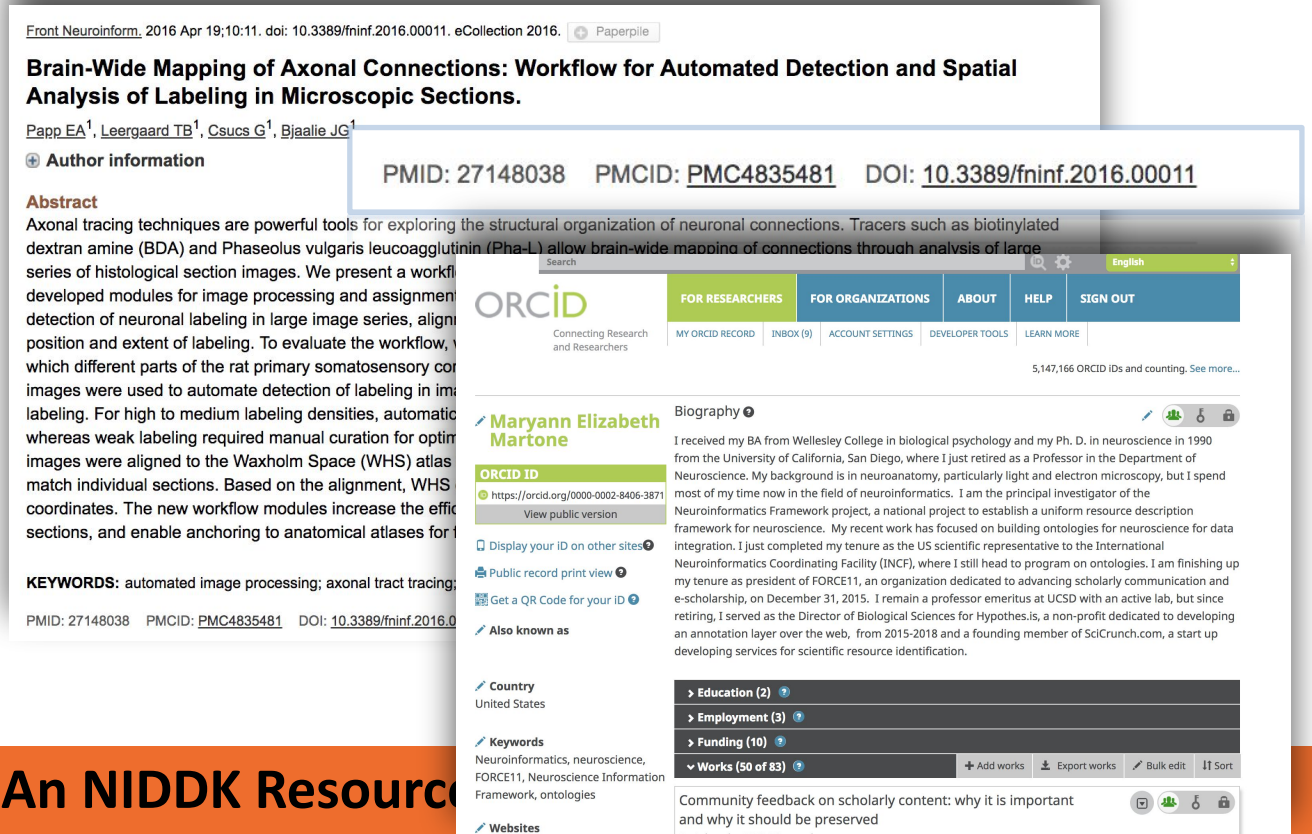
- **F1. (meta)data are assigned a *globally unique and persistent* identifier**
- **F2. data are described with rich metadata**
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

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

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

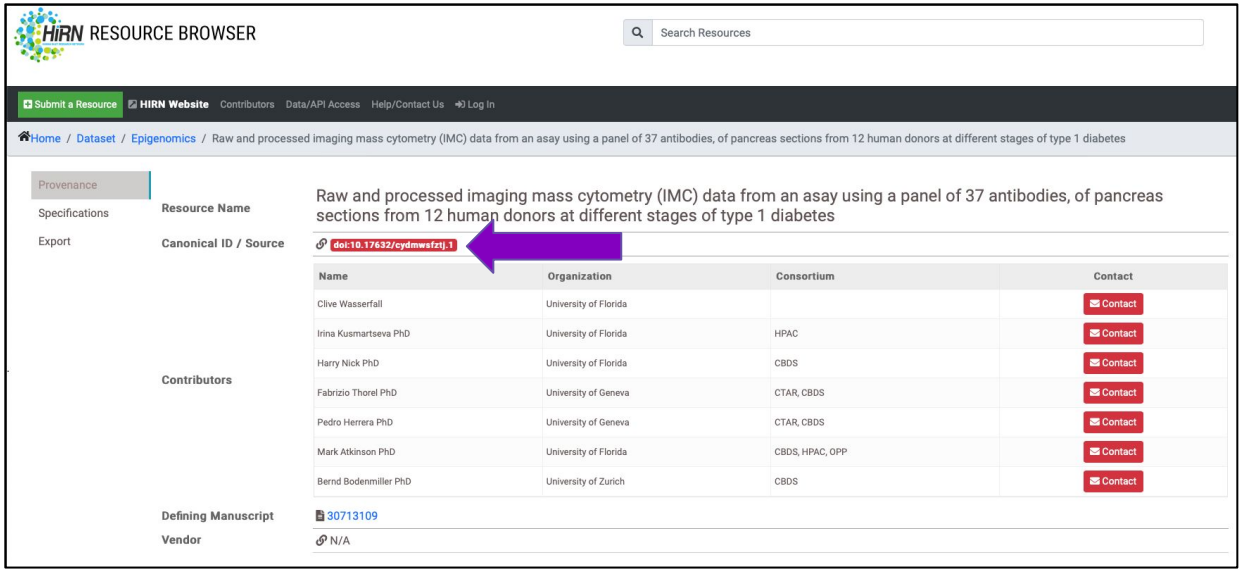
Some examples of PIDs



- 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**

PIDS in action: DOI

- Identifies an object regardless of its location
- Issued by authorities: DataCite and CrossRef who work with data repositories to make sure that the links don't break
- Issued by data repositories who agree to keep the links up to date
- A PID is a social contract!



DOI: 10.17632/cydmwsfztj.1

https://doi.org/10.17632/cydmwsfztj.1

Lesson: You cannot assign PIDs to your data yourself

You obtain a PID from a trusted repository that provides PIDs

https://dknet.org/about/Suggested-data-repositories-niddk

Suggested data repositories

Home / Suggested data repositories

Where can I deposit my data?

We have organized a list of data repositories that are recommended by the following sources: NIDDK domain experts, Nature Scientific Data, PLOS One, NLM NIH-Science. It is generally best practice to deposit data into a discipline-specific and community recognized repository if one is available, or into an institutional or gene specialist repository is available.

- NIDDK-specific repositories
- NIH-supported repositories
- Institutional repository
- Other NIDDK Project-specific or consortium-specific data or sample repositories

By scientific disciplines

- All research data types
- Chemistry and chemical biology and biochemistry
- Cytometry and Immunology
- Functional Genomics
- Imaging
- Metabolomics
- Molecular and supramolecular structure
- Neuroscience
- Nucleic acid sequence
- Other domain-specific repositories
- Protein sequence
- Proteomics

De-identified human clinical research data

Clinical trial data is encouraged to be submitted to the ClinicalTrials.gov even if it is not required. For studies include human genomic and associate phenotyp database of Genotypes and Phenotypes (dbGaP). Another repository that you can consider is ICPSR, which hosts a variety of human data, including many d science studies. Information on uploading data to ICPSR can be found here. Before uploading data, please note that the data should be de-identified, and you institutional IRB's requirements and receive approvals. For completed phase I-IV interventional studies, you can also share anonymized data at Vivli.

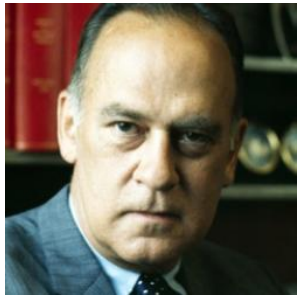
NIH-supported repositories (for complete and current list of NIH repositories click here)

Repository Name	RRID	Description	Type of Data	Recommended By
Accelerating Medicines Partnership Type 2 Diabetes Knowledge Portal (AMP-T2D)	RRID:SCR_003743 FAIR	Portal and database of DNA sequence, functional and epigenomic information, and clinical data from studies on type 2 diabetes and analytic tools to analyze these data. .Provides data and tool ...[more]	Array, exome sequencing, whole genome sequencing data	NLM, NIDDK
Analysis, Visualization, and Informatics Lab-space (AnVIL)	RRID:SCR_017469	Portal to facilitate integration and computing on and across large datasets generated by NHGRI programs, as well as initiatives funded by National Institutes of Health or by other agencies th ...[more]	Genomic data	NLM, NIDDK
Biological General Repository for Interaction Datasets (BioGRID)	RRID:SCR_007393 FAIR	Curated protein-protein and genetic interaction repository of raw protein and genetic interactions from major model organism species, with data compiled through comprehensive curation efforts.	Molecular interaction data	NLM, NIDDK, PLoS ONE, Sci Data
Cancer Imaging Archive (TCIA)	RRID:SCR_008927 FAIR	Archive of medical images of cancer accessible for public download. All images are stored in DICOM file format and organized as Collections, typically patients related by common disease (e.g. ...[more]	Primary DICOM image datasets from cancer patients and analysis datasets	NLM, NIDDK, PLoS ONE, Sci Data
Cancer Nanotechnology Laboratory (caNanoLab)	RRID:SCR_013717 FAIR	Data sharing portal designed to facilitate information sharing across international biomedical nanotechnology research community to expedite and validate use of nanotechnology in biomedicine.	Physico-chemical, in vitro and in vivo assays data that characterize nanomaterials. "This is a curated resource which may not accept direct submission of data. Contact the database directly f ...[more]	NLM, PLoS ONE, Sci Data
Cell Image Library (CIL)	RRID:SCR_003510 FAIR	Freely accessible, public repository of vetted and annotated microscopic images, videos, and animations of cells from a variety of organisms, showcasing cell architecture, intracellular funct ...[more]	Microscopic imaging data	NLM, NIDDK
ClinicalTrials.gov	RRID:SCR_002309 FAIR	Registry and results database of federally and privately supported clinical trials conducted in United States and around world. Provides information about purpose of trial, who may participat ...[more]	Clinical trial	NLM, NIDDK, Sci Data
DNA DataBank of Japan (DDBJ)	RRID:SCR_002359 FAIR	Maintains and provides archival, retrieval and analytical resources for biological information. Central DDBJ resource consists of public, open-access nucleotide sequence databases including r ...[more]	Gene sequence	NLM, NIDDK, Science, PLoS ONE, Sci Data
Database of Interacting Proteins (DIP)	RRID:SCR_003167	Database to catalog experimentally determined interactions between proteins combining information from a variety of sources to create a single, consistent	Protein interaction data	NLM, NIDDK, PLoS ONE, Sci Data

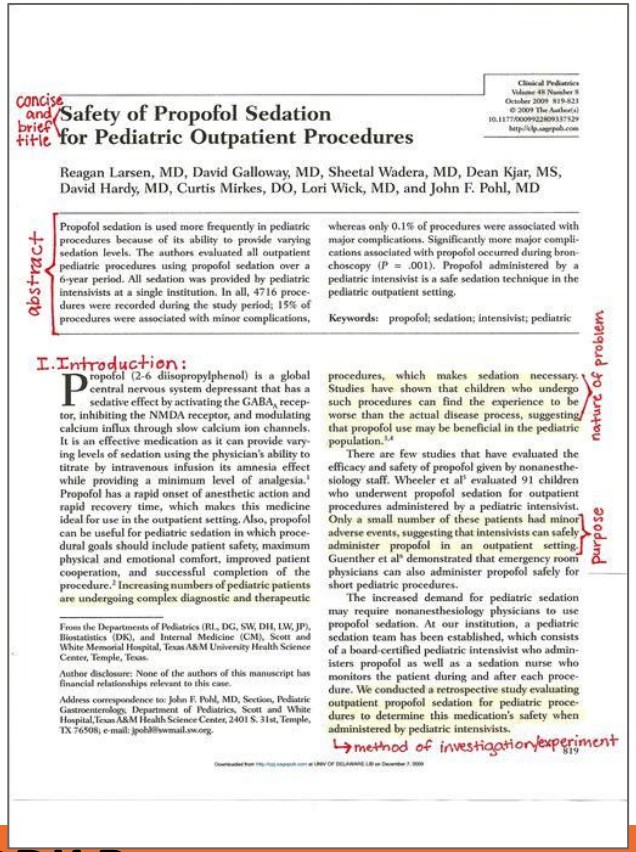
An NIDDK Resource

F2: Data are described with rich metadata

- What exactly are rich metadata?
- “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



We know how to publish papers so they can be found by search engines, understood and reused by people we don't know



concise and brief title

Clinical Pediatrics
Volume 48 Number 8
October 2009 819-823
© 2009 The Author(s)
10.1177/0009922809337529
http://clp.sagepub.com

Safety of Propofol Sedation for Pediatric Outpatient Procedures

Reagan Larsen, MD, David Galloway, MD, Sheetal Wadera, MD, Dean Kjar, MS, David Hardy, MD, Curtis Mirkes, DO, Lori Wick, MD, and John F. Pohl, MD

abstract

Propofol sedation is used more frequently in pediatric procedures because of its ability to provide varying sedation levels. The authors evaluated all outpatient pediatric procedures using propofol sedation over a 6-year period. All sedation was provided by pediatric intensivists at a single institution. In all, 4716 procedures were recorded during the study period; 15% of procedures were associated with minor complications, whereas only 0.1% of procedures were associated with major complications. Significantly more major complications associated with propofol occurred during bronchoscopy ($P = .001$). Propofol administered by a pediatric intensivist is a safe sedation technique in the pediatric outpatient setting.

Keywords: propofol; sedation; intensivist; pediatric

I. Introduction:

nature of problem

Propofol (2-6 diisopropylphenol) is a global central nervous system depressant that has a sedative effect by activating the $GABA_A$ receptor, inhibiting the NMDA receptor, and modulating calcium influx through slow calcium ion channels. It is an effective medication as it can provide varying levels of sedation using the physician's ability to titrate by intravenous infusion its amnesia effect while providing a minimum level of analgesia.[1] Propofol has a rapid onset of anesthetic action and rapid recovery time, which makes this medicine ideal for use in the outpatient setting. Also, propofol can be useful for pediatric sedation in which procedural goals should include patient safety, maximum physical and emotional comfort, improved patient cooperation, and successful completion of the procedure.[2] Increasing numbers of pediatric patients are undergoing complex diagnostic and therapeutic procedures, which makes sedation necessary.[3] Studies have shown that children who undergo such procedures can find the experience to be worse than the actual disease process, suggesting that propofol use may be beneficial in the pediatric population.[3,4]

purpose

There are few studies that have evaluated the efficacy and safety of propofol given by nonanesthesiology staff. Wheeler et al[5] evaluated 91 children who underwent propofol sedation for outpatient procedures administered by a pediatric intensivist. Only a small number of these patients had minor adverse events, suggesting that intensivists can safely administer propofol in an outpatient setting. Guenther et al[6] demonstrated that emergency room physicians can also administer propofol safely for short pediatric procedures.

The increased demand for pediatric sedation may require nonanesthesiology physicians to use propofol sedation. At our institution, a pediatric sedation team has been established, which consists of a board-certified pediatric intensivist who administers propofol as well as a sedation nurse who monitors the patient during and after each procedure. We conducted a retrospective study evaluating outpatient propofol sedation for pediatric procedures to determine this medication's safety when administered by pediatric intensivists.

↳ method of investigation/experiment

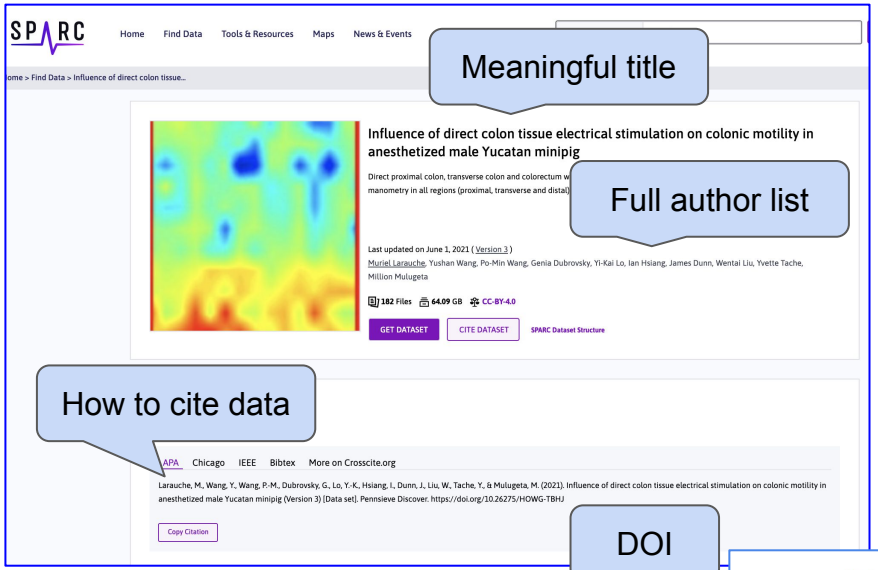
From the Departments of Pediatrics (RL, DG, SW, DH, LW, JP), Biostatistics (DK), and Internal Medicine (CM), Scott and White Memorial Hospital, Texas A&M University Health Science Center, Temple, Texas.

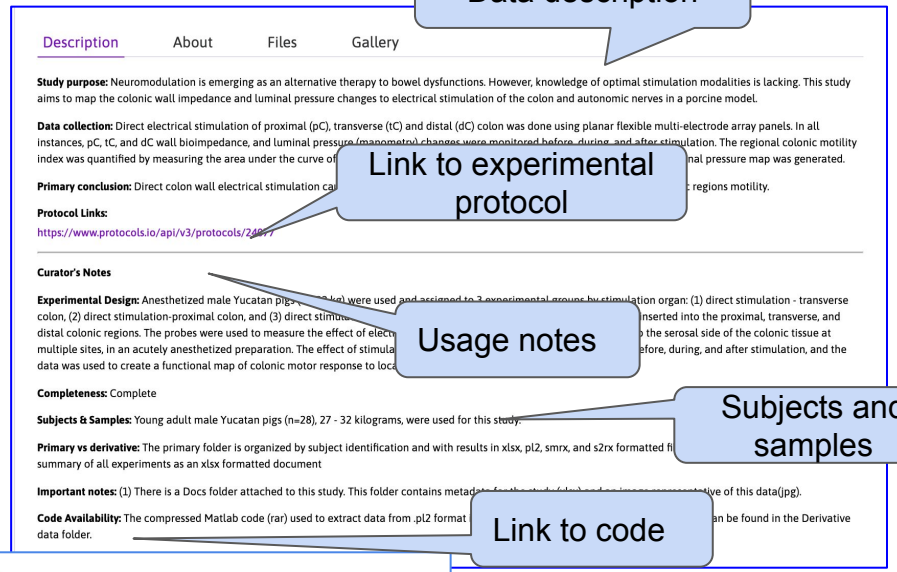
Author disclosure: None of the authors of this manuscript has financial relationships relevant to this case.

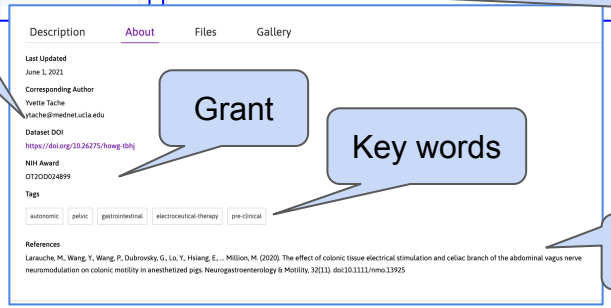
Address correspondence to: John F. Pohl, MD, Section, Pediatric Gastroenterology, Department of Pediatrics, Scott and White Hospital, Texas A&M Health Science Center, 2401 S. 31st, Temple, TX 76508; e-mail: jpohl@swmail.sw.org.

- Clear title
- Standard metadata:
  - Authors
  - Affiliations
  - Contact author
  - Journal
  - Issue
  - Pages (if appropriate)
  - DOI/Pub Med ID
- Standard structure
  - Abstract
  - Introduction
  - Methods
  - Results
  - Discussion
  - Figures/tables
  - References
  - Acknowledgements

PIDs + Rich Meta = FAIR[1]







Rich metadata aids in **search**, provides **provenance**, **credit** for those who produced it and valuable **context** for understanding the data

https://sparc.science/

Lesson: More metadata is better than less metadata

Think about publishing data like you would publish a paper; what information do people need to understand it and use it?

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

“As open as possible, as closed as necessary”

Structured vs unstructured (meta)data

- Structured (meta)data = Data organized according to a data model=an abstract model that organizes elements of data and standardizes how they relate to one another and to properties of the real world entities.

- Unstructured (meta)data = information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. This results in irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in fielded form in databases or annotated (semantically tagged) in documents. (Wikipedia)

Example

- Unstructured: “subjects comprised male (N=6) and female (N =6) C57BL/6 mice, aged 25 days”

- Structured:
  - Subject Group: 1
  - Organism: mouse
  - Age: 25
  - Age unit: days
  - Strain: C57BL/6
  - Sex: Male
  - Number: 6
  - Subject Group: 2
  - Organism: mouse
  - Age: 25
  - Age unit: days
  - Strain: C57BL/6
  - Sex: Female
  - Number: 6

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

Typical Research Scenario

Three different laboratories study a rat model of Type 1 diabetes

- Data set 1: Physiological measurements correlated with gene expression in kidney
- Data set 2: Gene expression in peripheral organs
- Data set 3: Gene expression in major organs

One of these things is not like the other. Maybe.

	Data set 1		Data set 2		Data set 3
Age	15-30 days	**Age range**	10-20 days	**Age**	adult
Organism	WF.BBDR-(*D4Arb 29-D4Rat44*)/Wor	**Subject**	N8 WF.*iddm4*	**Animal**	6
Gene name	Calretinin	**Gene**	Calb2	**Exp**	CR
Location	Ct Kidney	**Region**	Renal cortex	**Sample**	Cortex

- What are the problems here if you wanted to combine these data?
- What would a computer not "know"
- What could have been done differently?

FAIR vocabularies

- Have the characteristics of FAIR data:
  - PIDs
  - Metadata
  - Relationships
- Cortex of kidney = UBERON:0001225
- Can be read by both humans and machines

PIDs are for more than dataset IDs an articles

- Best practice: Use common identifiers to unambiguously identify the same entity across multiple data sets
  - Reagents and tools: RRIDs
  - Genes: Gene IDs
  - People: ORCIDs
  - Articles: DOIs, PMIDs
  - Concepts: Ontologies/controlled vocabularies

PID Power!

Attribute	Data set 1	Data set 2	Data set 3
Age	15-30 days	10-20 days	adult
Organism OBI_0100026	WF.BBDR-(*D4Arb29-D4Rat44*)/Wor RRID:RGD_1357172	N8 WF.*iddm4* RRID:RGD_1357172	6 RRID:RGD_1357172
Gene SO_0000704	Calretinin Gene ID: 117059	Calb2 Gene ID: 117059	CR Gene ID: 117059
Anatomical structure UBERON:0000061	Ctx kidney UBERON:0001225	Renal cortex UBERON:0001225	Cortex UBERON:0000956

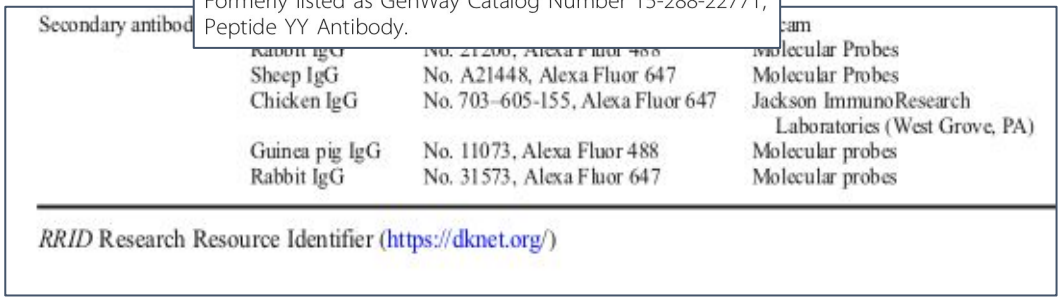
What are RRIDS?

- A persistent identifier for research resources: antibodies, digital tools, cell lines, organisms, plasmids, biosamples
- Required by many journals: Supplied by authors to identify resources in the materials and methods section
- Designed to answer two simple questions:
  - What resources were used in a study
  - Who else has published with this resource?

RRID:AB_1855196

Anti-PYY (ab1)
antibody produced in chicken,
affinity isolated antibody

Catalog Number GW22771
Formerly listed as GenWay Catalog Number 15-288-22771,
Peptide YY Antibody.

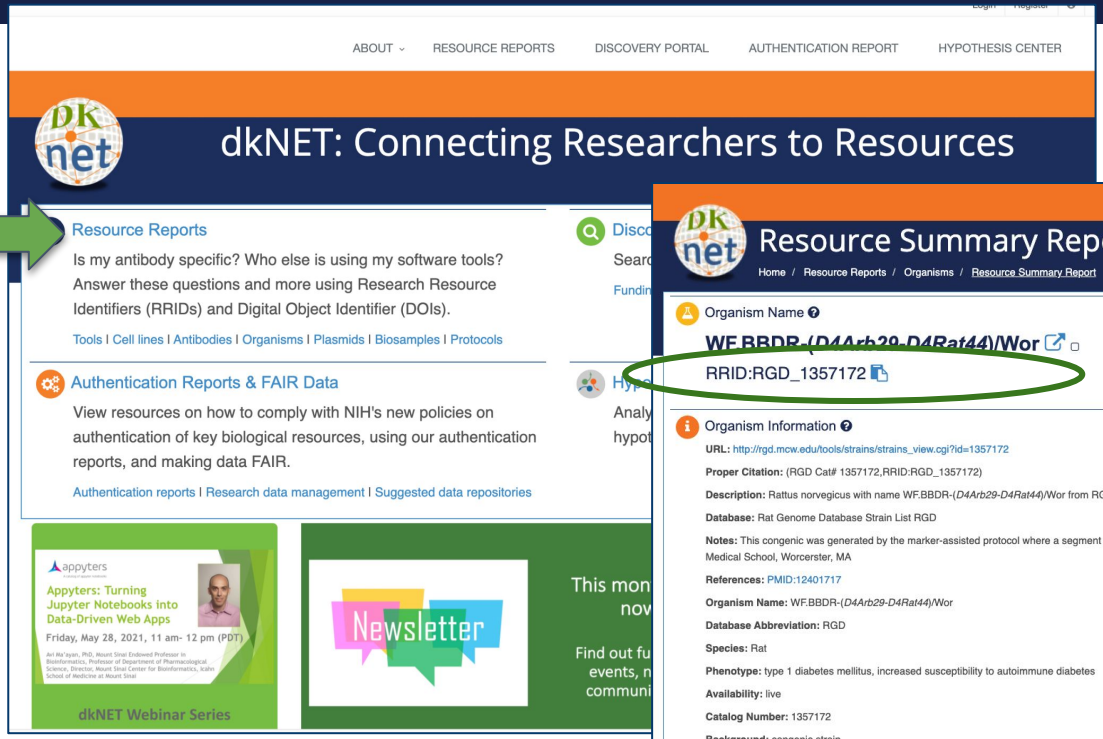


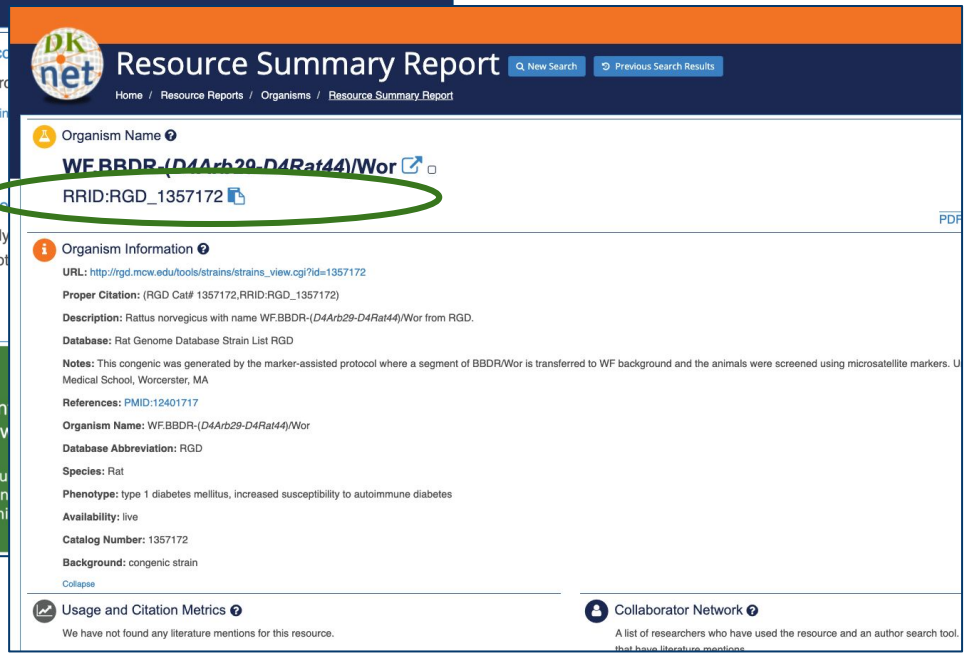
Secondary antibod		cam
	Rabbit IgG No. 21206, Alexa Fluor 488	Molecular Probes
Sheep IgG	No. A21448, Alexa Fluor 647	Molecular Probes
Chicken IgG	No. 703-605-155, Alexa Fluor 647	Jackson ImmunoResearch Laboratories (West Grove, PA)
Guinea pig IgG	No. 11073, Alexa Fluor 488	Molecular probes
Rabbit IgG	No. 31573, Alexa Fluor 647	Molecular probes

RRID Research Resource Identifier (https://dknet.org/)

Fothergill LJ et al. Cell and Tissue Research. 375 (2) 359-69, 2019.

Where can I find RRIDs? Ask dkNET!

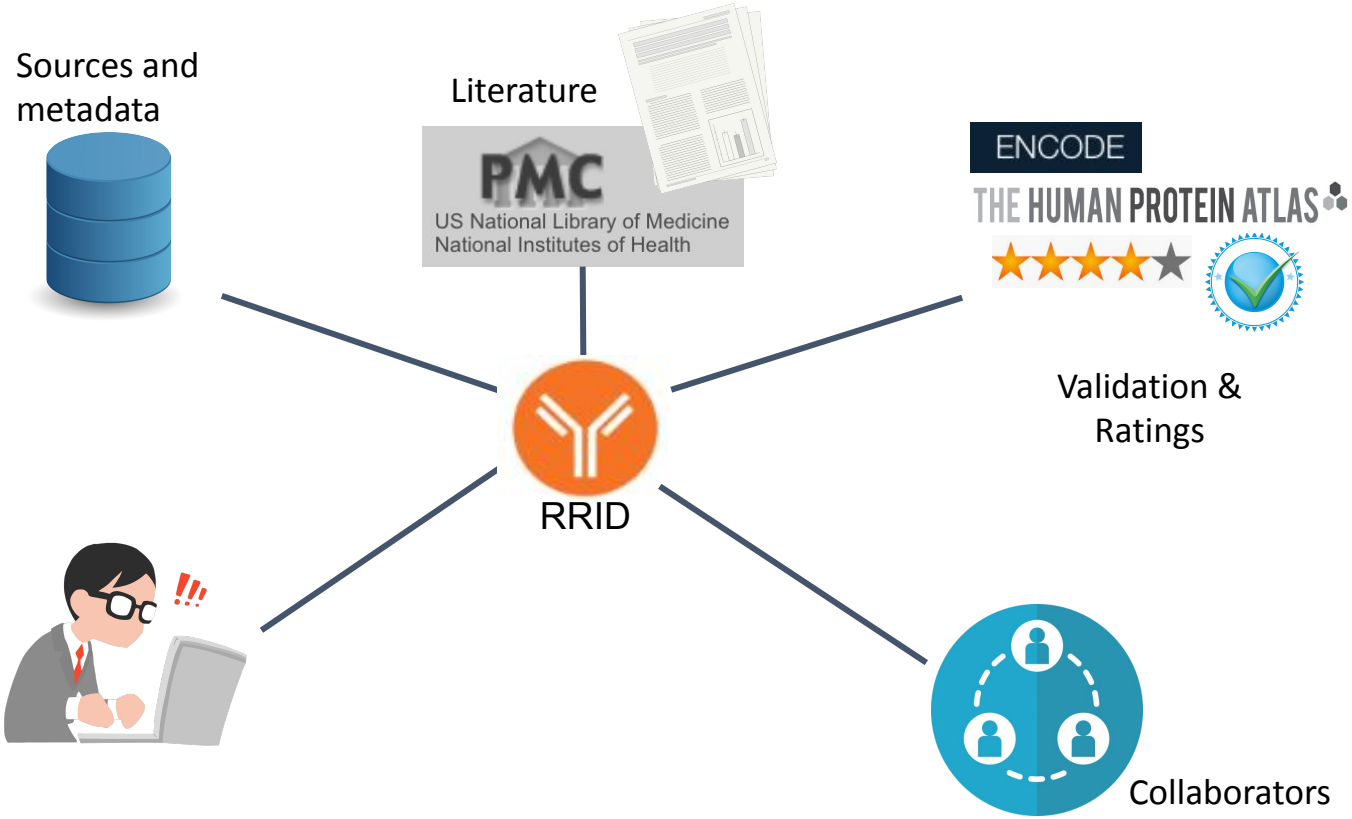




https://dknet.org/rin/rrids

PIDs make it easier to combine data across sources

- Computers PIDs like RRIDs
- dkNET uses RRIDs to create "Resource reports" which provide additional information about how these resources are used



Lesson: Use RRIDs and other IDs in your lab notebooks, spreadsheets, data dictionaries and papers!

Helps you and others (including computers) identify the tools that you use

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**

Lesson: Think about where your data will end up in the beginning

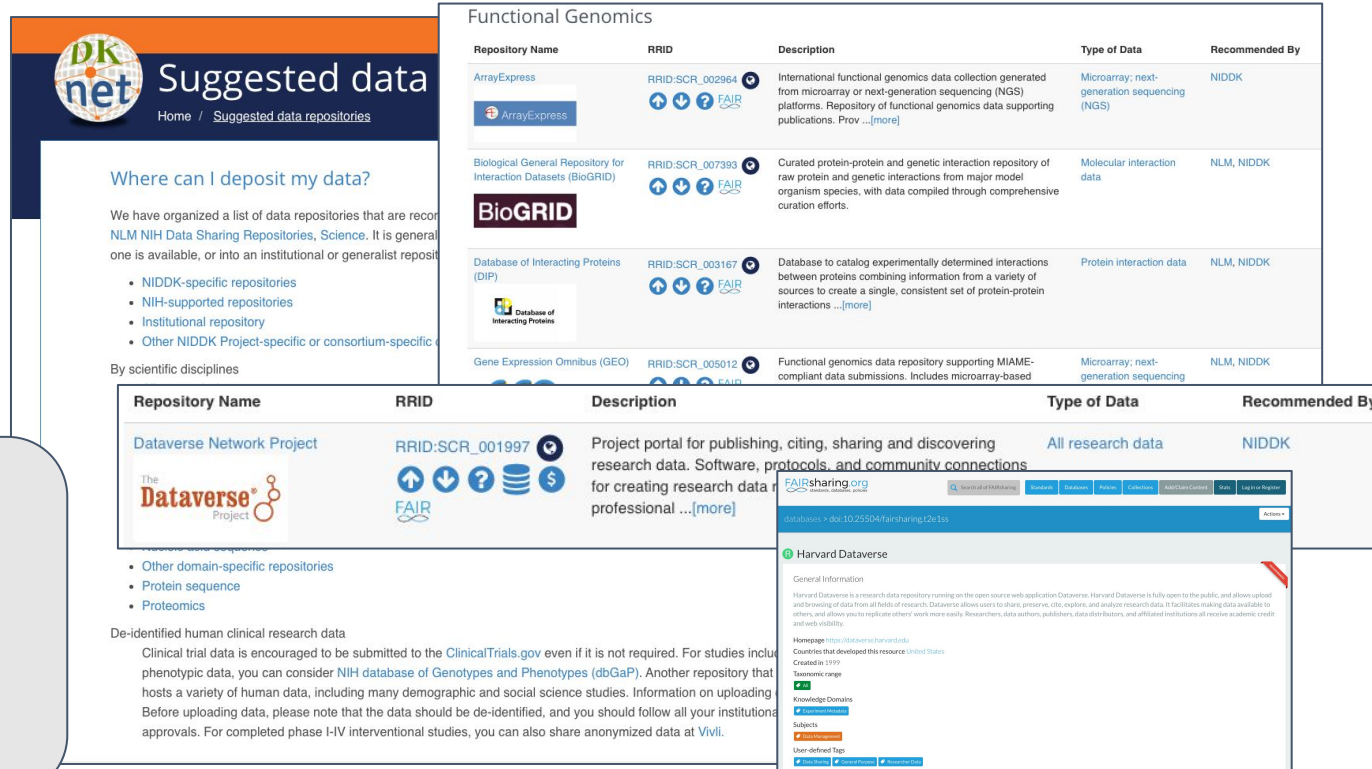
Best practice: Submit your data to repository specialized for your type of data or your domain

..if there isn't one, then there are also general purpose repositories available

Where Can I Deposit My Data?

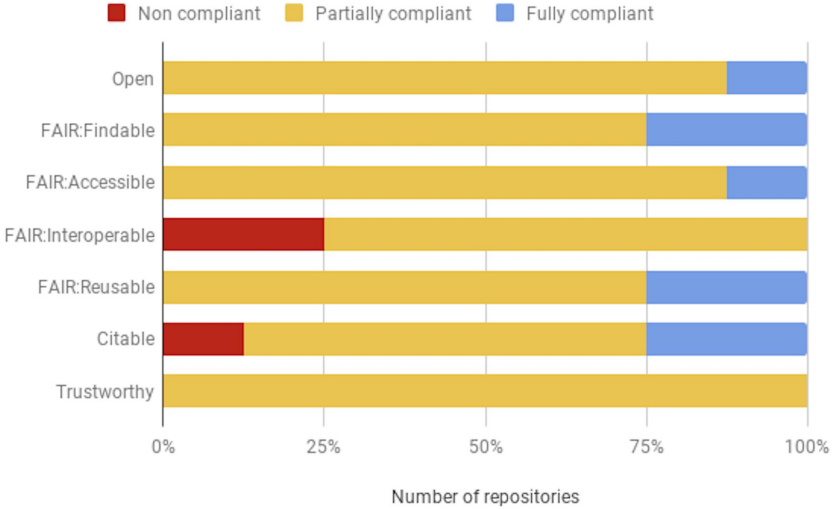
- List of DK relevant repositories, recommended by NLM and various journals
- Created in conjunction with NIDDK
- **Coming soon: FAIR data wizard**

- FAIR Standards
- Clinical Repositories Information
- Data maintenance
- Data size limit and cost
- Dynamic database



Data repositories

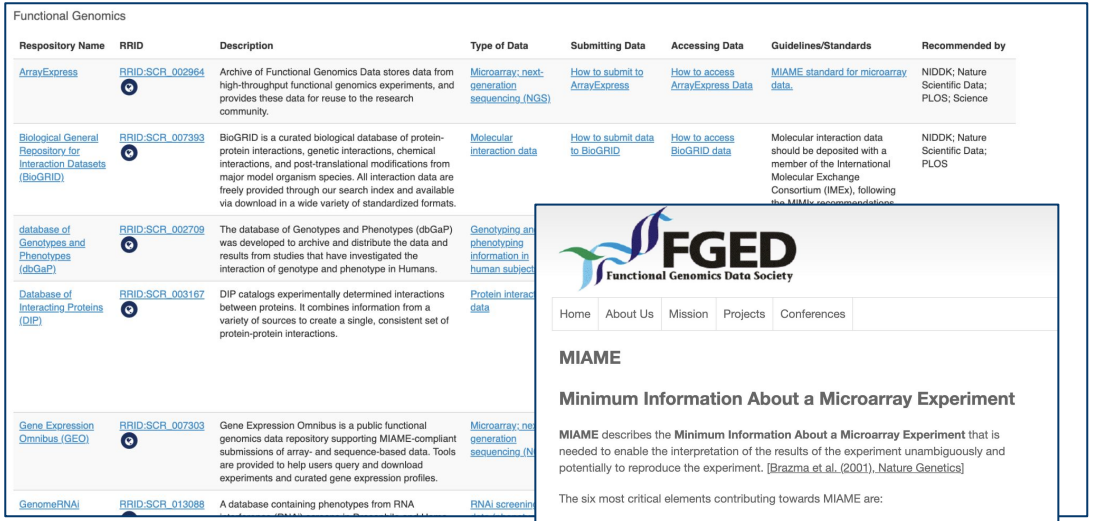
- “Publishing platform” for data
- Run by governments, researchers, commercial entities, non-profit entities, universities
- Hundreds of them across biomedicine
- Four major types:
  - Data type specific
  - Domain specific
  - Generalist
  - Institutional
- A stated goal of NIH’s strategic data plan is to modernize biomedical repositories
- Ensure persistence, searchability, FAIRness etc



Murphy et al., PLoS One, in press. Evaluation of biomedical data repositories against Open, FAIR, Citable and Trustworthy criteria
Preprint:
https://www.biorxiv.org/content/10.1101/2021.01.19.427362v2

What standards should I use?

- Repositories often enforce specific standards for metadata and data
- Thinking about where your data will end up before you start your experiments will help you determine how to collect, annotate and organize your (meta)data
- Fairsharing.org maintains a database of standards and policies across biomedicine



Functional Genomics

Respository Name	RRID	Description	Type of Data	Submitting Data	Accessing Data	Guidelines/Standards	Recommended by
ArrayExpress	RRID:SCR_002964	Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.	Microarray; next-generation sequencing (NGS)	How to submit to ArrayExpress	How to access ArrayExpress Data	MIAME standard for microarray data.	NIDDK; Nature Scientific Data; PLOS; Science
Biological General Repository for Interaction Datasets (BioGRID)	RRID:SCR_007393	BioGRID is a curated biological database of protein-protein interactions, genetic interactions, chemical interactions, and post-translational modifications from major model organism species. All interaction data are freely provided through our search index and available via download in a wide variety of standardized formats.	Molecular interaction data	How to submit data to BioGRID	How to access BioGRID data	Molecular interaction data should be deposited with a member of the International Molecular Exchange Consortium (IMEx), following the MIMIx recommendations	NIDDK; Nature Scientific Data; PLOS
database of Genotypes and Phenotypes (dbGaP)	RRID:SCR_002709	The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.	Genotyping and phenotyping information in human subject				
Database of Interacting Proteins (DIP)	RRID:SCR_003167	DIP catalogs experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions.	Protein interac data				
Gene Expression Omnibus (GEO)	RRID:SCR_007303	Gene Expression Omnibus is a public functional genomics data repository supporting MIAME-compliant submissions of array- and sequence-based data. Tools are provided to help users query and download experiments and curated gene expression profiles.	Microarray; ne generation sequencing (N				
GenomeRNAi	RRID:SCR_013088	A database containing phenotypes from RNA	RNAi screenin				

FGED Functional Genomics Data Society

Home | About Us | Mission | Projects | Conferences

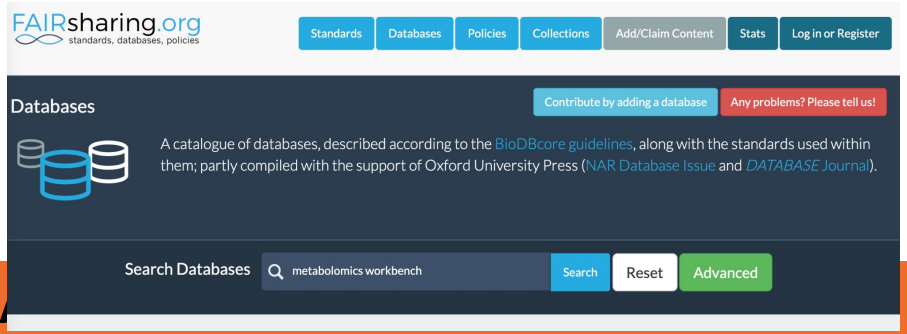
MIAME

Minimum Information About a Microarray Experiment

MIAME describes the **Minimum Information About a Microarray Experiment** that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. [Brazma et al. (2001), Nature Genetics]

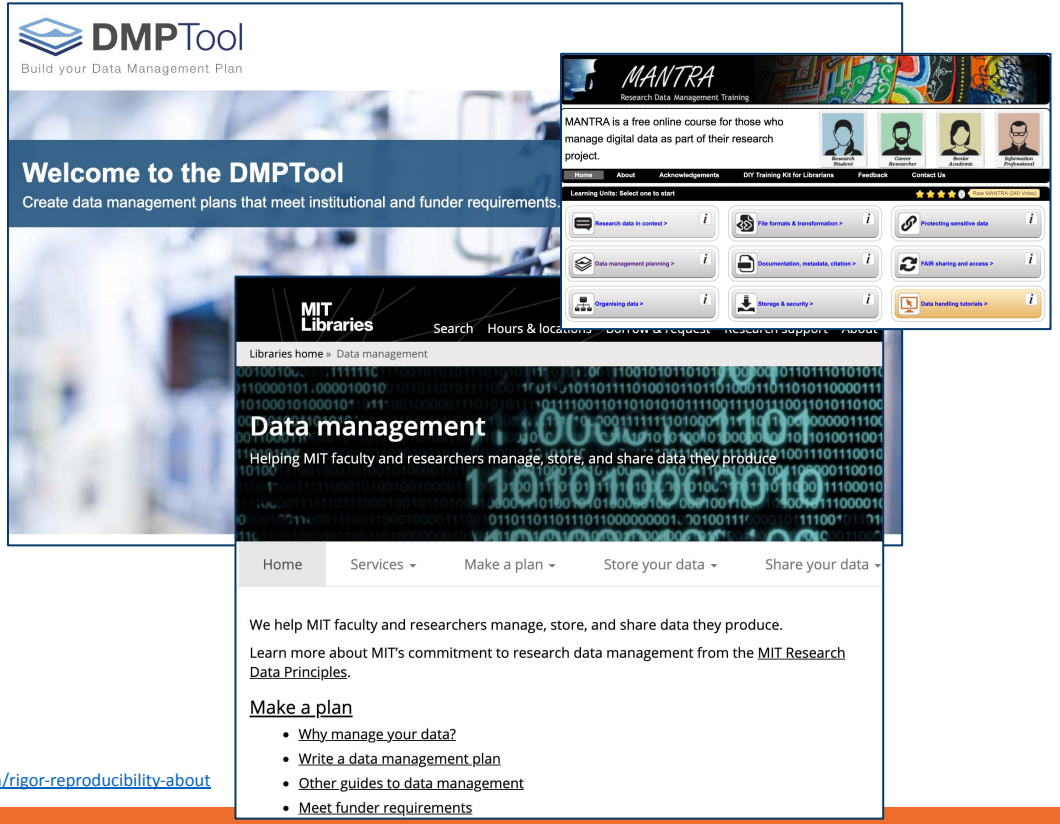
The six most critical elements contributing towards MIAME are:

1. The raw data for each hybridisation (e.g., CEL or GPR files)
2. The final processed (normalised) data for the set of hybridisations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
3. The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment)
4. The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridisations are technical, which are biological replicates)
5. Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
6. The essential laboratory and data processing protocols (e.g., what normalisation method has been used to obtain the final processed data)



Data Management and Sharing Plan

- Creating a good data management and sharing plan allows you to:
  - Comply with NIH mandates
  - Ensure that you allocate enough resources for preparing and sharing your data
  - Ensure that you collect your data in a FAIR manner
  - Easily share data with yourself, future you, your colleagues and the scientific community
- dkNET provides links to resources that can help



https://dknet.org/rin/rigor-reproducibility-about

R1.2: Provenance

- Metadata supplied with a dataset should include information about where and how the data were obtained
- Detailed experimental protocols
- Associated code
- Associated publications
- Standard authoring metadata
  - Also makes the data citable!

Effects of a fecal transplant from anxious donors on rehabilitative training, microbiota composition, systemic inflammation and behaviour following a unilateral cervical spinal contusion (C5, 125kdyn) in female Lewis rats

DOI:10.34945/F5XW2P

DATASET CITATION

Schmidt E. K. A.., Raposo P. J. F.., Madsen K. L., Fenrich K. K., Kabarchuk G., Fouad K. (2021) Effects of a fecal transplant from anxious donors on rehabilitative training, microbiota composition, systemic inflammation and behaviour following a unilateral cervical spinal contusion (C5, 125kdyn) in female Lewis rats. ODC-SCI:578 http://doi.org/10.34945/F5XW2P

ABSTRACT

STUDY PURPOSE: The purpose of this study was to determine whether optimal donor selection would influence the outcome of a fecal microbiota transplant (FMT) and the efficacy of rehabilitative training after a unilateral C5 spinal contusion (infinite horizons impactor, 125 kdyns) in female Lewis rats. Uninjured, genetically identical FMT donors (n=10) were selected as rats who displayed naturally reduced baseline activity levels and increased anxiety-like behaviour. Experimental groups consisted of a group that was gavaged with the FMT solution for 3 days following SCI (SCI+FMT, n=15) and a group that received a vehicle control solution (SCI+Vehicle, n=15).

DATA COLLECTED: Data collected for this study include:
16s rRNA sequencing of fecal matter collected pre-injury, on the day of injury, then 3, 7, 14 and 56 days post injury. This data includes the bacteria bacteroidetes, firmicutes, proteobacteria, cyanobacteria, lactobacillus, as well as the alpha diversity.
Rehabilitation training on a single pellet reaching task. This includes the number of attempts and success rate at baseline (pre-injury) and throughout 8 weeks of rehab training after SCI (10 min per day, 5 days per week, analyzed once per week).
Von Frey test at 1 and 9 weeks post SCI (normalized to baseline values).
The percentage of ipsilesional paw placements in the cylinder test (9 weeks post SCI).
The total distance travelled in the open field (9 weeks post SCI).
The total distance travelled and the percentage of time spent in the open arms of an elevated plus maze (9 weeks post SCI).
The time spent in the light component of the light-dark box test (9 weeks post SCI).
The percentage of sucrose solution consumed over 2 hours (9 weeks post SCI).
The time spent interacting in the social interaction test (9 weeks post SCI).
Lesion analysis.
IBA1 density immediately caudal to the lesion, at and immediately rostral to the lesion.
Plasma analytes were measured at baseline, 3, 21 and 77 days post SCI.
FITC dextran assay for intestinal permeability (7 days post SCI displayed as a fold change from baseline, this was a separate experiment which also includes lesion analysis and plasma analytes at 7 days post SCI).

DATA USAGE NOTES: The data presented in this dataset show that a fecal transplant from anxious donors (with decreased levels of Lactobacillus in their stool) does not prevent SCI-induced dysbiosis. This fecal transplant had some long-term side effects on systemic and local inflammation and also increased anxiety-like behaviour in the recipient rats. Overall, this study shows that optimal donor selection is critical for successful FMT treatment following SCI.

KEYWORDS

Spinal Cord Injury; Inflammation; Anxiety; Fecal transplant

PROVENANCE / ORIGINATING PUBLICATIONS

DATASET INFO

Contact: Fouad Karim (kfouad@ualberta.ca)

Lab: Karim Fouad
ODC-SCI Accession:578

Records in Dataset: 431
Fields per Record: 134

Files: 2

LICENSE

Creative Commons Attribution License (CC-BY 4.0)

FUNDING AND ACKNOWLEDGEMENTS

CONTRIBUTORS

Schmidt, Emma K. A.. [ORCID:0000-0001-9803-6391]
University of Alberta
Raposo, Pamela J. F.. [ORCID:0000-0001-6350-5223]
University of Alberta
Madsen, Karen L. [ORCID:0000-0002-8636-0714]
University of Alberta
Fenrich, Keith K. [ORCID:0000-0003-4360-064X]
University of Alberta
Kabarchuk, Gillian
University of Alberta
Fouad, Karim [ORCID:0000-0003-3654-7852]
University of Alberta

DOI:10.34945/F5XW2P

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.

DC1 Data Citation Principles

Joint Declaration of Data Citation Principles

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

A data citation looks like a regular citation

DOI

Full citation

DOI:10.34945/F5XW2P

Effects of a fecal transplant from anxious donors on rehabilitative training, microbiota composition, systemic inflammation and behaviour following a unilateral cervical spinal contusion (C5, 125kdyn) in female Lewis rats

DOI:10.34945/F5XW2P

DATASET CITATION

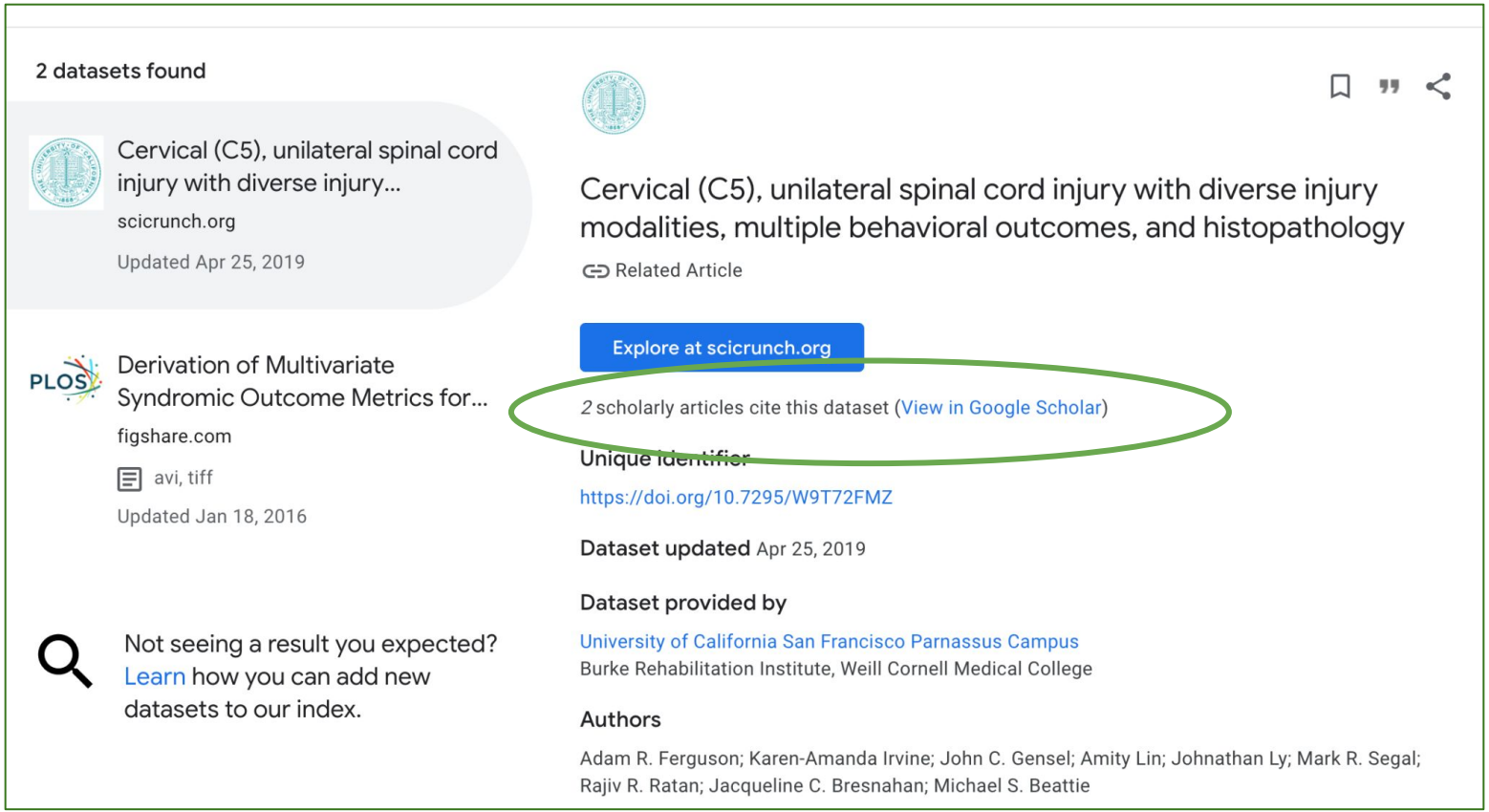
Schmidt E. K. A.., Raposo P. J. F.., Madsen K. L., Fenrich K. K., Kabarchuk G., Fouad K. (2021) Effects of a fecal transplant from anxious donors on rehabilitative training, microbiota composition, systemic inflammation and behaviour following a unilateral cervical spinal contusion (C5, 125kdyn) in female Lewis rats. ODC-SCI:578 http://doi.org/10.34945/F5XW2P

ABSTRACT

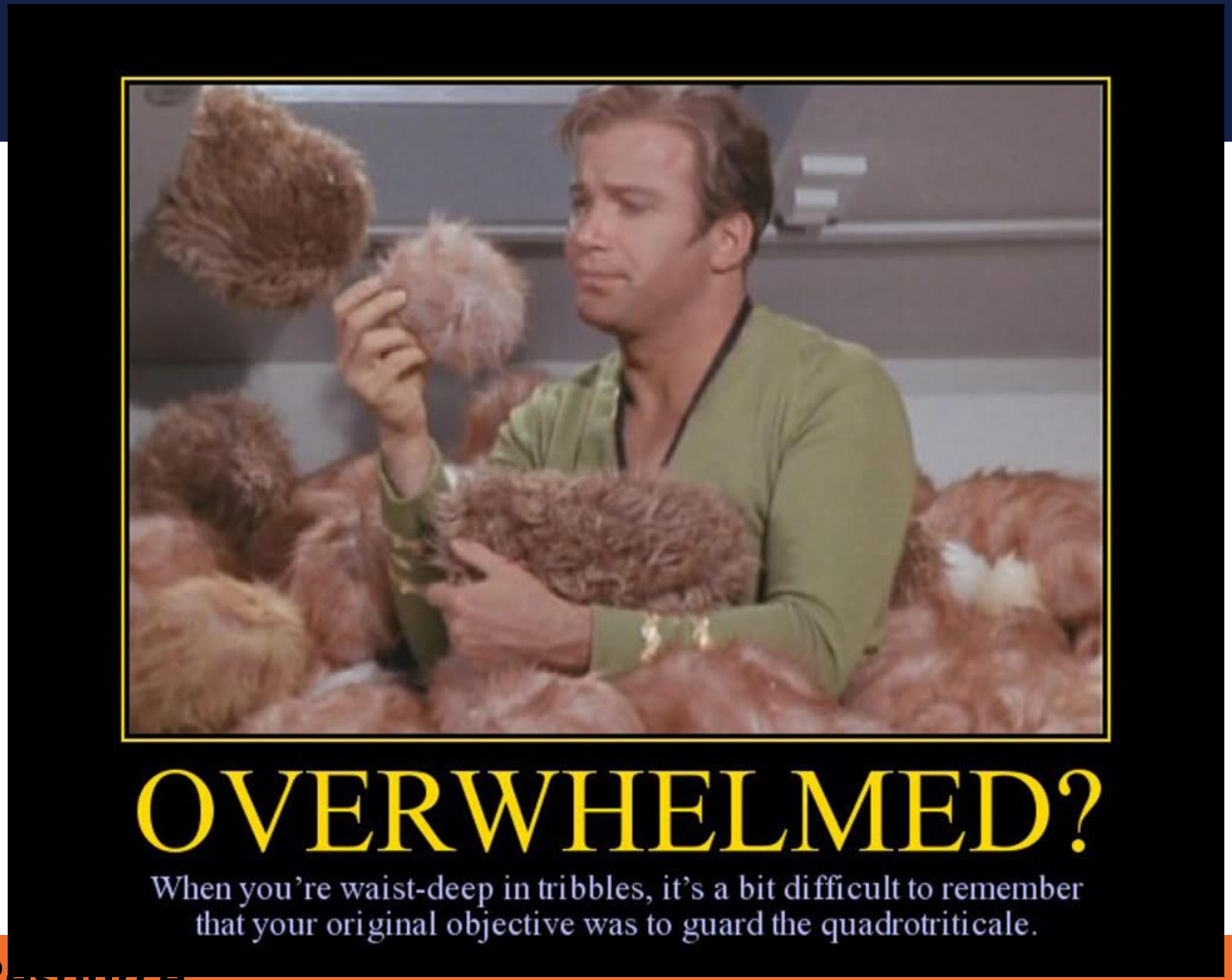
STUDY PURPOSE: The purpose of this study was to determine whether optimal donor selection would influence the outcome of a fecal microbiota transplant (FMT) and the efficacy of rehabilitative training after a unilateral C5 spinal contusion (infinite horizons impactor, 125 kdyns) in female Lewis rats. Uninjured, genetically identical FMT donors (n=10) were selected as rats who displayed naturally reduced baseline activity levels and increased anxiety-like behaviour. Experimental groups consisted of a group that was gavaged with the FMT solution for 3 days following SCI (SCI+FMT, n=15) and a group that received a vehicle control solution (SCI+Vehicle, n=15).

FDI Lab
FAIR Data Informatics

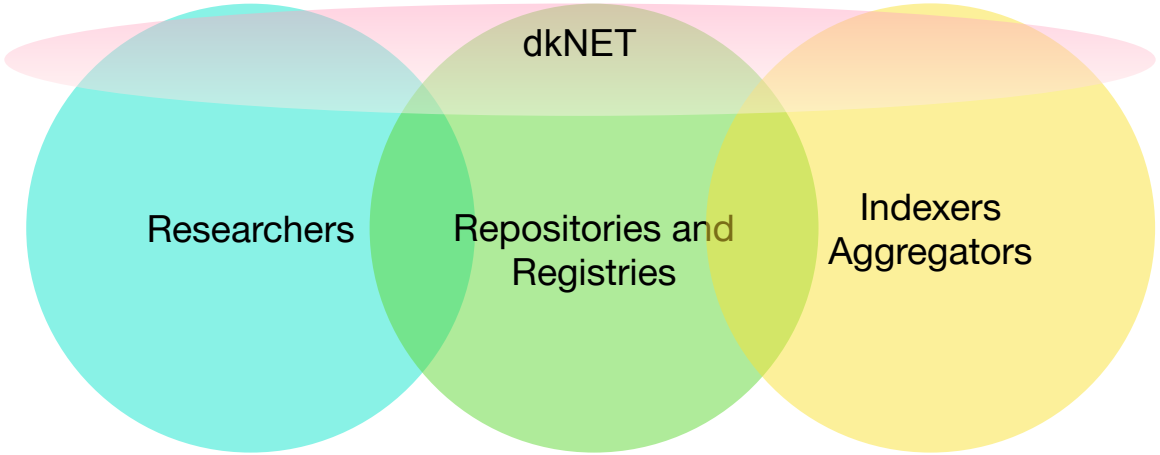
Proper data citation = data citation metrics



https://datasetsearch.research.google.com/


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

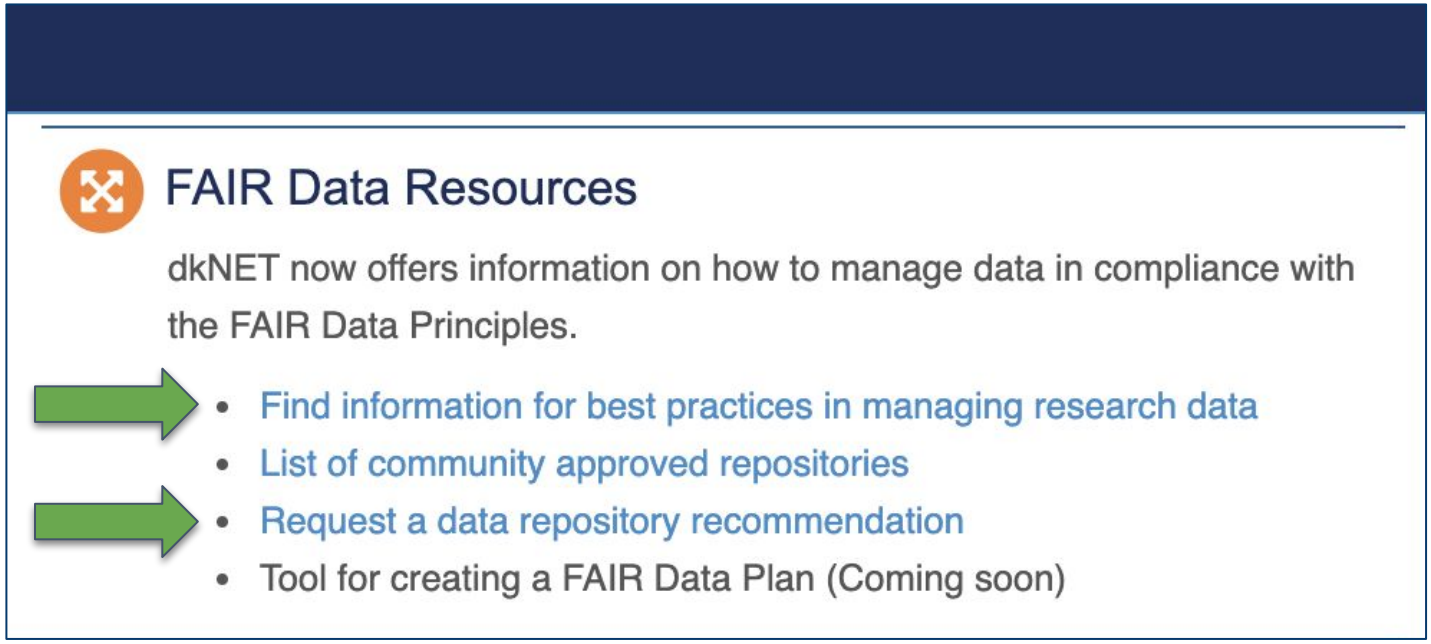


- **Good data management**
- **Rich metadata**
- **Prepare to share**
- Open formats
- **Adopt/align to standards**
- **Submit to repository**

- 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

Having trouble? Ask dkNET



Coming soon: The FAIR data wizard!

You are here

Changing the culture of biomedicine one step at a time...


thank you
danke
謝謝
ngiyabonga
شكراً جزيلاً
teşekkür ederim
gracias
merci
obrigado
dziękuję
bedankt
go raibh maith agat
sukriya
kop khun krap
terima kasih
grazie
arigatô
dakujem
mochchakkeram
tapadh leat
спасибо
рахмат
Баярлалаа
faafetai lava
kiitos
dankie
dhanyavad
nanni
nandri
bayarlalaa
gracie
hvala
mauruuru
köszönöm
enkosi
vinaka
спасибі
blagodaram
mersi
kia ora
barka
welalin
tack
spas
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
хвала
asante
manana
obrigada
murakoze
tenki
djiere dieuf
tau
akun
dankon
aciu
chnorakaloutioun
gratias ago
gràcies
sulpáy
дякую
mamnun
chokrane
sobodi
dēkuji
sagolun
mèsi
didi madloba
kam sah hamnida
najis tuke
rahmat
ありがとう
tanemirt
rahmet
taiku
takk
dakujem
trugarez
diolch
dhanyavadagalu
shukriya
merci
мерси
xièxie
감사합니다
ευχαριστώ
তোমাকে ধন্যবাদ