Sharing Data - Open MRI, A Big Data Case

Adam Thomas
Data Science and Sharing Team, Intramural Research Program, NIMH

Slides available at:
https://github.com/agt24/2023-02-06_big_data_sharing_UCinn

National Institute
of Mental Health

Who am I?

Adam Thomas
Team Lead

- Data Science and Sharing Team (DSST) NIMH Intramural Program, NIH Campus, Bethesda, MD
- Intramural != Extramural
- "~~But Adam said this was OK!~~"
- Team Mission: To support and advance the creation, distribution, and use of large, open datasets to accelerate discovery & help researchers learn and practice open and reproducible science practices.

http://cmn.nimh.nih.gov/dsst

NIMH IRP Data Science and Sharing Team

Adam Thomas
Team Lead

Dustin Moraczewski
Data Scientist

Arshitha Basavaraj
Data Engineer

Jessica Dafflon
Data Scientist

Eric Earl
Data Scientist

Anthony Galassi
Software Engineer

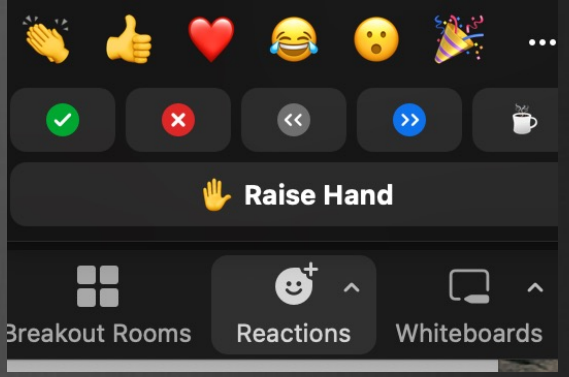
Carl Harris
Post-bac IRTA

Francisco Pereira
Machine Learning Team Lead

http://cmn.nimh.nih.gov/dsst

The 2023 Data Sharing Policy

Poll



What's your experience with data sharing?

👏 I've written/published papers using one or more public shared dataset

👍 I've downloaded shared data sets, but haven't yet included shared data in a publication

❤️ I've deposited data into a public data repository

😂 My trainees/staff have uploaded data from our group, but I don't know what's involved

😮 I've drafted a Data Management & Sharing Plan for submission to the NIH

Topics

- The 2023 NIH Data Sharing
  - How we got here
- Technical challenges
- Privacy Challenges

The 2023 Data Sharing Policy

https://sharing.nih.gov/

2023 – The Year of Open Science

THE WHITE HOUSE

MENU

JANUARY 11, 2023

FACT SHEET: Biden-Harris Administration Announces New Actions to Advance Open and Equitable Research

OSTP ▸ BRIEFING ROOM ▸ PRESS RELEASES

OSTP launches Year of Open Science to advance national open science policies across the federal government in 2023

Today, the White House Office of Science and Technology Policy (OSTP) announced new actions to advance open and equitable research, including new grant funding, improvements in research infrastructure, broadened

Under the DMS policy, NIH expects that investigators and institutions:

https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research/

Data Sharing Policy History

FINAL NIH STATEMENT ON SHARING RESEARCH DATA

RELEASE DATE: February 26, 2003

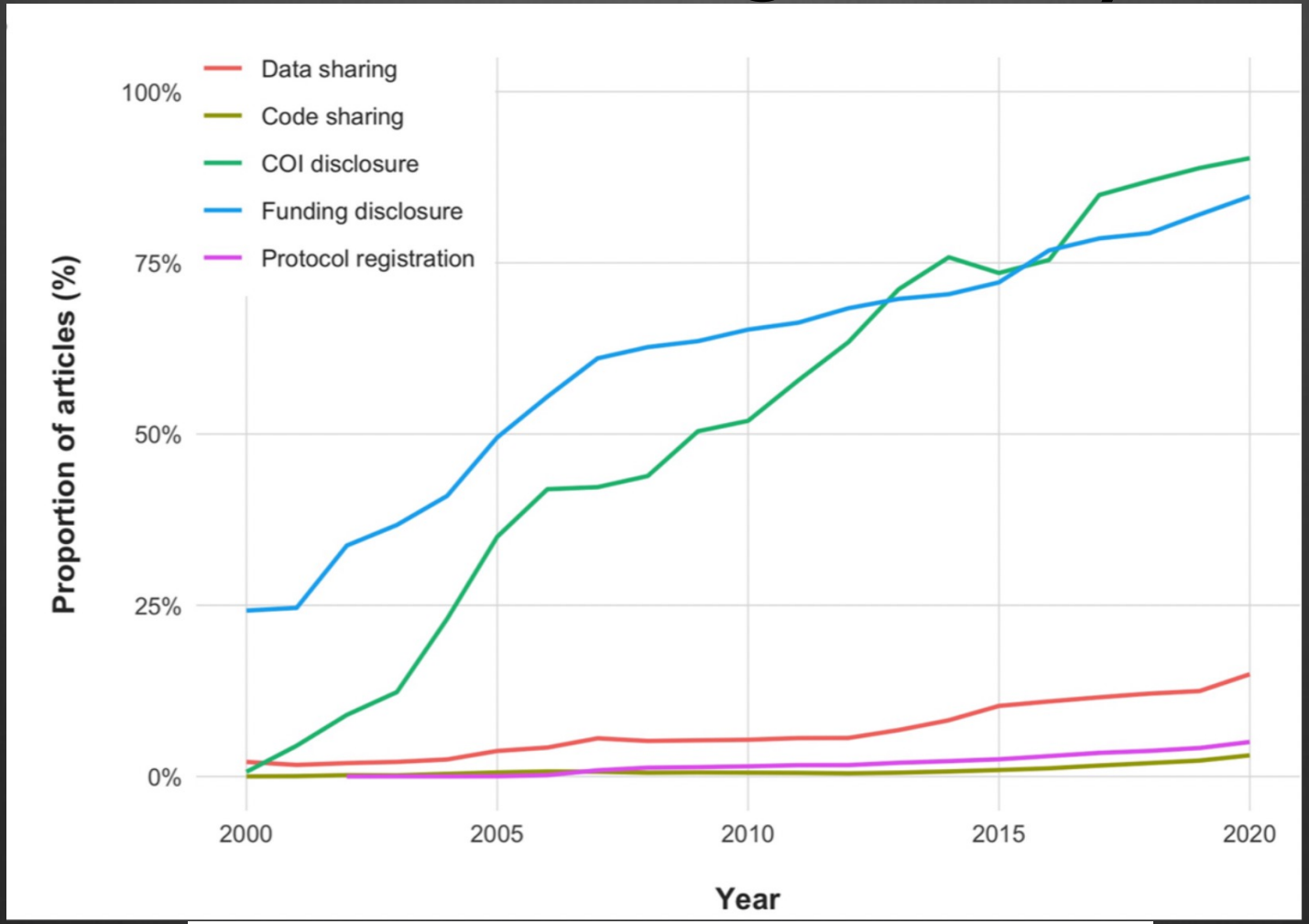
NOTICE: NOT-OD-03-032

National Institutes of Health ((NIH)

As part of NIH's long-standing policy to share and make available to the public the results and accomplishments of the activities that it funds, NIH announced and invited comments on a draft statement about the sharing of final research data on March 1, 2002. Since that time, NIH has received and reviewed many thoughtful comments from a range of scientific organizations and over 150 individuals. Additionally, during

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

Data Sharing History



META-RESEARCH ARTICLE

https://doi.org/10.1371/journal.pbio.3001107

Assessment of transparency indicators across the biomedical literature: How open is open?

Stylianos Serghiou[1,2], Despina G. Contopoulos-Ioannidis[3], Kevin W. Boyack[4], Nico Riedel[5], Joshua D. Wallach[6], John P. A. Ioannidis[1,2,7,8,9]*

Data Repositories: History

NeuroImage

Volume 82, 15 November 2013, Pages 677-682

Review

Why share data? Lessons learned from the fMRIDC

John Darrell Van Horn [a], Michael S. Gazzaniga [b, 1]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.neuroimage.2012.11.010

Data Sharing History

"A growing chorus of concern, from scientists and laypeople, contends that the complex system for ensuring the reproducibility of biomedical research is failing and is in need of restructuring."
"As leaders of the US National Institutes of Health (NIH), we share this concern and here explore some of the significant interventions that we are planning."

Published: 27 January 2014

Policy: NIH plans to enhance reproducibility

Francis S. Collins & Lawrence A. Tabak

Nature **505**, 612–613 (2014) | Cite this article https://doi.org/10.1038/505612a

Data Repositories: Present

NEWS | 16 February 2022 | Correction 16 February 2022

NIH issues a seismic mandate: share data publicly

The data-sharing policy could set a global standard for biomedical research, scientists say, but they have questions about logistics and equity.

Max Kozlov

https://doi.org/10.1038/d41586-022-00402-1

Data Repositories: Present

https://sharing.nih.gov/

2023 NIH Grants Virtual Conference

videos available online tomorrow

2-Day Virtual Conference

NIH Grants Conference: Wednesday, February 1 - Thursday, February 2

Experience the NIH Grants Conference first-hand…there is no other event that brings hundreds of NIH experts and others in the research community together like this event!

- 25 Sessions with live Q&A
- 45 NIH Institute and Office Exhibitors
- NIH Program, Grants Management, Review, & Policy Officials
- 1:1 Meet the Experts chats
- Peer-to-Peer Networking opportunities
- Downloadable resources
- And so much more!

https://grants.nih.gov

Wed, Feb 1st : 4:00 - 4:45 PM

The NIH Final Policy for Data Management and Sharing is in Effect: Planning for Success!

What are you most concerned about?

a) Technical challenges of sharing large, complicated MRI datasets

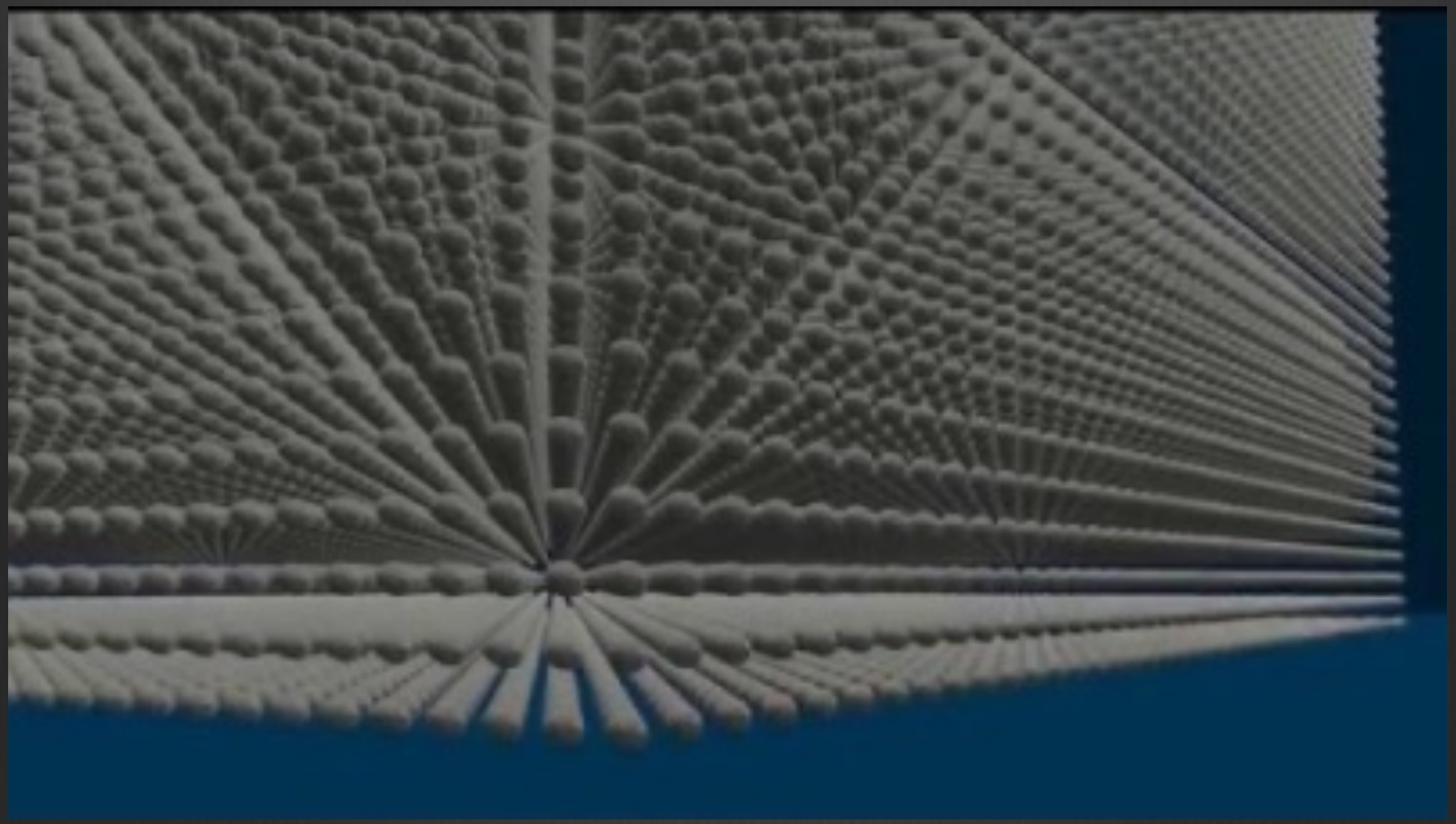
b) Privacy and ethics challenges of sharing MRI data (de-facing)

c) Something else

Topics

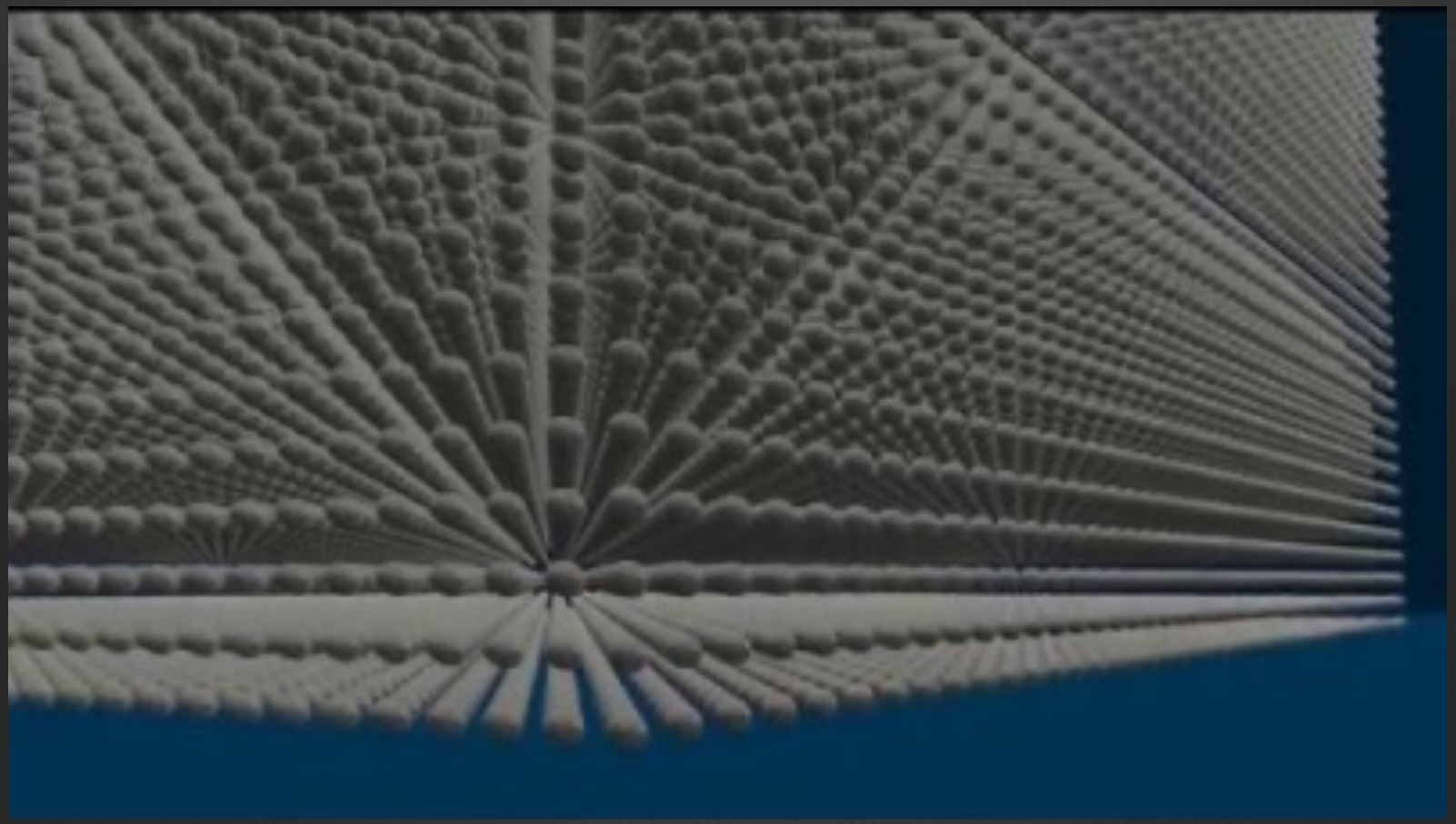
- The 2023 NIH Data Sharing
  - History
- Technical challenges
  - Are your data too big?
- Privacy Challenges

Is your data too big to share?



https://www.youtube.com/watch?v=DbPNscjIC6U

UK Biobank, 100,000 MRI Scans

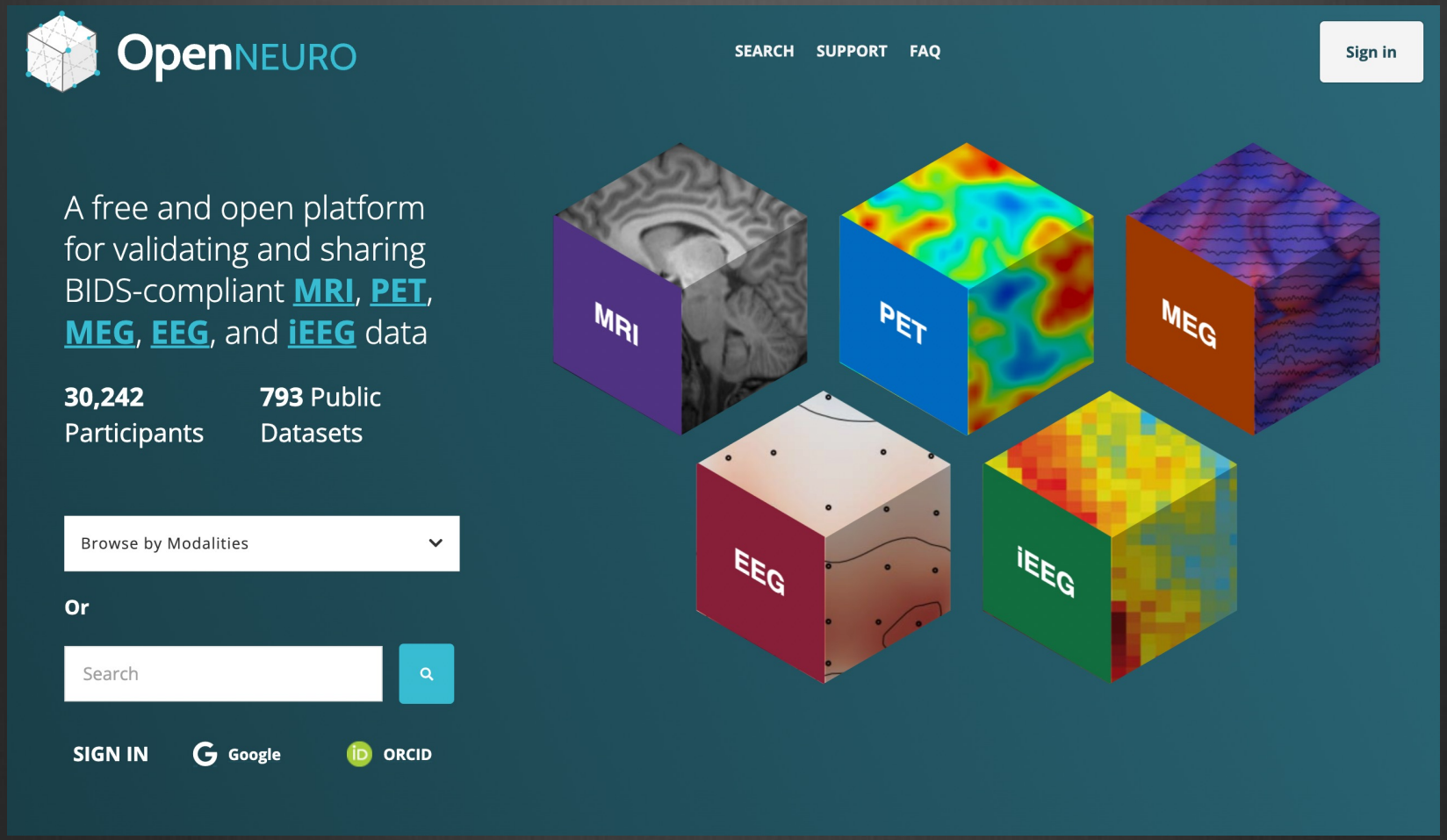


https://www.youtube.com/watch?v=DbPNscjIC6U

UK Biobank, 100,000 MRI scans

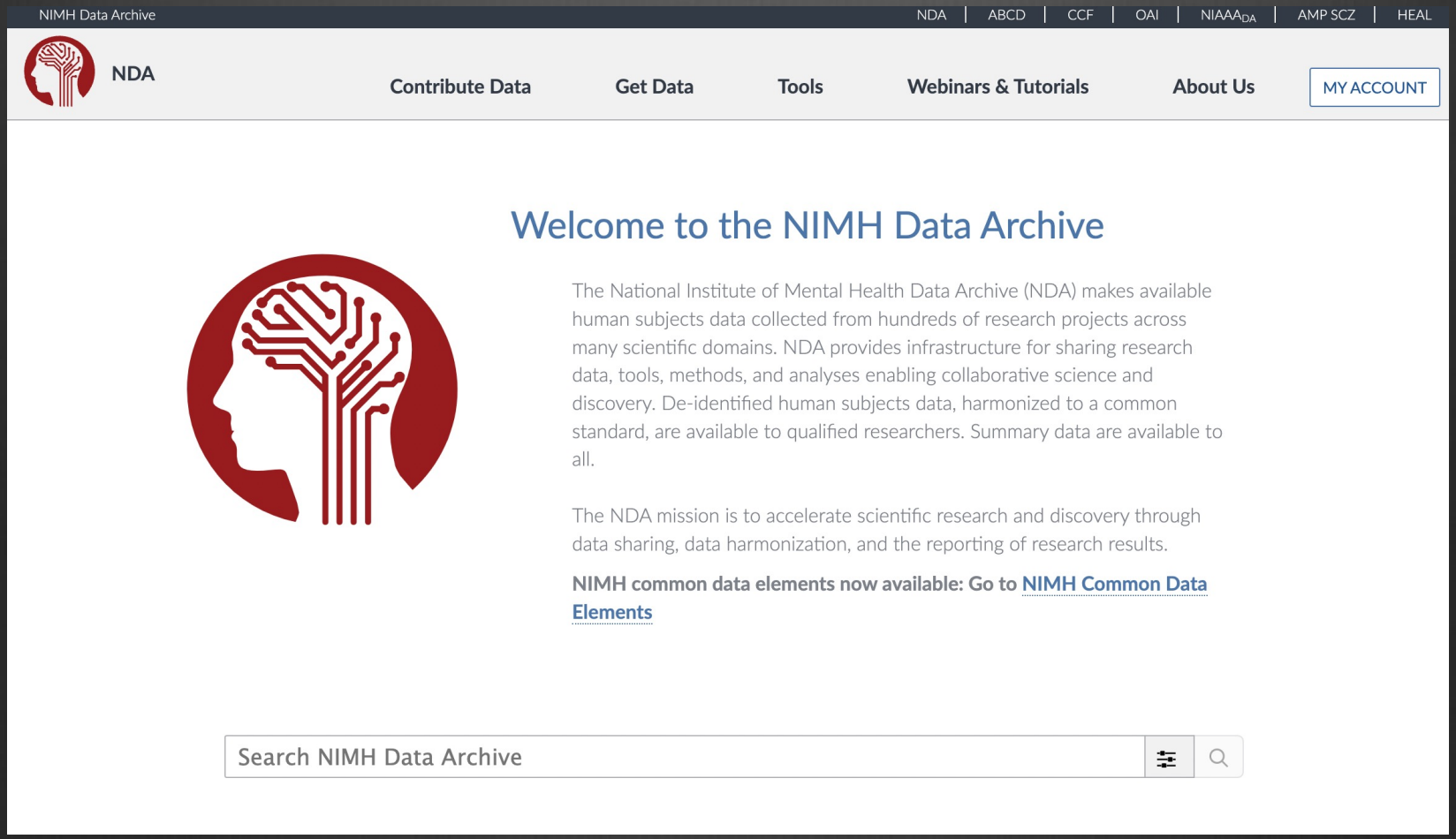
https://www.youtube.com/watch?v=DbPNscjIC6U

OpenNeuro.org



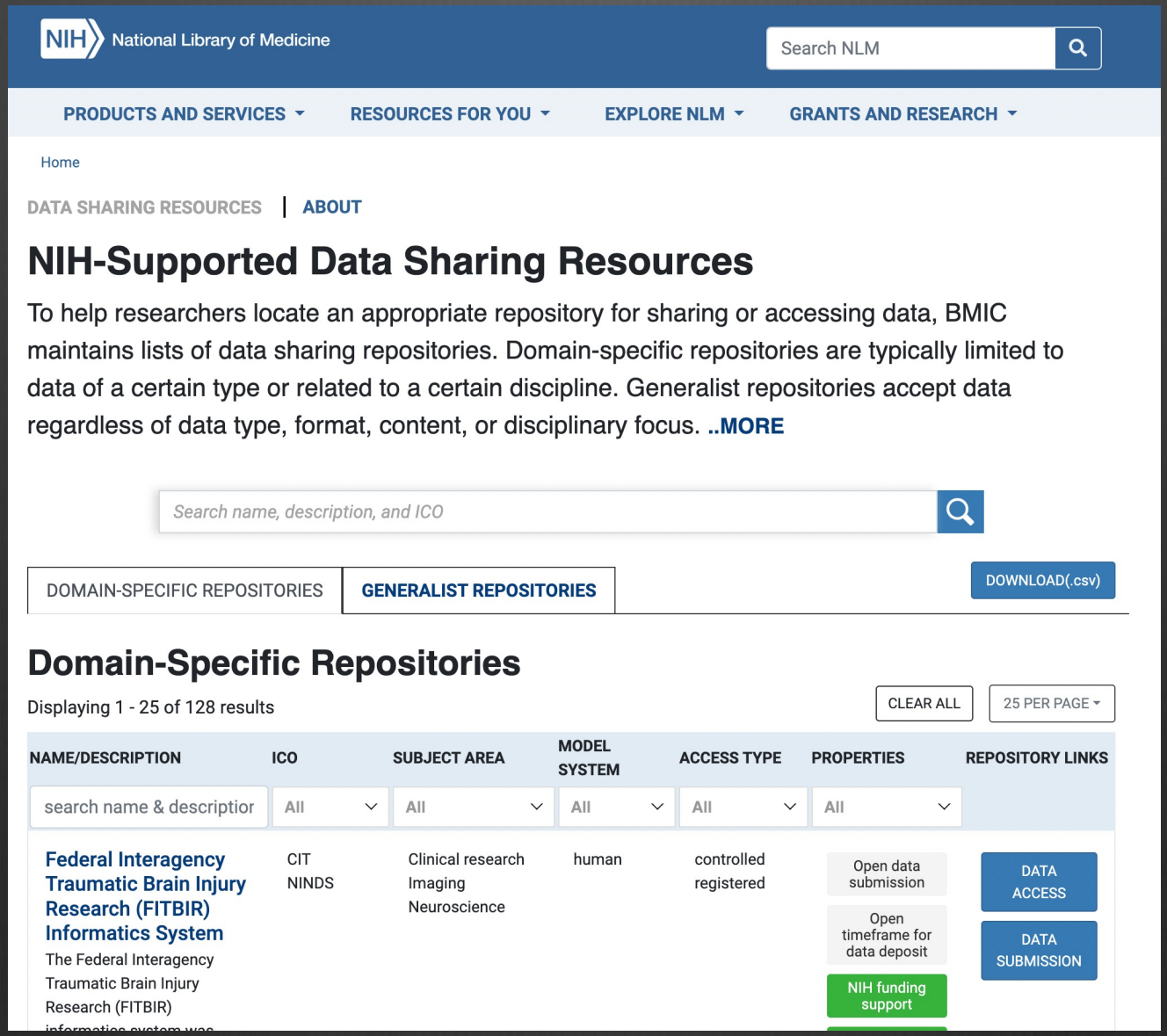
https://openneuro.org/

nda.nih.gov



https://nda.nih.gov/

NIH listing of Repositories



https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html

Finding a Good Repository

NIH SCIENTIFIC DATA SHARING

Search

NIH Staff | FAQ | Contacts & Help

DATA MANAGEMENT AND SHARING POLICY | GENOMIC DATA SHARING POLICY | OTHER SHARING POLICIES | ACCESSING DATA | ABOUT

Home > Data Management and Sharing Policy > Sharing Scientific Data > Repositories for Sharing Scientific Data

Repositories for Sharing Scientific Data

In general, NIH does not endorse or require sharing data in any particular repository, although some initiatives and funding opportunities will have individual requirements. **Overall, NIH encourages researchers to select the repository that is most appropriate for their data type and discipline. See Selecting a Data Repository.**

Browse through this listing of NIH-supported repositories to learn more about some places to share scientific data. Note that this list is not exhaustive. Select the link provided in the "Data Submission Policy" column to find data submission instructions for each repository.

Can't find a repository that suits your data? Here are several more resources:

- A listing of generalist repositories that accepts all data types
- Nature's Data Repository Guidance
- The Registry of Research Data Repositories

See **Accessing Scientific Data** for more information about accessing data from NIH-supported repositories.

https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/selecting-a-data-repository

Finding a Good Repository

Desirable Characteristics for All Data Repositories

When choosing a repository to manage and share data resulting from Federally funded research, here are some desirable characteristics to look for:

- **Unique Persistent Identifiers:** Assigns datasets a citable, unique persistent identifier, such as a digital object identifier (DOI) or accession number, to support data discovery, reporting, and research assessment. The identifier points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.
- **Long-Term Sustainability:** Has a plan for long-term management of data, including maintaining integrity, authenticity, and availability of datasets; building on a stable technical infrastructure and funding plans; and having contingency plans to ensure data are available and maintained during and after unforeseen events.
- **Metadata:** Ensures datasets are accompanied by metadata to enable discovery, reuse, and citation of datasets, using schema that are appropriate to, and ideally widely used across, the community(ies) the repository serves. Domain-specific repositories would generally have more detailed metadata than generalist repositories.
- **Curation and Quality Assurance:** Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.
- **Free and Easy Access:** Provides broad, equitable, and maximally open access to datasets and their metadata free of charge in a timely manner after submission, consistent with legal and ethical limits required to maintain privacy and confidentiality, Tribal sovereignty, and protection of other sensitive data.
- **Broad and Measured Reuse:** Makes datasets and their metadata available with broadest possible terms of reuse; and provides the ability to measure attribution, citation, and reuse of data (i.e., through assignment of adequate metadata and unique PIDs).
- **Clear Use Guidance:** Provides accompanying documentation describing terms of dataset access and use (e.g., particular licenses, need for approval by a data use committee).
- **Security and Integrity:** Has documented measures in place to meet generally accepted criteria for preventing unauthorized access to, modification of, or release of data, with levels of security that are appropriate to the sensitivity of data.
- **Confidentiality:** Has documented capabilities for ensuring that administrative, technical, and physical safeguards are employed to comply with applicable confidentiality, risk management, and continuous monitoring requirements for sensitive data.
- **Common Format:** Allows datasets and metadata downloaded, accessed, or exported from the repository to be in widely used, preferably non-proprietary, formats consistent with those used in the community(ies) the repository serves.
- **Provenance:** Has mechanisms in place to record the origin, chain of custody, and any modifications to submitted datasets and metadata.
- **Retention Policy:** Provides documentation on policies for data retention within the repository.

https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/selecting-a-data-repository#desirable-characteristics-for-all-data-repositories

BIDS: Brain Imaging Data Structure

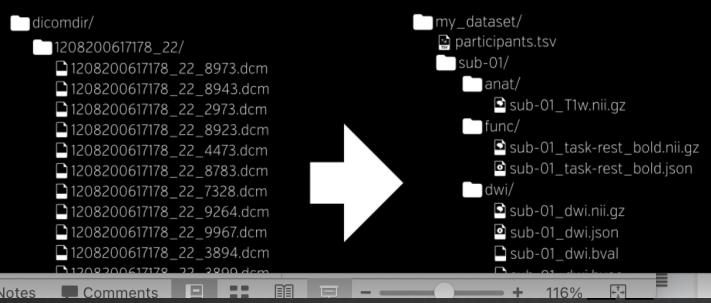
Brain Imaging Data Structure

A simple and intuitive way to organize and describe your neuroimaging and behavioral data.

ABOUT NEWS BENEFITS SPECIFICATION GET STARTED GET INVOLVED GOVERNANCE ACKNOWLEDGMENTS

About BIDS

Neuroimaging experiments result in complicated data that can be arranged in many different ways. So far there is no consensus how to organize and share data obtained in neuroimaging experiments. Even two researchers working in the same lab can opt to arrange their data in a different way. Lack of consensus (or a standard) leads to misunderstandings and time wasted on rearranging data or rewriting scripts expecting certain structure. With the Brain Imaging Data Structure (BIDS), we describe a simple and easy to adopt way of organizing neuroimaging and behavioral data.



Tweets from @BIDSstandard

BIDS-standard @BIDSstandard · Dec 15, 2022

BIDS is now also on Mastodon.

@bidsstandard@fosstodon.org

14

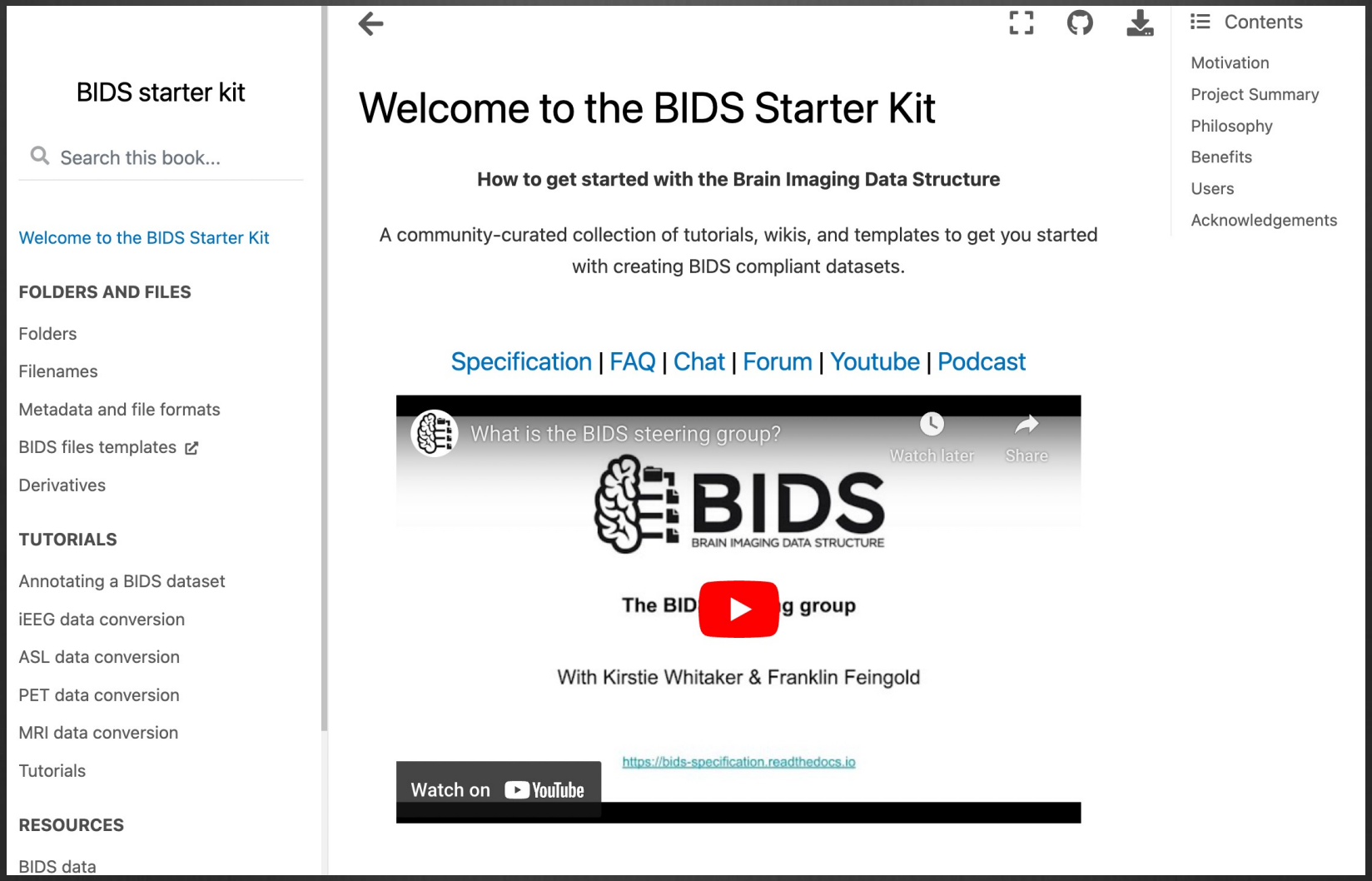
BIDS-standard Retweeted

Cyril R Pernet @CyrilRPernet · Nov 8, 2022

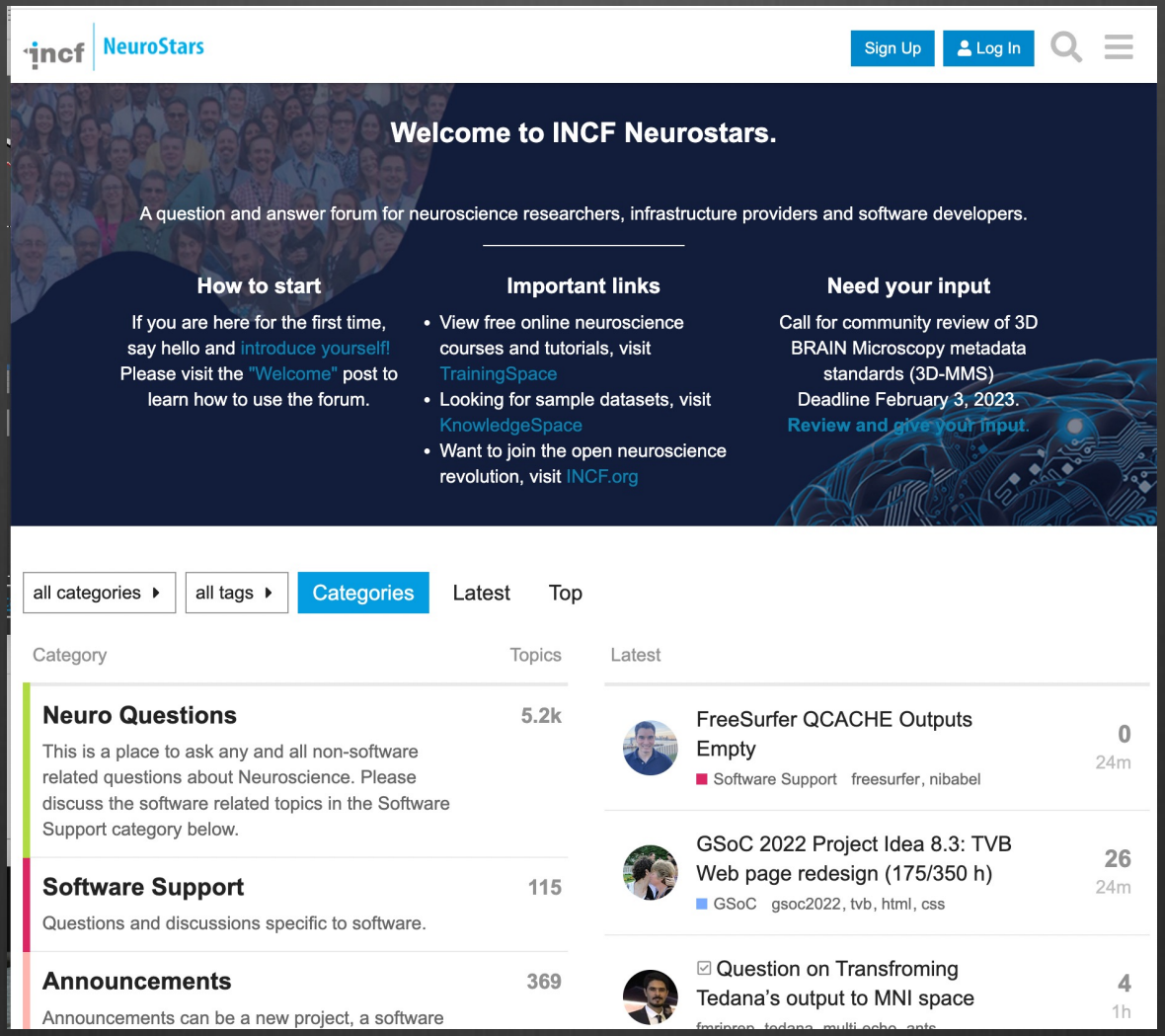
Results are in app.rankedvote.co/contests/24735... Thank you all who voted, for me or not. Turns out I will be helping steer BIDSstandard@mastodon.online (where is the Mastodon handle btw?) for the next few years

Notes Comments 116% Sheet1

The BIDS Starter Kit



Support: Neurostars.org



https://neurostars.org/

Topics

- The 2023 NIH Data Sharing
  - History
- Technical challenges
  - Are your data too big?
- Privacy Challenges

Privacy Concerns

Received: 9 July 2021 | Revised: 19 January 2022 | Accepted: 28 January 2022

DOI: 10.1002/hbm.25803

REVIEW ARTICLE

WILEY

The spectrum of data sharing policies in neuroimaging data repositories

Anita S. Jwa | **Russell A. Poldrack**

Department of Psychology, Stanford University, Stanford, California, USA

Correspondence
Anita S. Jwa, Department of Psychology, Stanford University, 450 Jane Stanford Way, Building 420, Stanford, CA 94305, USA.
Email: anniejwa@stanford.edu

Funding information
National Institute of Mental Health, Grant/Award Number: 3R24MH117179-03S1

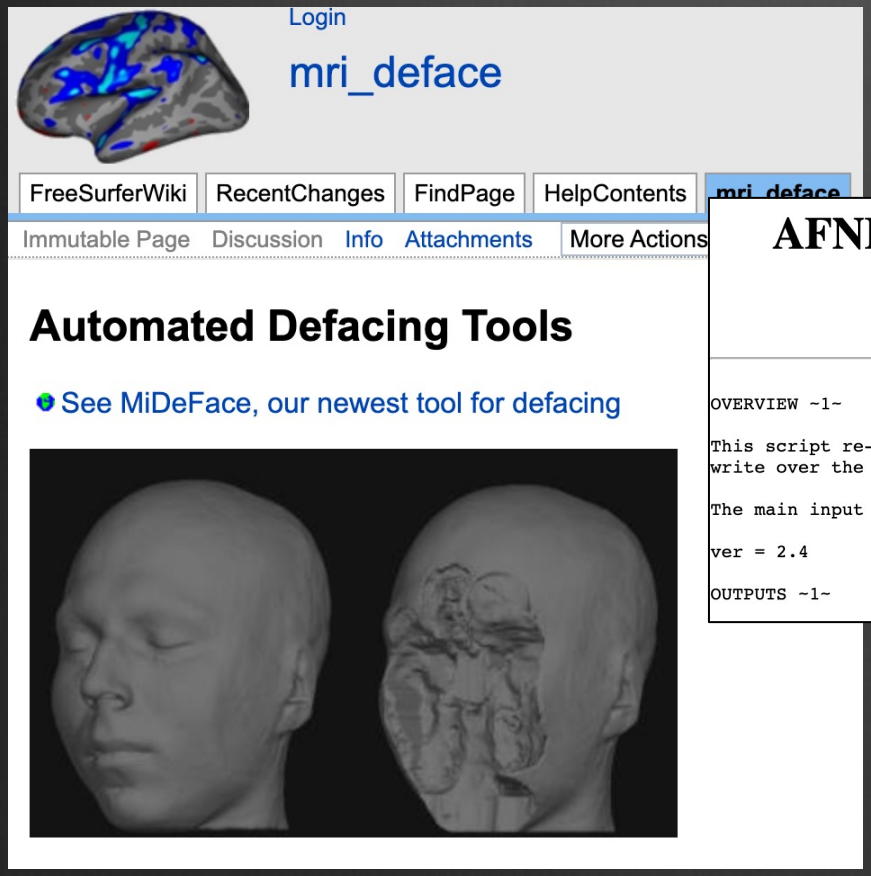
Abstract

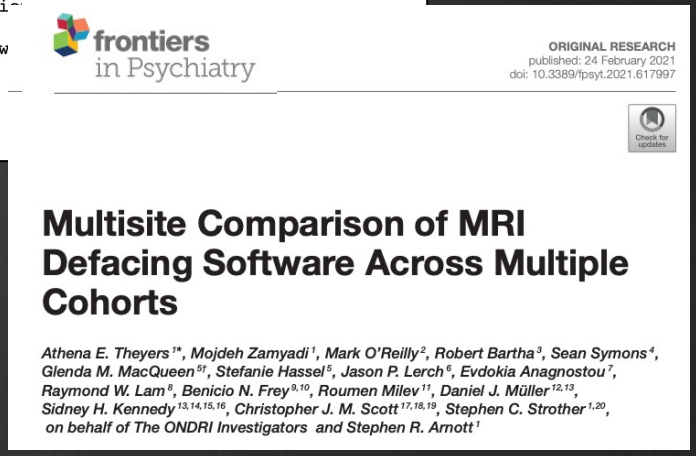
Sharing data is a scientific imperative that accelerates scientific discoveries, reinforces open science inquiry, and allows for efficient use of public investment and research resources. Considering these benefits, data sharing has been widely promoted in diverse fields and neuroscience has been no exception to this movement. For all its promise, however, the sharing of human neuroimaging data raises critical ethical and legal issues, such as data privacy. Recently, the heightened risks to data privacy posed by the rapid advances in artificial intelligence and machine learning techniques have

http://dx.doi.org/10.1002/hbm.25803

Privacy Concerns

- All MRI data shared without a data use agreement should be de-faced and de-identified





Privacy Concerns

- All MRI data shared without a data use agreement should be de-faced and de-identified
- Techniques exist to re-identify de-identified data, but data on their effectiveness is limited
- The fact that it is possible to re-identify subjects does not necessarily mean that it is likely or harmful
- Subjects suffer harms only when sensitive information disclosed from re-identification is used to jeopardize the subjects' interests (disease state or biomarker used against subject's interest)

Privacy Concerns

- Secondary analysis on de-identified neuroimaging data is not considered human subject research under the Common Rule thus informed consent and IRB review are not required.
- De-identified data do not constitute protected health information under the HIPAA.
- Once data are de-identified, the current US regulatory system primarily relies on researchers to make ethical and judicious decisions on how and where to share their data (Ross et al.,2018).
- Researchers are expected to follow ethical principles and community norms in selecting data repositories and limiting subsequent use of data.

Privacy Concerns

- Participants should be made aware of re-identification risks, no matter how small, as part of the consent procedures
- Additional regulations against malicious use of all biomedical data collected in research studies (e.g. Genetic Information Non-discrimination Act, GINA) would help relieve concerns around remaining risk of re-identification

Take Homes

- NIH's new data management and sharing policy asks grant recipients to make a considerable investment of time and resources to make their datasets Findable, Accessible, Interoperable, Re-usable (FAIR)
- Many resources are available to ease the burden and limit privacy concerns
- Planning ahead will dramatically decrease the time and resources required

Thanks!

See online slides for more URLs and references:

https://github.com/agt24/

http://cmn.nimh.nih.gov/dsst

Questions?